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 poster*

## Sommario


ELISA DI DOMENICO
I *si* di Univers-ITA 3

DARIA EVANGELISTA
Metodi di annotazione per l'analisi funzionale delle metafore in *ClimIta*, un corpus di testi scritti prodotti da media e istituzioni sul tema della crisi climatica 6

ANDREA FARINA E MICHELE CILETTI
Dati diacronici sui preverbi in italiano: un'estensione del dataset PREMOVE 9

STEFANO FIORI, MATTEO RIVOIRA, ALBERTO GHIA, BARBARA TURCHETTA E LORENZO FERRAROTTI
Per una valorizzazione del dato geolinguistico alla luce della digitalizzazione dell'Atlante Linguistico Italiano 13

GUGLIELMO INGLESE
Dati sincronici e dati diacronici in tipologia: la scala di spontaneità e i verbi labili 16

MARIA LAURENTI, SARAH BIGI, SIBILLA PARLATO, CHIARA MIDEA, ALESSANDRO CACCIATORE E RAMONA BONGELLI
PragMed-ONCO: Sfide teoriche e metodologiche nella creazione di un corpus di interazioni medico-paziente annotato pragmaticamente 19

ANNA MANTOVANI
Dal quaderno al corpus: costruzione e analisi del dato scritto nell'italiano L2 in contesti scolastici multilingui 22

JOHANNA MIECZNIKOWSKI ED ELENA BATTAGLIA
Il corpus TIGR: una risorsa multimodale per studiare l'italiano parlato 25

LAURA OCCHIPINTI
Misurare l'efficacia comunicativa: dati linguistici oggettivi e soggettivi nella leggibilità infantile 29

LUCREZIA SABIA
Corpora e dati sperimentali al servizio dell'antonimia: sviluppo teorico di una relazione semantica 32

FLORIANA CARLOTTA SCIUMBATA
Dalla parola al dato: metodi e problemi nella creazione di liste lessicali 35

MARTINA VERDELLI
Sfide nella raccolta dati di verbi sintagmatici nelle varietà bergamasche: metodologie a confronto 37

PAOLA VERNILLO
Dalla distinzione teorica alla variabile osservabile: la specificità lessicale nel giudizio metaforico 39

**ELISA DI DOMENICO**

## I *si* di Univers-ITA

La ricerca linguistica si avvale di dati di diversa natura per cogliere generalizzazioni, formulare teorie o spiegazioni di un fenomeno particolare. I *corpora* – opportunamente interrogabili – sono uno strumento prezioso per cogliere la diffusione e le varietà di realizzazione di un fenomeno specifico. Partendo da un'indagine effettuata sul *corpus* scritto Univers-ITA (Grandi et al. 2023), ci concentreremo sia sulle possibilità che sui limiti di uno studio effettuato su *corpora*. Due sono infatti, a nostro avviso, i problemi. Da un lato, se un fenomeno non è osservato in quel *corpus*, non è detto che non possa occorrere. Questo problema può essere in parte risolto con considerazioni statistico/ distribuzionali e con il confronto con altri tipi di dati. Il secondo problema (che riguarda anche i dati elicitati sperimentalmente) è che i dati non contengono in sé la loro spiegazione: il ricorso alla teoria (anche in fasi intermedie dell'analisi) è quindi imprescindibile per raggiungere un'adeguatezza esplicativa.

L'indagine da cui partiamo riguarda il pronome clitico *si* dell'italiano in frasi complesse con verbi 'a ristrutturazione', a confronto con altri clitici. Com'è noto in queste frasi un clitico può trovarsi in proclisi al verbo della principale o in enclisi al verbo dell'infinitiva (*lo voglio fare / voglio farlo*). Alcuni studi suggeriscono che l'opzionalità sia legata a variabili geografiche, di registro o del parlante, come la sua dilalia (e.g. Benincà 1986; Cardinaletti, Giusti & Lebani 2025; Rholfs 1967). Le caratteristiche d'interrogabilità di Univers-ITA hanno consentito di indagare su una nuova possibile variabile: il tipo di clitico.

La procedura di estrazione è stata in primo luogo automatica, e ha consentito l'identificazione di 19.642 clitici (di cui 10.123 *si*), suddivisi per macro-aree (Nord, Centro, Sud).

In una seconda fase (manuale-oggettiva) sono stati identificati i contesti a ristrutturazione in cui i clitici comparivano (2.618, di cui 1.756 con *si*), e la loro distribuzione in enclisi o proclisi. In generale si osserva una prevalenza della proclisi (1.557) sull' enclisi (1.051), uniforme nelle tre macro-aree. L'analisi per tipo di clitico rivela tuttavia che questa prevalenza riguarda solo *si* e *vi*: escludendoli, osserviamo una netta prevalenza dell'enclisi, diffusa nelle tre macro-aree.

Infine, in una terza fase (manuale-teoricamente motivata), sono stati distinti due tipi di *si*: IMP (nelle varianti impersonale, media e passiva) e REFL (nelle varianti riflessiva, inerente e benefattiva). IMP si ha l'interpretazione di una variabile libera, REFL si ha un'interpretazione anaforica (Manzini

1986). Per i clitici *vi* e *ci* abbiamo distinto il loro valore 'personale' e 'non personale' (locativo, esistenziale, etc.).

Il quadro che emerge suggerisce due generalizzazioni indipendenti:

(1) I clitici con un tratto di persona tendono a stare in enclisi; gli altri (cioè quelli riferentisi a luoghi, cose, eventi, parti, o con interpretazione anaforica) sono più liberi di scegliere enclisi o proclisi.

(2) a. IMP *si* occorre quasi sempre in proclisi (cfr. Berretta 1986).
b. Nei rari casi in cui occorre in enclisi (28/1.756) ha sempre un'interpretazione passiva.

Per ragioni di spazio non scenderemo nei dettagli della loro interpretazione, (1) spiegata nei termini di un principio di precocità di verifica del tratto di Persona, (2) (in particolare b.) come impossibilità per *si* di ricevere caso nominativo in enclisi.

In conclusione, l'analisi per tipo di clitico, resa possibile dalle caratteristiche di Univers-ITA, ha messo in luce variabili nuove nel fenomeno osservato. La numerosità dei dati analizzati contribuisce alla solidità delle generalizzazioni osservate. Alcune scelte teoricamente fondate sono state tuttavia fondamentali per portare a termine l'analisi. Un ultimo dato, infine, necessita di ulteriori osservazioni: la distribuzione di *si* in enclisi potrebbe essere più diffusa di quanto osservato in Univers-ITA.

**Riferimenti bibliografici**

Benincà, Paola. 1986. Punti di sintassi comparata dei dialetti italiani settentrionali. Holtus, Günter & Ringger, Kurt (a cura di), *Raetia Antiqua et Moderna: W. Theodor Elwert zum 80. Geburztag*. 457–479. Tübingen: Niemeyer.

Berretta, Monica. 1986. Struttura informativa e sintassi dei pronomi atoni: Condizioni che favoriscono la 'risalita'. Stammerjohan, Harro (a cura di), *Tema-Rema in italiano*. 71–83. Tübingen: Narr.

Cardinaletti, Anna, Giusti, Giuliana & Lebani, Gianluca. 2025. Clitic Climbing across Italy. Variation, optionality, and the role of bilectalism. *Linguistic Variation* 26(1). 76–96. https://doi.org/10.1075/lv.24065.car

Grandi, Nicola, Ballarè, Silvia, Chiusaroli, Francesca, Gallina, Francesca, Pascoli, Matteo & Pistolesi, Elena. 2023. *Corpus Univers-ITA 2023*, https://doi.org/10.60760/unibo/univers-ita

Manzini, Maria Rita. 1986. On Italian *Si*. Borer, Hagit (a cura di), *Syntax and Semantics 19: The syntax of pronominal clitics*, 241–262. New York: Academic Press.

Rohlfs, Gerhard. 1968. *Grammatica storica dell'italiano e dei suoi dialetti. Morfologia.* Torino: Einaudi.

**DARIA EVANGELISTA**

## Metodi di annotazione per l'analisi funzionale delle metafore in *ClimIta*, un corpus di testi scritti prodotti da media e istituzioni sul tema della crisi climatica

La metafora è un potente strumento concettuale capace di veicolare un punto di vista all'interno del testo e sorreggerne l'argomentazione (Evangelista 2025a). Nell'ultimo ventennio, linguisti (Lakoff 2010) ed ecologisti (Larson 2011) hanno evidenziato come l'uso delle metafore influenzi la percezione di fenomeni ambientali con ingenti conseguenze sociali come quello della crisi climatica. In particolare, viene sottolineata la capacità della metafora di reiterare l'*endoxa*, e in generale un punto di vista dominante, creando potenzialmente inerzia, precludendo lo spazio cognitivo che rende possibili punti di vista alternativi e smorzando la percezione dell'urgenza (Larson 2011, 18). Su questa base, e alla luce della svolta cognitiva della metafora (cfr. Lakoff & Johnson 1980), il contributo approfondisce i metodi con cui è possibile analizzare le metafore all'interno del discorso pubblico di sensibilizzazione sul clima, concentrandosi sull'annotazione del fenomeno all'interno di *ClimIta*, un corpus bilanciato di testi scritti in lingua italiana, tematicamente incentrato sulla crisi climatica (Evangelista 2026).

In particolare, verranno esplorati metodi di annotazione pertinenti a livelli linguistici differenti, con l'obiettivo di farli interagire in una tipologia delle funzioni pragmatiche della metafora nel discorso di sensibilizzazione sulla crisi climatica. Ci si soffermerà: i) sul piano lessicale, considerando le differenze tra le tipologie della metafora convenzionale e della metafora creativa (Prandi 2023), poiché esse veicolano due diversi gradi di implicitezza (Evangelista 2025b, 91) e vengono dunque utilizzati in contesti comunicativi differenti; ii) sul piano argomentativo, osservando i processi di *framing* (Entman 1993; Fillmore 1976) messi in atto dalle metafore presenti all'interno del corpus, in relazione agli argomenti da essi introdotti; iii) sul piano testuale, focalizzandosi sulla posizione che le metafore vengono ad occupare all'interno della gerarchia del testo, al livello macro, meso e microtestuale (Ferrari 2024, 78–79). Inoltre, verrà osservato in che modo le metafore contribuiscono alla polifonia, considerata nei suoi tre diversi piani: interazione discorsiva, discorso riportato e punto di vista (Ferrari, Lala & Pecorari 2024). Dall'interazione di queste categorie di annotazione verrà tratta, con metodologia *corpus driven* basata su *ClimIta*, una categorizzazione concernente le funzioni pragmatiche delle metafore nella tipologia di discorso analizzato.

I metodi di annotazione trattati sono rigorosamente manuali, e basati sul procedimento di individuazione della metafora per le unità lessicali, MIP (Pragglejaz Group 2007), che consiste nell'individuare il significato contestuale dell'unità lessicale per paragonarlo con le sue accezioni

canoniche. Il corpus di analisi *ClimIta* è ad oggi uno dei pochissimi corpora in italiano dedicato al discorso pubblico sulla crisi climatica, e comprende: i) le principali pagine web e campagne ambientali pubblicate da Ministeri e Dipartimenti dell'Ambiente, come pure dalle principali ONG ambientaliste, in Italia e in Svizzera; ii) articoli di giornale online tratti da importanti quotidiani generalisti e scientifici in Italia e in Svizzera; iii) i principali testi normativi internazionali, dell'Unione europea e nazionali (italiani e svizzeri) attualmente in vigore in materia di cambiamento climatico. Il corpus e la ricerca fanno parte del progetto biennale *The Awareness-Raising Discourse about the Climate Crisis. An Analysis of Written Texts in Italian at the Interplay between Rhetoric, Argumentation Theory and Text Linguistics*, finanziato dal Fondo nazionale svizzero per la ricerca scientifica (FNS, n. 230661).

**Riferimenti bibliografici**


Entman, Robert. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication* 43(4). 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Evangelista, Daria. 2025a. Metaphors in persuasion and argumentation: A theoretical systematisation at the intersection of rhetoric and textuality with examples from the climate crisis discourse. *Topoi. Special Issue "Theorizing Rhetoric: Bridging Concepts Across Multiple Perspectives, Different Practices and Contexts"*. https://doi.org/10.1007/s11245-025-10312-2

Evangelista, Daria. 2025b. *La legge ornata. Un'analisi dei testi normativi tra retorica e linguistica del testo*. Alessandria: Edizioni dell'Orso.

Evangelista, Daria. 2026. *ClimIta*: a corpus of media and institutional written communication in Italian on the climate crisis. Dataset. https://doi.org/10.5281/zenodo.18335890 .

Ferrari, Angela (a cura di). 2024. *Dizionario di linguistica del testo a uso delle scienze umane*. Roma: Carocci.

Ferrari, Angela, Lala, Letizia & Pecorari, Filippo. 2024. Il piano enunciativo-polifonico della strutturazione del testo scritto. Gli ambiti dell'interazione discorsiva, del riporto e del punto di vista. *Studi di grammatica italiana*. *A cura dell'Accademia della Crusca* XLIII, 231–256.

Fillmore, Charles. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences. Origin and development of language and speech*, 280(1). 20–32. https://doi.org/10.1111/j.1749-6632.1976.tb25467.x

Lakoff, George & Johnson, Mark. 1980. *Metaphors we live by*. Chicago/London: The University of Chicago Press.

Lakoff, George. 2010. Why it matters how we frame the environment. *Environmental Communication* 4(1). 70–78.

Larson, Brendon. 2011. *Metaphors for environmental sustainability. Redefining our relationship with nature.* New Haven: Yale University Press.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse, *Metaphor and Symbol* 22(1). 1–39.

Prandi, Michele. 2023. *Retorica. Una disciplina da rifondare*. Bologna: Il Mulino.

**ANDREA FARINA E MICHELE CILETTI**

## Dati diacronici sui preverbi in italiano: un'estensione del dataset PREMOVE

I preverbi, cioè prefissi che si combinano con basi verbali (*circum-* in *circumnavigare*), rappresentano un tratto strutturale centrale delle lingue indoeuropee (Booij & Van Kemenade 2003). Già nelle lingue classiche mostrano ampia variabilità semantica: a usi composizionali (*ab-eo* 'andare via') si affiancano casi di lessicalizzazione o grammaticalizzazione, in cui il significato non è più trasparente (*e-venio* 'accadere', *com-edo* 'divorare'; Cuzzolin 1995). Questa complessità continua anche nelle lingue moderne, in particolare in greco (Efthymiou 2001; 2018) e nelle lingue romanze (Araujo Silva 2009; García Sánchez 2018; Jekl 2020; Kopecka 2006), dove i preverbi restano centrali nell'organizzazione del lessico verbale.

In italiano, lo studio dei preverbi è ancora qualitativo, spesso focalizzato su singoli preverbi o verbi preverbati (Dell'Oro 2022; Iacobini & Masini 2007; Todaro 2019). Ciò ostacola analisi su larga scala, soprattutto in prospettiva diacronica, e uno studio sistematico dell'interazione semantica tra preverbo e base verbale. Servono quindi *dataset* strutturati, fondati su *framework* consolidati anche in ambito computazionale, che permettano analisi comparabili e replicabili. In questa direzione, come mostrano studi sulle lingue classiche (Farina & Ciletti *in stampa*), annotazioni (semi)automatiche tramite LLMs, finora inesplorate per l'italiano, possono offrire un valido supporto.

Questo contributo analizza in prospettiva diacronica la semantica dei verbi italiani con preverbo derivati da basi originariamente di movimento latine, tramite l'impiego di LLMs. L'attenzione si concentra sulla metodologia e sui risultati, interpretati anche qualitativamente e confrontati con il latino. La metodologia si basa sull'espansione del *dataset* PREMOVE (Farina 2025), concepito per lingue classiche ma progettato per l'estensione ad altre lingue grazie a linee guida di annotazione esplicitamente comparabili (Farina 2024). Il *dataset* originale si fonda su un corpus diacronico di 541.620 *token* (dal VI sec. a.C. al II sec. d.C.), rappresentativo di diversi generi in prosa e poesia. Analogamente, l'estensione all'italiano si basa su un subcorpus diacronico (XIII-XX sec. d.C.) tratto dal corpus MIDIA (D'Achille & Iacobini 2022), che include 173 tra testi poetici, narrativi e teatrali di 59 autori. A partire da un elenco di ventuno verbi italiani con preverbo di origine latina (lat. *adcurro* > it. 'accorrere'), sono state individuate ed estratte 1.299 occorrenze.

Un sottoinsieme bilanciato di 300 *token* è stato annotato manualmente e indipendentemente dagli autori secondo Farina (2024). Per ciascun *token* sono stati identificati (i) il significato del preverbo, tramite una traduzione inglese, e (ii) il significato del verbo, tramite un *synset* dell'*Open English WordNet* (McCrae et al. 2020). Su un campione di 100 occorrenze, gli autori hanno raggiunto

un *Inter-Annotator Agreement* perfetto per entrambi i *task*, a conferma dell'affidabilità dell'annotazione. Per l'annotazione automatica è stato impiegato il modello *DeepSeek-V3.2-Thinking*, tra i modelli *open-source* più performanti in *task* analoghi (Farina & Ciletti *in stampa*), con *few-shot prompting* (Brown et al. 2020). Sul campione di 100 occorrenze, il modello ha ottenuto una precisione dell'84% per il *task* (i) e dell'80% per il *task* (ii).

Un'analisi preliminare evidenzia differenze significative nella distribuzione dei preverbi tra latino e italiano, riconducibili anche al mutamento semantico dei verbi preverbati durante il passaggio dal latino all'italiano: *convenire* rappresenta il 48,8% delle occorrenze italiane, dato l'ampio uso della forma impersonale *conviene* nel senso di 'è vantaggioso', mentre il lat. *convenio* 'riunirsi' costituisce solo il 5% dei *token* di PREMOVE. Inoltre, nella maggior parte dei casi, i preverbi in italiano hanno perso il loro valore lessicale originario e risultano ampiamente lessicalizzati, con alcune eccezioni (es. *pre-*, *circum-*), che saranno discusse qualitativamente nel contributo.

**Riferimenti bibliografici**


Araujo Silva, Ana Paula. 2009. *Prefixos, prefixóides ou elementos de composição? Uma análise crítico-comparativa dos prefixos registrados no* Novo dicionário Aurélio da língua portuguesa *e no* Dicionário Houaiss da língua portuguesa. Rio de Janeiro: Universidade do Estado do Rio de Janeiro. (Tesi di dottorato).

Booij, Geert & Van Kemenade, Ans. 2003. Preverbs: An introduction. In Booij, Geert & Van Marle, Jap (a cura di), *Yearbook of Morphology 2003 (Yearbook of Morphology)*, 1–12. Dordrecht: Kluwer. https://doi.org/10.1007/978-1-4020-1513-7_1.

Brown, Tom B., Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared, Dhariwal, Prafulla, Neelakantan, Arvind, Shyam, Pranav, Sastry, Girish, Askell, Amanda, Agarwal, Sandhini, Herbert-Voss, Ariel, Krueger, Gretchen, Henighan, Tom, Rewon, Child, Aditya, Ramesh, Ziegler, Daniel M., Wu, Jeffrey, Winter, Clemens, Hesse, Christopher, Chen, Mark, Sigler, Eric, Litwin, Mateusz, Gray, Scott, Chess, Benjamin, Clark, Jack, Berner, Christopher, McCandlish, Sam, Radford, Alec, Sutskever, Ilya & Amodei, Dario. 2020. Language models are few-shot learners. In Larochelle, Hugo, Ranzato, Marc'Aurelio, Hadsell, Raia Thais, Balcan, Maria-Florina & Lin, Hsuan-Tien Lin (a cura di), *Advances in neural information processing systems* 33, *Proceedings of the 34th International Conference on Neural Information Processing Systems*, *Online, 6-12 dicembre 2020*, 1877–1901. San Diego: NeurIPS - Neural Information Processing Systems Foundation, Inc.

Cuzzolin, Pierluigi. 1995. A proposito di sub vos placo e della grammaticalizzazione delle preposizioni. *Archivio Glottologico Italiano* 80 (1–2). 12–143.

D'Achille, Paolo & Iacobini, Claudio. 2022. Il corpus MIDIA: Concezione, realizzazione, impieghi. In Cresti, Emanuela & Moneglia, Massimo (a cura di), *Corpora e Studi Linguistici. Atti del LIV Congresso Internazionale di Studi della Società di Linguistica Italiana*, *Online, 8-10 settembre 2021*, 207–221. Milano: Officinaventuno. https://doi.org/10.17469/O2106SLI000014

Dell'Oro, Francesca. 2022. Verbs of motion and intermediate source domains of modality: The understudied case of Italian occorrere 'to be necessary, to be needed'. *Folia Linguistica* 56(43–1). 65–95. https://doi.org/10.1515/flin-2022-2017.

Efthymiou, Angeliki (Ευθυμίου, Αγγελική). 2001. The Modern Greek prefix *kse-*: The concepts of move away and change of state [Το νεοελληνικό πρόθημα *ξε-*: οι έννοιες της απομάκρυνσης και της αλλαγής κατάστασης]. *Studies in Greek Linguistics* [*Μελέτες για την ελληνική γλώσσα*] 21. 202–213.

Efthymiou, Angeliki (Ευθυμίου, Αγγελική). 2018. *Verb derivation in Modern Greek* [*Ο σχηματισμός των ρημάτων στη νέα ελληνική*]. Thessaloniki: Epikentro [Επίκεντρο].

Farina, Andrea. 2024. *Guidelines for a linguistic annotation of preverbed verbs of motion.* London: King's College London. https://doi.org/10.18742/25055573.

Farina, Andrea. 2025. PREMOVE – A diachronic dataset of Ancient Greek and Latin annotated PREverbed MOtion VErbs. *Oxford Text Archive*. Dataset. http://hdl.handle.net/20.500.14106/2579.

Farina, Andrea & Ciletti, Michele. *In stampa*. Probing preverbs: Evaluating Large Language Models on Latin and Ancient Greek preverbed motion verbs. *Proceedings of Historical Languages and AI, Berlin, 5-6 marzo 2026.*

García Sánchez, Jairo J. 2018. El preverbio *ex-*: Morfología léxica, continuidad y variación del latín al español. *Estudios de lingüística del español* 39. 31–53.

Iacobini, Claudio & Masini, Francesca. 2007. Verb-particle constructions and prefixed verbs in Italian: Typology, diachrony and semantics. In Booij, Geert, Ducceschi, Luca, Fradin, Bernard, Guevara, Emiliano, Ralli, Angeliki & Scalise, Sergio (a cura di), *On-line proceedings of the Mediterranean Morphology Meetings (MMM5)*, *Fréjus, 15-18 settembre 2005*, 157–184. Bologna: Università degli Studi di Bologna,

Jekl, Ágnes. 2020. I prefissi verbali dal latino classico all'italiano: problemi di classificazione su base etimologica. Il prefisso *ad-*. *Acta Antiqua Academiae Scientiarum Hungaricae* 59(1–4). 377–386. https://doi.org/10.1556/068.2019.59.1-4.33.

Kopecka, Anetta. 2006. The semantic structure of motion verbs in French: Typological perspectives. In Hickmann, Maya & Robert, Stéphane (a cura di), *Space in Languages Linguistic Systems and Cognitive Categories (Typological Studies in Language, vol. 66)*, 83–102. Amsterdam/Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/tsl.66.06kop.

McCrae, John P., Rademaker, Alexandre, Rudnicka, Ewa & Bond, Francis. 2020. English WordNet 2020: Improving and extending a WordNet for English using an open-source methodology. In Declerk, Thierry, Gonzalez-Dios, Itziar & Rigau, German (a cura di), *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020), Marseille, 11-16 maggio2020*, 14–19. Paris: European Language Resources Association (ELRA). https://aclanthology.org/2020.mmw-1.3.pdf

Todaro, Giuseppina. 2017. *Nomi (e aggettivi) che diventano verbi tramite prefissazione: quel che resta della parasintesi.* Toulouse/Roma: Université Toulouse le Mirail-Toulouse II/Università degli studi Roma Tre. (Tesi di dottorato.)

**STEFANO FIORI, MATTEO RIVOIRA, ALBERTO GHIA, BARBARA TURCHETTA E LORENZO FERRAROTTI**

## Per una valorizzazione del dato geolinguistico alla luce della digitalizzazione dell'Atlante Linguistico Italiano

La comunicazione intende sviluppare alcune riflessioni metodologiche nate nel corso della digitalizzazione dell'Atlante Linguistico Italiano (ALI), intrapresa grazie al progetto PNRR CHANGES "DigitALI". Dal punto di vista tecnico, la disponibilità di file digitali preparati dalla redazione per la stampa dei volumi cartacei (Rivoira 2016) ha consentito di evitare una laboriosa acquisizione dei dati e di passare direttamente alla fase di "retro-standardizzazione" dell'archivio (Cerruti et al. 2025). Un nodo cruciale che il contributo vuole affrontare è che questo passo non consiste in un semplice cambio di supporto o di manipolazione strutturale, dalla quale i dati passano relativamente indenni, ma conduce inevitabilmente a riflettere sulla natura dei dati stessi. Nella nostra esperienza, questo ha riguardato almeno i seguenti ambiti:

1) I metadati e le marche metalinguistiche. I verbali delle inchieste (Massobrio et al. 1995) contengono preziose informazioni sul contesto del rilievo dialettale, sulla località visitata e sul profilo socio-biografico degli informatori; rappresentano dunque i "metadati" delle risposte dialettali. A queste, spesso gli informatori stessi aggiungono osservazioni "metalinguistiche", riportate in corsivo nelle carte, che forniscono indicazioni sullo stato variazionale della forma linguistica (ad esempio, "*v*" per "forma usata dai vecchi"). L'integrazione di questi elementi rappresenta un tratto qualificante del progetto DigitALI, il cui scopo è suggerire una lettura "profonda" del dato atlantistico, spesso accolto, nonostante l'impostazione originaria, senza riserve e senza adeguatamente problematizzarne e storicizzarne i dati.
2) Il concetto di "punto (geo)linguistico". I punti cartografati nei volumi rappresentano spesso più di una località d'inchiesta, e d'altra parte una stessa inchiesta può comprendere coordinate spaziotemporali diverse. Per questo si è deciso che l'entità più sicuramente localizzabile, alla quale attribuire delle coordinate geografiche, è il parlante stesso. Anche questo cambio di prospettiva, come la scelta di valorizzare le marche metalinguistiche, va nella direzione di portare l'informatore al centro dell'analisi; inoltre, nel caso di inchieste che coinvolgono diverse località vicine può rivelare più chiaramente dinamiche di microvariazione.
3) La trascrizione fonetica. L'alfabeto fonetico dell'ALI è molto complesso, concepito per consentire al raccoglitore di registrare impressionisticamente le più minute sfumature del parlato degli informatori. La sua traduzione in sistemi di largo uso come l'IPA è un passaggio cruciale per favorire l'interoperabilità dei dati ALI con altre banche dati, ma è irto di difficoltà sia tecniche

che teoriche, dal momento che l'IPA è concepito anche per fornire un modello universalmente valido di rappresentazione delle opposizioni fonematiche pertinenti nelle lingue del mondo (Moran & Cysouw 2018, 38-40), il che contrasta con l'operato dei raccoglitori dell'ALI. La sfida è dunque giungere a un compromesso tra il massimo di aderenza al dettaglio fonetico delle trascrizioni ALI e il minimo di soluzioni ad hoc da adottare per piegare l'IPA a questo scopo.

Illustrando e discutendo le strategie adottate per far fronte a questi problemi, vorremmo dunque proporre alcuni spunti per una gestione il più possibile coordinata dei dati geolinguistici digitali, in un'ottica di progressiva FAIR-ificazione (Krefeld & Lücke 2023). Ciò è particolarmente necessario in questo periodo ricco di progetti di digitalizzazione e di federazione di risorse, che favorisce un approccio agli atlanti linguistici come veri e propri *corpora* (Baiwir & Renders 2013, Julià & Rost 2025 & Pustka 2023) e apre a prospettive nuove di ricerca e analisi, non da ultimo sul versante quantitativo (Rubehn, Montemagni & Nerbonne 2024).

**Riferimenti bibliografici**


ALI = Bartoli, Matteo, Vidossi Giuseppe, Grassi Corrado, Genre Arturo, Massobrio Lorenzo, Rivoira Matteo (direttori), *Atlante Linguistico Italiano*. Voll. I-VIII, Roma, Istituto Poligrafico e Zecca dello Stato, 1995-2011. Vol. IX-X, Torino: Istituto dell'Atlante Linguistico Italiano, 2018-.

Baiwir, Esther & Renders, Pascale. 2013. Les atlas linguistiques sont-ils des corpus?. *Corpus* 12. 27–37. https://doi.org/10.4000/corpus.2301

Cerruti, Massimo, Ferrarotti, Lorenzo, Fiori, Stefano, Rivoira, Matteo & Turchetta, Barbara. 2025. La digitalizzazione dell'Atlante Linguistico Italiano (ALI): problemi e prospettive. *Linguistik Online* 136(4). 37–67. https://doi.org/10.13092/v3yehd21

Krefeld, Thomas & Lücke, Stephan. 2023. FAIRNESS – Medien im methodologischen Zentrum der Geolinguistik. Version 3 (22.06.2023). In Krefeld, Thomas, Lücke, Stephan & Mutter, Christina (a cura di), *Berichte aus der digitalen Geolinguistik (II): Vernetzung und Nach-haltigkeit*. (Korpus im Text 9), Version 30. https://www.kit.gwi.uni-muenchen.de/?p=45335&v=3

Julià Luna, Carolina & Rost Bagudanch, Assumpciò (a cura di). 2025. *Geografía lingüística y corpus en las lenguas románicas*. Berlin/Boston: De Gruyter. https://doi.org/10.1515/9783111439181.

Massobrio, Lorenzo, Ronco, Giovanni, Nosengo, Maria Carla & Tuninetti, Germana. 1995. *ALI. Verbali delle inchieste*. 2 voll. Roma: Istituto Poligrafico e Zecca dello Stato.

Moran, Steven & Cysouw, Michael. 2018. *The Unicode Cookbook for Linguists.* Berlin: Language Science Press.

Pustka, Elissa, Quijada Van de Berghe, Carmen & Weiland, Verena (a cura di). 2023. *Corpus Dialectology.* Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/scl.110?locatt=mode:legacy.

Rivoira, Matteo. 2016. L'archivio dell'ALI: informatizzazione dei dati e prospettive di sviluppo. In Avolio, Francesco & Cigno, Antonino (a cura di), *Archivi Etnolinguistici Multimediali dalla formazione alla gestione e al dialogo col territorio, Atti del Convegno, Pescara, 6 ottobre 2012*. 49–61. Pescara: Museo delle genti d'Abruzzo.

Rubehn, Arne, Montemagni, Simonetta & Nerbonne, John. 2024. Extracting Tuscan phonetic correspondences from dialect pronunciations automatically. *Language Dynamics and Change* 14(1). 1–33. https://doi.org/10.1163/22105832-bja10034.

**GUGLIELMO INGLESE**

## Dati sincronici e dati diacronici in tipologia: la scala di spontaneità e i verbi labili

Questo contributo indaga l'interazione tra dati sincronici e dati diacronici in tipologia, prendendo in esame il caso della distribuzione dei verbi labili.

La labilità, cioè l'uso della stessa forma verbale in strutture transitive e intransitive (es. *aumentare*), è una delle strategie impiegate dalle lingue del mondo per codificare l'alternanza causale-noncausale (Creissels 2024; Letuchiy 2009; Tubino-Blanco 2020). Dal punto di vista sincronico, la distribuzione della labilità può essere spiegata facendo ricorso alla scala di spontaneità. La scala di spontaneità è una gerarchia tipologica formulata da Haspelmath (1993, 2016) per descrivere la distribuzione interlinguistica delle costruzioni causative e anticausative. Sulla base dell'osservazione del comportamento di 31 significati in un campione di 21 lingue, Haspelmath (1993) mostra come questi significati possano essere organizzati in una scala che va da quelli più propensi alla causativizzazione (es. *bollire,* it. 'far bollire') a quelli più propensi all'anticausativizzazione (es. *rompere*, it. 'rompersi'). Secondo Haspelmath, questa distribuzione si può spiegare facendo ricorso alla nozione di spontaneità: in base a un principio di ottimizzazione, gli eventi che più probabilmente accadono spontaneamente richiedono una marca esplicita quando usati in contesti che coinvolgono un cambiamento di stato indotto, e viceversa.

La stessa scala può anche spiegare la distribuzione dei verbi labili. Contrariamente ai risultati di Letuchiy (2010), se consideriamo una più ampia base di dati, che include i dati dal *World Atlas of Transitivity Pairs* (WATP)[1], bilanciati su base genealogica e integrati con dataset aggiuntivi (= 68 lingue per 23 famiglie), possiamo osservare una distribuzione sbilanciata dei lessemi labili. In particolare, se dividiamo i 31 significati analizzati da Haspelmath (1993) in due classi (1–15 vs. 16–31), notiamo come i verbi labili sembrino concentrarsi nella porzione inferiore della scala, in corrispondenza ai lessemi meno spontanei, es. ROMPERE e APRIRE (cfr. Tab. 1).

| Classe verbale | Numero di verbi labili (media) | Numero di verbi labili |
|---|---|---|
| Spontanei (1-15) | 7,33 | 110 |
| Non spontanei (16-31) | 10,53 | 158 |

Tabella 1. I verbi labili e la scala di spontaneità

[1] World Altlas of Transitivity Pairs. https://watp.ninjal.ac.jp/en/

Se osservati da una prospettiva sincronica, i dati tipologici suggeriscono quindi come la scala di spontaneità possa spiegare, almeno in parte, la distribuzione del pattern labile. Tuttavia, negli ultimi decenni, gli studi di tipologia diacronica, in particolare la tipologia *source-oriented* (Cristofaro 2019; Cristofaro & Zúñiga 2018), hanno messo in luce come le motivazioni addotte per spiegare la configurazione sincronica delle gerarchie tipologiche siano spesso incompatibili con ciò che sappiamo riguardo a come le distribuzioni descritte dalle gerarchie effettivamente nascono e si sviluppano nel tempo. Anche nel caso della labilità, per quanto i dati siano ancora disponibili per un numero esiguo di lingue (Kulikov 2021), le dinamiche diacroniche che portano alla nascita dei verbi labili non sembrano supportare direttamente la scala di spontaneità. I verbi labili, infatti, emergono tramite diversi tipi di mutamento linguistico, che includono il mutamento fonologico e noti meccanismi di mutamento sintattico (Harris & Campbell 1995), come la reinterpretazione di strutture argomentali, l'estensione analogica e il contatto (cfr. Inglese 2026; Kehayov & Vihman 2014). Questi processi, tuttavia, operano indipendentemente dalla semantica del verbo: ne segue che la spontaneità non svolge un ruolo centrale nella diacronia della labilità.

Riassumendo, il contributo riafferma la centralità del dato empirico nella tipologia linguistica. In primo luogo, la disponibilità di ampie basi di dati è fondamentale per confermare o rivalutare generalizzazioni interlinguistiche. Dall'altro, si evidenzia come dati di natura diversa (sincronici vs. diacronici) non sempre portino alle stesse conclusioni. Risulta quindi necessario, in sede di analisi, bilanciare dati diversi, e talvolta conflittuali, per raggiungere una migliore comprensione dei fenomeni linguistici.

**Riferimenti bibliografici**

Creissels, Denis. 2024. *Transitivity, valency and voice*. Oxford: Oxford University Press.

Cristofaro, Sonia. 2019. Taking diachronic evidence seriously: Result-oriented vs. Source-oriented explanations of typological universals. In Schmidtke-Bode, Karsten, Levshina, Natalia, Michaelis, Susanne Maria & Seržant, Ilja A. (a cura di), *Explanation in typology*, 25–46. Berlin: Language Science Press.

Cristofaro, Sonia & Zúñiga, Fernando. 2018. Synchronic vs. diachronic approaches to typological hierarchies. In Cristofaro, Sonia & Zúñiga, Fernando (a cura di), *Typological Hierarchies in Synchrony and Diachrony*, 3–30. Amsterdam/Philadelphia: John Benjamins.

Harris, Alice C. & Campbell, Lyle. 1995. *Historical Syntax in Cross-Linguistic Perspective*. Cambridge: Cambridge University Press.

Haspelmath, Martin. 1993. More on the typology of the inchoative/causative verb alternations. In Comrie, Bernard & Polinsky, Maria (a cura di), *Causatives and Transitivity*, 87–120. Amsterdam/Philadelphia: John Benjamins.

Haspelmath, Martin. 2016. Universals of causative and anticausative verb formation and the spontaneity scale. *Lingua Posnaniensis* 58(2). 33–63. https://doi.org/10.1515/linpo-2016-0009

Inglese, Guglielmo. 2026. Lability in Hittite and Indo-European: a diachronic perspective. *Studia Linguistica* 80:e70010. https://doi.org/10.1111/stul.70010

Kehayov, Petar & Vihman, Virve. 2014. The Lure of Lability: A Synchronic and Diachronic Investigation of the Labile Pattern in Estonian. *Linguistics* 52(4). 1061–1105.

Kulikov, Leonid. 2021. *Les verbes labiles dans l'histoire de la famille indo-européenne: une perspective typologique diachronique*. In Lauwers, Peter, Paykin, Katia, Ilioaia, Mihaela, Meulleman, Machteld & Hadermann, Pascale (a cura di), *Quand le syntagme nominal prend ses marques: du prédicat à l'argument*, 191–216. Reims: EPURE.

Letuchiy, Alexander. 2009. Towards a typology of labile verbs: lability vs. derivation. In Epps, Patience & Arkhipov, Alexandre (a cura di), *New Challenges in Typology: Transcending the Borders and Refining the Distinctions*, 223–244. Berlin/New York: de Gruyter.

Letuchiy, Alexander. 2010. Lability and spontaneity. In Brandt, Patrick & García García, Marco (a cura di), *Transitivity: Form, Meaning, Acquisition, and Processing*, 237–256. Amsterdam/Philadelphia: John Benjamins.

Tubino-Blanco, Mercedes. 2020. Causative-Inchoative in Morphology. In Aronoff, Mark (a cura di), *Oxford Research Encyclopedia of Linguistics*. Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.595

**MARIA LAURENTI, SARAH BIGI, SIBILLA PARLATO, CHIARA MIDEA, ALESSANDRO CACCIATORE E RAMONA BONGELLI**

## PragMed-ONCO: Sfide teoriche e metodologiche nella creazione di un corpus di interazioni medico-paziente annotato pragmaticamente

Il contributo che presentiamo si inserisce nell'ambito della linguistica medica, intesa come "un filone della ricerca che comprende la tradizionale attenzione della linguistica applicata al linguaggio medico nel contesto delle cosiddette lingue per scopi speciali o linguaggi settoriali, che non si limita tuttavia agli studi sul lessico e sulla terminologia medica, ma considera anche il discorso medico e dell'informazione medica nel suo complesso (sia scritto che orale), la comunicazione medico-paziente e, più in generale, il rapporto paziente-operatore [...]." (Orletti, Cardinaletti & Dovetto 2015, 393).

La linguistica medica si è potuta sviluppare e ha preso forza grazie alla disponibilità di dati da conversazioni reali nei contesti di cura, fattore che è anche quello più sfidante per lo studioso: è noto, infatti, che una delle maggiori difficoltà per il linguista che si occupi di interazioni nei contesti di cura stia proprio nella raccolta, archiviazione, trattamento e condivisione dei dati di analisi.

In questo contributo vorremmo offrire risultati iniziali e considerazioni metodologiche relative a un lavoro in corso, finalizzato alla creazione di un corpus di trascrizioni di visite mediche annotate pragmaticamente, raccolte in un setting oncologico. Il lavoro si inserisce tra gli obiettivi del progetto PRIN 2022 intitolato, *Effective communication between doctors and patients in oncological settings* (PI e prima unità di ricerca, Ramona Bongelli, Università di Macerata; seconda unità di ricerca, Sarah Bigi, Università Cattolica del Sacro Cuore, Milano). Il lavoro preparatorio alla creazione del corpus che presentiamo è iniziato nel maggio 2025 e si concluderà in ottobre 2026.

La creazione di un corpus annotato pragmaticamente pone alcune sfide sul piano metodologico e interpretativo. Da una parte è necessario costruire un sistema di annotazione in grado di rappresentare la multifunzionalità pragmatica degli enunciati e la multidimensionalità del dato linguistico, dall'altra è fondamentale interpretare in maniera univoca gli atti e le mosse dialogiche dell'interazione. Nel corso degli anni ci sono stati vari tentativi di creazione di sistemi annotativi in grado di coprire la grande mole di atti e di funzioni dialogiche definite dalla teoria linguistica, citiamo tra gli altri DAMSL – *Dialogue Act Markup in Several Layers* e DiAML – *Dialogue Act Markup Language*, ISO 24617-2 (Allen & Core 1997; Bunt et al. 2020); tuttavia ad oggi non sembra che si sia giunti alla creazione di uno schema standardizzato e adattabile a ogni contesto di interazione linguistica.

A partire da queste premesse, abbiamo strutturato uno schema di annotazione multilivello e semi-standardizzato, adattato dal DiAML – ISO 24617-2 (Bunt et al. 2012, 2017, 2020), del quale si riprendono le funzioni comunicative (per esempio, *offer, request, inform*) a partire dalle macro-classi funzionali, definite in base allo scopo comunicativo (quali *Information-seeking functions, Commissive functions, Directive functions*). Utilizzando questo schema e sulla base della dinamica interazionale presente nei nostri dati, abbiamo scelto di annotare le strutture interrogative, le repliche e le strategie di *repair* (es. *self-initiated self-repair*, *other-initiated self-repair*) (Kitzinger 2012; Sacks, Schegloff & Jefferson 1977).

Prevediamo di adottare un processo di annotazione semi-automatico: a una prima fase manuale, seguirà una seconda automatica, utilizzando la piattaforma INCEpTION (Klie et al. 2018) e operando interamente in locale per salvaguardare i dati in possesso, altamente sensibili. Un'ultima fase di revisione manuale permetterà di controllare quanto ottenuto automaticamente.

Alla presentazione dei dati e dello schema di annotazione, seguiranno alcune considerazioni in merito alle sfide metodologiche poste dai contesti di cura: in particolare riguardo alla selezione, raccolta, strutturazione, conservazione e riutilizzo dei dati.

**Riferimenti bibliografici**


Allen, James & Core, Mark. 1997. *DAMSL: Dialogue Act Markup in Several Layers* (Draft 2.1). http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual

Bunt, Harry, Alexandersson, Jan, Jae-Woong, Choe, Fang, Alex C., Koiti, Hasida, Petukhova, Volha, Popescu-Belis, Andrei, Traum & David R. 2012. ISO 24617-2: A Semantically-Based Standard for Dialogue Annotation. In Calzolari, Nicoletta, Choukri, Khalid, Declerck, Thierry, Uğur, Doğan, Mehmet, Maegaard, Bente, Mariani, Joseph, Moreno, Asuncion, Odijk, Jan & Piperidis, Stelios (a cura di), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, *Istanbul, 21-27 maggio 2012,* 430–437. Paris: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/530_Paper.pdf

Bunt, Harry, Petukhova, Volha, Traum, David R. & Alexandersson, Jan. 2017. Dialogue act annotation with the ISO 24617-2 standard. In Dahl, Deborah A. (a cura di), *Multimodal interaction with W3C standards*, 109–135. Cham: Springer. https://doi.org/10.1007/978-3-319-42816-1_6

Bunt, Harry, Petukhova, Volha, Gilmartin, Emer, Pelachaud, Catherine, Fang, Alex, Keizer, Simon & Prévot, Laurent. 2020. The ISO Standard for Dialogue Act Annotation, Second Edition. In Calzolari, Nicoletta, Béchet, Frédéric, Blache, Philippe, Choukri, Khalid, Cieri, Christopher, Declerck, Thierry, Goggi, Sara, Isahara, Hitoshi, Maegaard, Bente, Mariani, Joseph, Mazo, Hélène, Moreno, Asuncion, Odijk, Jan & Piperidis, Stelios (a cura di), *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, *Marseille, 11-16 maggio 2020,* 549–558. Paris: European Language Resources Association (ELRA). https://aclanthology.org/2020.lrec-1.69

Kitzinger, Celia. 2012. Repair. In Sidnell, Jack & Stivers, Tanya (a cura di), *The Handbook of Conversation Analysis*. 229–256. Chichester, UK: Wiley-Blackwell. https://doi.org/10.1002/9781118325001.ch12

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 di), *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018): System Demonstrations*, *Santa Fe (New Mexico), 20-26 agosto 2018,* 5–9. Stroudsburg: Association for Computational Linguistics. https://aclanthology.org/C18-2002

Orletti, Franca, Cardinaletti, Anna & Dovetto, Francesca. 2015. Tra linguistica medica e linguistica clinica. Il ruolo del linguista. *Studi Italiani di Linguistica Teorica e Applicata* 44(3). 392–394.

Sacks, Harvey, Schegloff, Emanuel A. & Jefferson, Gail. 1977. The Preference for self-correction in the organization of repair in conversation. *Language* 53(2). 361–382. https://doi.org/10.2307/413107

**ANNA MANTOVANI**

## Dal quaderno al corpus: costruzione e analisi del dato scritto nell'italiano L2 in contesti scolastici multilingui

Il contributo propone una riflessione epistemologica sullo statuto del dato scritto nella ricerca sull'italiano L2 nelle classi multilingui della scuola secondaria di primo grado (cfr. Fiorentini, Gianollo & Grandi 2020). Il lavoro mette in dialogo le produzioni scritte con l'osservazione didattica e l'analisi computazionale, integrando il dato scritto in L2 e la modellizzazione esplorativa attraverso il linguaggio di programmazione Python e librerie *open-source* (Pedregosa et al. 2011), in una prospettiva di linguistica acquisizionale (Vedovelli & Villarini 2003).

Due domande di ricerca orientano il contributo:

(i) Quali aspetti dell'interlingua degli apprendenti di livello elementare (A1-A2) risultano maggiormente analizzabili a partire dai dati scritti rispetto ad altri canali diamesici?

(ii) In che modo la tipologia di task incide sulla natura, sulla controllabilità e sull'interpretabilità del dato scritto nell'italiano per lo studio?

La ricerca ha coinvolto 262 discenti adolescenti (età media = 14 anni e mezzo; anni di permanenza media in Italia = 3 anni e mezzo) alloglotti di livello QCER A1-A2, portatori di 38 L1, provenienti da 3 Centri Provinciali per l'Istruzione degli Adulti del Veneto e da 9 scuole medie delle regioni Veneto (n = 7), Emilia-Romagna (n = 1) e Lombardia (n = 1).

Il corpus multilingue è stato manualmente annotato e analizzato da chi scrive. La raccolta dati ha incluso 150 ore di osservazione e somministrazione di 16 Unità di Apprendimento (UdA) progettate in percorsi CLIL con pratiche *translanguaging* implementate per l'italiano per lo studio come lingua veicolare (Santipolo 2021), nelle discipline di antologia, storia, geografia, scienze, matematica e geometria.

All'interno di ciascuna UdA sono stati impiegati cinque macro-tipi di task finalizzati a elicitare l'interlingua e i microlinguaggi disciplinari: (a) esercizi a completamento e cloze (produzione controllata); (b) domande a risposta breve (produzione guidata); (c) scelte binarie (vero/falso; sì/no; selezione tra alternative); (d) produzioni scritte semi-guidate; (e) attività di abbinamento.

L'analisi privilegia la dimensione scritta assumendola come variabile diamesica metodologicamente rilevante: il dato scritto offre stabilità, conservabilità e possibilità di annotazione sistematica e replicabile, consentendo un'analisi fine dei fenomeni morfosintattici e una maggiore controllabilità comparativa (Solier et al. 2019). Tuttavia, esso implica una maggiore monitorizzazione

metalinguistica e può riflettere processi pianificati, diversi dalla produzione spontanea, oltre a escludere dimensioni prosodiche, interazionali e riparative tipiche dell'oralità, restituendo un'interlingua mediata dal canale grafico.

Le produzioni raccolte si collocano in una zona intermedia tra dati naturali ed ecologici e dati controllati, poiché generati all'interno di pratiche didattiche autentiche osservate in presenza, ma al contempo orientate da dispositivi di elicitazione strutturati (Lenci, Montemagni & Pirrelli 2016). La varietà tipologica dei task consente di osservare l'interlingua lungo un continuum che spazia dalla produzione controllata a quella libera, includendo manifestazioni di competenza lessicale e strategie di categorizzazione. In tale prospettiva, particolare rilievo assume la presenza di pratiche *translanguaging* nella produzione scritta, come dimostra l'esempio (1) tratto dall'output di una studentessa L1 Bini (partecipante n. 87):

(1) *Pounded yam, è Egusy soup fried rice è pollo (chicken)*
Pounded yam COP.3SG Egusi soup fried rice COP.3SG chicken (chicken)
*Black soup*
black soup
'Pounded yam, zuppa Egusi, riso fritto con pollo, zuppa nera e Poundo'

L'esempio evidenzia fenomeni di alternanza e uso integrato e dinamico di più lingue per organizzare e mediare i processi cognitivi, mostrando come il *translanguaging* contribuisca alla costruzione del significato anche nel canale scritto, entro una prospettiva di multilinguismo ed eteroglossia (García & Lin 2017). Il contributo problematizza infine la natura del dato linguistico distinguendo tra dati grezzi, osservabili e annotabili; dati interpretativi inferiti (strategie individuali, traiettorie interlinguistiche, repertori plurilingui); dati trattati, risultanti da processi di trascrizione, normalizzazione e annotazione mediante un sistema di tagging; e metadati (L1, età, tempo di esposizione alla L2).

**Riferimenti bibliografici**

Fiorentini, Ilaria, Gianollo, Chiara & Grandi, Nicola. 2020. La classe plurilingue in Italia: sfide e opportunità. In Fiorentini, Ilaria, Gianollo, Chiara & Grandi, Nicola (a cura di), *La classe plurilingue*, 9–19. Bologna: Bononia University Press.

García, Ofelia & Lin, Mei Yi Angel. 2017. Translanguaging in Bilingual Education. In García, Ofelia & Lin, Mei Yi Angel (a cura di), *Bilingual and Multilingual Education. Encyclopedia of Language and Education*, 117–130. Cham: Springer International Publishing.

Lenci, Alessandro, Montemagni, Simonetta & Pirrelli, Vito. 2016. *Testo e computer*, 4ª ed. Roma: Carocci Editore.

Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Müller, Andreas, Nothman, Joel, Louppe, Gilles, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, Vanderplas, Jake, Passos, Alexandre, Cournapeau, David, Brucher, Matthieu, Perrot, Matthieu & Duchesnay, Édouard. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12. 2825–2830.

Santipolo, Matteo. 2021. The Content of CLIL and Beyond: Classifying Typologies of Microlanguages for Teaching and Learning Purposes. In Graziano, Alba, Turchetta, Barbara, Benedetti, Fausto & Cinganotto, Letizia (a cura di), *Pedagogical and Technological Innovations in (and through) Content and Language Integrated Learning*, 18–28. Cambridge: Cambridge Scholars Publishing.

Solier, Clara, Perret, Cyril, Baqué, Lorraine & Soum-Favaro, Christiane. 2019. Written Training Tasks Are Better than Oral Training Tasks at Improving L2 Learners' Speech Production. *Applied Psycholinguistics* 40(6). 1455–1480. https://doi.org/10.1017/S014271641900033X

Vedovelli, Massimo & Villarini, Andrea. 2003. Dalla linguistica acquisizionale alla didattica acquisizionale: le sequenze sintattiche nei materiali per l'italiano L2 destinati agli immigrati stranieri. In Giacalone Ramat, Anna (a cura di), *Verso l'italiano. Percorsi di acquisizione*, 270–304. Roma: Carocci Editore.

**JOHANNA MIECZNIKOWSKI ED ELENA BATTAGLIA**

## Il corpus TIGR: una risorsa multimodale per studiare l'italiano parlato

Il corpus TIGR (Miecznikowski, Battaglia & Geddo *in fase di valutazione*) è stato raccolto nei cantoni svizzeri Ticino e Grigioni negli anni 2021-2022. Comprende 23 eventi videoregistrati in contesti quotidiani e istituzionali (conversazioni a tavola, preparazione di cibo, interviste, lezioni teoriche e pratiche) e trascritti in ELAN, per un totale di 23,5 ore / 223.076 parole, con un apparato di metadati sociolinguistici relativi ai parlanti. In questo contributo presentiamo la risorsa, che rappresenta una novità nel panorama italiano dove i principali corpora accessibili di parlato sono audioregistrati (es. LIP, De Mauro et al. 1993; CLIPS, Albano Leoni 2007; KIParla, Ballarè, Goria & Mauri 2022), e la inquadriamo in una prospettiva di linguistica interazionale basata su corpus multimodale.

La linguistica interazionale esamina il linguaggio, soprattutto il parlato, tenendo conto del ruolo costitutivo dell'interazione sociale organizzata sequenzialmente (Couper-Kuhlen & Selting 2017; Mondada 2001). L'integrazione della dimensione incorporata, resa possibile dall'uso di dati video (es. Deppermann 2014; Mondada 2013; Pekarek Doehler 2022), è uno sviluppo naturale di questo paradigma, che tuttavia solleva problemi metodologici importanti: fra questi, l'organizzazione di corpora e metadati (cfr. Deppermann & Reineke 2023); l'impostazione tecnica della videoregistrazione; questioni etiche e tecniche legate alla gestione dell'immagine video (Profazi & Miecznikowski 2023); la trascrizione dei fenomeni multimodali (Greco et al. 2019; Mondada 2018); il coding semplificato di tali fenomeni nell'annotazione collaborativa (Jacquin et al. 2022); l'esigenza infrastrutturale di creare piattaforme atte a esplorare, interrogare e condividere corpora multimodali (cfr. CLAPI, Baldauf-Quillatre 2016; AGD presso il Leibniz-Institut für Deutsche Sprache[2]; ZuMult[3]; LCP-Videoscope, Graën et al. 2024).

Presenteremo alcune soluzioni adottate di fronte a questi problemi nella costruzione, gestione e analisi del corpus TIGR. Il corpus design distingue diversi generi di interazione sulla base di parametri situazionali come +/- istituzionalità o +/- asimmetria dei ruoli accanto a parametri relativi alle attività pratiche (+/- manipolazione di oggetti, +/- multi-attività, +/-spostamento nello spazio). I dati personali sono stati gestiti ottenendo dai partecipanti il consenso informato alla condivisione e all'uso dei documenti ai fini della ricerca scientifica e offrendo – oltre alla pseudonimizzazione delle trascrizioni – opzioni di de-identificazione come l'alterazione della voce e il mascheramento del viso.

[2] Leibniz-Institut für Deutsche Sprache. Archiv für Gesprochenes Deutsch (AGD). https://agd.ids-mannheim.de/index.shtm

[3] ZuMult. Zugang zu multimodlen Korpora gesprochener Sprache. http s://zumult.org/

I dati sono stati trascritti in ELAN secondo le convenzioni GAT 2 per la trascrizione fine (Selting et al. 2011); le trascrizioni sono poi state predisposte in altri tre formati per anticipare esigenze sia dell'analisi qualitativa sequenziale e multimodale, sia dell'annotazione collaborativa: documenti TXT impaginati con una quantità ridotta di timecode (TIGRformat 2025); documenti XML tokenizzati e conformi allo standard ISO/TEI 20624:2016 (Hedeland & Schmidt 2022), che rendono interrogabile l'informazione codificata dalle convenzioni di trascrizione; dopo modellazione dei dati, tabelle ottenute dai documenti XML che costituiscono l'input per la rappresentazione nella piattaforma multimediale LCP-Videoscope (Miecznikowski et al. *in preparazione*).

Accenneremo infine alle prime analisi condotte sui dati del corpus TIGR per illustrarne il potenziale: misurazioni condotte a partire dai documenti XML sulla lunghezza dei turni e la varietà lessicale nei vari generi di interazione (Miecznikowski et al. *in preparazione*) e un caso studio sul segnale discorsivo *ah* (Miecznikowski & Battaglia 2026) che integra i dati video allineati alle trascrizioni nell'annotazione. Complessivamente, i risultati mostrano correlazioni tra fenomeni linguistici e parametri situazionali o relativi alle attività pratiche, che sono pienamente apprezzabili solo in presenza del dato multimodale, sollecitando dunque ulteriori sforzi in questa direzione.

**Riferimenti bibliografici**


Albano Leoni, Federico. 2007. Un frammento di storia recente della ricerca (linguistica) italiana. Il corpus CLIPS. *Bollettino d'Italianistica* 4(2). 122–130.

Baldauf-Quilliatre, Heike, Carvajal, Isabel Colón de, Etienne, Carole, Jouin-Chardon, Elwys, Teston-Bonnard, Sandra & Traverso, Véronique. 2016. CLAPI, une base de données multimodale pour la parole en interaction: apports et dilemmes. *Corpus* 15. https://doi.org/10.4000/corpus.2991

Ballarè, Silvia, Goria, Eugenio & Mauri, Caterina. 2022. *Italiano parlato e variazione linguistica. Teoria e prassi nella costruzione del corpus KIParla*. Bologna: Pàtron Editore.

Couper-Kuhlen, Elizabeth & Selting, Margret. 2017. *Interactional Linguistics: Studying Language in Social Interaction*. Cambridge: Cambridge University Press.

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

Deppermann, Arnulf. 2014. Multimodal participation in simultaneous joint projects: interpersonal and intrapersonal coordination in paramedic emergency drills. In Haddington, Pentti, Keisanen,

Tiina, Mondada, Lorenza & Nevile, Maurice (a cura di), *Multiactivity in Social Interaction: Beyond Multitasking*, 247–282. Amsterdam: John Benjamins.

Deppermann, Arnulf & Reineke, Silke. 2023. Zur Verwendung von Metadaten in der interaktionsanalytischen Arbeit mit Korpora - am Beispiel einer Untersuchung anhand des Korpus FOLK. In Beißwenger, Michael, Gredel, Eva, Lemnitzer, Lothar & Schneider, Roman (a cura di), *Korpusgestützte Sprachanalyse: Grundlagen, Anwendungen und Analysen*, 245–259. Tübingen: Leibniz-Institut für Deutsche Sprache (IDS).

Graën, Johannes, Schaber, Johannes, McDonald, David, Mustač, Ivan, Rajović, Nikola, Schneider, Gerold, Vuković, Teodora, Zehr, Jérémy & Bubenhofer, Noah. 2024. The LiRI Corpus Platform. In Lindén, Krister, Kontino, Thalassia & Niemi, Jyrki (a cura di), *Selected papers from the CLARIN Annual Conference 2023 (Linköping Electronic Conference Proceedings 210)*, 62–75. https://doi.org/10.3384/ecp210010

Greco, Luca, Galatolo, Renata, Horlacher, Anne-Sylvie, Piccoli, Vanessa, Ticca, Anna Claudia & Ursi, Biagio. 2019. Some theoretical and methodological challenges of transcribing touch in talk-in-interaction. *Social Interaction. Video-Based Studies of Human Sociality* 2(1). https://hal.science/hal-02956481v2

Hedeland, Hanna & Schmidt, Thomas. 2022. The TEI-based ISO Standard 'Transcription of spoken language' as an Exchange Format within CLARIN and beyond. In Monachini, Monica & Eskevich, Maria (a cura di), *Selected papers from the CLARIN Annual Conference 2021*, *(Linköping Electronic Conference Proceedings 189)*, 34–45. https://doi.org/10.3384/ecp1894

Jacquin, Jérôme, Keck, Ana, Robin, Clotilde & Roh, Sabrina. 2022. *Guide d'annotation du projet POSEPI*. (Manuel technique). https://doi.org/10.5281/zenodo.7266737

Miecznikowski, Johanna. 2025. TIGRformat 2025, v. 1.0 (Software). https://github.com/JohannaMiecznikowski/Infinitaplus/releases/tag/v.1.0.0

Miecznikowski, Johanna & Battaglia, Elena. 2026. Aspetti evidenziali, sequenziali e multimodali di *ah* come segnale di acquisizione della conoscenza. Poster presentato a *XIV Italianistiktag, Graz, 28 febbraio 2026*.

Miecznikowski, Johanna, Battaglia, Elena, Schmidt, Thomas & Zehr, Jérémy. *In preparazione*. *The TIGR corpus of spoken Italian*.

Miecznikowski-Fuenfschilling, Johanna, Battaglia, Elena & Geddo, Christian. *In fase di valutazione*. *General TIGR documentation (Version 1.0)*. Dataset. LaRS - Language Repository of Switzerland. https://doi.org/10.48656/mgq4-7p77

Mondada, Lorenza. 2001. Pour une linguistique interactionnelle. *Marges Linguistiques* 1. 142–162.

Mondada, Lorenza. 2013. Deixis: an integrated interactional multimodal analysis. In Bergmann, Pia, Brenning, Jana, Pfeiffer, Martin & Reber, Elisabeth (a cura di), *Prosody and Embodiment in Interactional Grammar*, 173-206. Berlin/Boston: De Gruyter.

Mondada, Lorenza. 2018. Multiple Temporalities of Language and Body in Interaction: Challenges for Transcribing Multimodality. *Research on Language and Social Interaction* 51(1). 85–106. https://doi.org/10.1080/08351813.2018.1413878

Pekarek Doehler, Simona. 2022. Multimodal action formats for managing preference: *chais pas* 'dunno' plus gaze conduct in dispreferred responses to questions. *Journal of Pragmatics* 197. 81–99. https://doi.org/10.1016/j.pragma.2022.05.010

Profazi, Nina & Miecznikowski, Johanna. 2023. Social interaction is among people. Legal, technical, and ethical explorations about personal information and its removal in talk-in-interaction as data. (Post su blog, 28 agosto 2023). https://www.chord-talk-in-interaction.usi.ch/news/feeds/36387

Selting, Margret, Auer, Peter, Barth-Weingarten, Dagmar, Bergmann, Jörg, Bergmann, Pia, Birkner, Karin, Couper-Kuhlen, Elizabeth, Deppermann, Arnulf, Gilles, Peter, Günthner, Susanne, Hartung, Martin, Kern, Friederike, Mertzlufft, Christine, Meyer, Christian, Morek, Miriam, Oberzaucher, Frank, Peters, Jörg, Quasthoff, Uta, Schütte, Wilfried & Uhmann, Susanne. 2011. A system for transcribing talk-in-interaction: GAT 2 translated and adapted for English by Elizabeth Couper-Kuhlen and Dagmar Barth-Weingarten. *Gesprächsforschung* 12. 1–51. http://www.gespraechsforschung-online.de/fileadmin/dateien/heft2011/px-gat2-englisch.pdf

**LAURA OCCHIPINTI**

## Misurare l'efficacia comunicativa: dati linguistici oggettivi e soggettivi nella leggibilità infantile

La leggibilità è comunemente definita come la facilità o difficoltà con cui i lettori possono comprendere il contenuto di una frase o di un testo scritto (Weiss & Meurers 2019). Le metriche di leggibilità mirano dunque a rappresentare la complessità percepita operazionalizzandola attraverso caratteristiche formali dei testi, quali lunghezza delle parole, la densità lessicale o complessità sintattica (Štajner et al. 2012). Tali indici si fondano prevalentemente su parametri quantitativi e strutturali, con una marcata attenzione alle dimensioni sintattiche e superficiali del testo. Le variabili semantiche, pur essendo centrali per la costruzione del significato, sono invece meno frequentemente integrate nei modelli di leggibilità, in particolare per la lingua italiana (Dell'Orletta Montemagni & Venturi 2011; Palmero Aprosio & Moretti 2018), anche in ragione della maggiore difficoltà di definirle e misurarle in modo sistematico (Occhipinti 2024). Tuttavia, la nozione di "complessità percepita" è intrinsecamente relativa (Dahl 2004; Miestamo, Karlsson & Sinnemäki 2008; Pallotti 2015), poiché dipende dalle risorse linguistiche e cognitive dei destinatari. Di conseguenza, una concezione esclusivamente quantitativa e oggettiva del dato linguistico rischia di risultare riduttiva.

Il presente contributo affronta tale questione a partire da uno studio sperimentale sull'efficacia comunicativa di testi destinati a lettori in età evolutiva, proponendo di integrare nel concetto di dato linguistico anche dimensioni semantiche e valutative.

Tradizionalmente la leggibilità è stata trattata come proprietà intrinseca del testo, misurabile attraverso indici strutturali. L'efficacia comunicativa, però, non coincide con tali parametri formali ma emerge dall'interazione tra caratteristiche linguistiche, incluse quelle semantiche, e processi cognitivi del lettore. In questa prospettiva, i giudizi di chiarezza (facilità di comprensione, assenza di confusione) e di informatività (quantità e rilevanza dell'informazione) costituiscono dati soggettivi che riflettono dimensioni parzialmente indipendenti del processamento linguistico.

Nel presente studio indaghiamo il rapporto tra due proprietà lessicali, concretezza e specificità (Bolognesi, Burgers & Caselli 2020; Villani & Bolognesi 2025) e valutazioni esplicite sulla comprensibilità testuale fornite da bambini di età compresa tra i 9 e i 13 anni. Sono stati coinvolti 103 bambini, ai quali sono stati presentati brevi testi narrativi costruiti manipolando sistematicamente due proprietà lessicali formalmente definite: concretezza (astratto vs. concreto) e specificità (generale, di base, o specifico). I testi, controllati per struttura morfosintattica, sono stati valutati mediante scale Likert a 5 punti su quattro dimensioni riconducibili a chiarezza (facilità e assenza di confusione) e

informatività (quantità e rilevanza dell'informazione). I dati sono stati analizzati mediante modelli lineari misti, mettendo in relazione proprietà linguistiche oggettive e giudizi soggettivi, con effetti casuali per partecipanti e item.

I risultati mostrano una dissociazione sistematica: la concretezza predice principalmente le valutazioni di chiarezza, mentre la specificità incide soprattutto sull'informatività percepita. Emergono inoltre differenze legate allo sviluppo, suggerendo che il peso attribuito alle diverse dimensioni informative varia con le risorse cognitive e metacognitive del lettore. Proponiamo quindi un modello integrato in cui il dato linguistico non coincide esclusivamente con proprietà strutturali del testo, ma include anche misure graduali di esperienza soggettiva, la cui analisi congiunta consente di ridefinire in modo più articolato il concetto di leggibilità e di efficienza comunicativa.

**Riferimenti bibliografici**


Aprosio, Alessio Palmero & Moretti, Giovanni. 2018. Tint 2.0: an All-inclusive Suite for NLP in Italian. In Cabrio, Elena, Mazzei, Alessandro & Tamburini, Fabio (a cura di), *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, *Torino, 10-12 dicembre 2018*, 312–318. CEUR Workshop Proceedings. https://aclanthology.org/2018.clicit-1.0/

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

Dahl, Östen. 2004. *The growth and maintenance of linguistic complexity*. Amsterdam/Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/slcs.71

Dell'Orletta, Felice, Montemagni, Simonetta & Venturi, Giulia. 2011. READ–IT: Assessing readability of Italian texts with a view to text simplification. In Alm, Norman (a cura di), *Proceedings of the second workshop on speech and language processing for assistive technologies, Edinburgh, 30 luglio 2011.* 73–83. Stroudsburg: The Association for Computational Linguistics. https://aclanthology.org/W11-2308/

Miestamo, Matti, Karlsson, Fred & Sinnemäki, Kaius. 2008. *Language complexity*. Amsterdam: John Benjamins Publishing Company.

Occhipinti, Laura. 2025. *Lexical complexity analysis from a computational perspective: a case study on institutional Italian language*. Bologna: Alma Mater Studiorum - Università di Bologna (Tesi di dottorato).

Pallotti, Gabriele. 2015. A simple view of linguistic complexity. *Second Language Research* 31(1). 117–134. https://doi.org/10.1177/0267658314536435

Štajner, Sanja, Evans, Richard, Orasan, Constantin & Mitkov, Ruslan. 2012. What can readability measures really tell us about text complexity. In Rello, Luz, Saggion, Horacio & Baeza-Yates Ricardo (a cura di), *Proceedings of workshop on natural language processing for improving textual accessibility, Atlanta, 14 giugno 2013,* 14–22. Stroudsburg: The Association for Computational Linguistics.

Villani, Caterina & Bolognesi, Marianna. 2025. Specificity: Metrics and Norms. In Nesi, Hilary & Milin, Petar (a cura di), *International Encyclopedia of Language and Linguistics 3rd Edition-Reference Module in Social Sciences, 2025*. Elsevier, 1–6. https://dx.doi.org/10.1016/b978-0-323-95504-1.00894-2

Weiss, Zarah & Meurers, Detmar. 2019. Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school. In Yannakoudakis, Helen, Kochmar, Ekaterina, Leacock, Claudia, Madnani, Nitin & Pilán, Ildikó & Zesch, Torsten (a cura di), *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications, Firenze, 2 agosto 2019*, 380–393. Stroudsburg: The Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4440

**LUCREZIA SABIA**

## Corpora e dati sperimentali al servizio dell'antonimia: sviluppo teorico di una relazione semantica

*Buono*/*cattivo*, *amore*/*odio*, *su*/*giù*, *vendere*/*comprare* sono esempi di antonimia, una relazione semantica di immediata individuazione, ma di difficile definizione teorica. A sottolinearne l'automatismo, la letteratura si riferisce al riconoscimento intuitivo del rapporto di antonimia con l'espressione *clang phenomenon* (Muehleisen 1997, 4). Tuttavia, alla sua natura autoesplicativa non corrisponde una definizione teorica uniforme. Questo lavoro si propone di indagare il ruolo svolto dai dati linguistici nella costruzione e riformulazione dei modelli teorici sul concetto di antonimia. Infatti, l'osservazione di dati elicitati da esperimenti psicolinguistici, da un lato, e di quelli estratti da corpora digitali, dall'altro, ha influenzato ed orientato l'illustrazione dell'antonimia e delle sue caratteristiche principali, in uno sviluppo teorico progressivo e in dialogo continuo con le teorie linguistiche succedutesi nella storia della disciplina.

Seguendo un filo cronologico, l'intervento illustrerà inizialmente l'antonimia come relazione associativa. L'approccio categoriale (Cruse 1986; Lyons 1977), insieme a lavori psicolinguistici su test di associazione libera (Charles & Miller 1989; Deese 1965), sostiene questa definizione di antonimia automatica e stabile, al punto da costituire le fondamenta di risorse digitali come il Princeton WordNet (Gross & Miller 1990). In questo contesto, l'antonimia è percepita come relazione semantico-lessicale. In seguito, nuovi dati, sia sperimentali che estratti da corpora, hanno condotto al passaggio da una concezione statica della relazione antonimica ad una più dinamica e in linea con gli elementi fondanti della linguistica cognitiva (Croft & Cruse 2004; Langacker 1987; Talmy 2000). In questo contesto, invece, l'antonimia diventa una relazione concettuale. A sostegno di ciò, studi che combinano linguistica dei corpora e metodi sperimentali (es. Paradis, Willners & Jones 2009) mettono al centro la nozione, gradabile, di canonicità. Ancora studi basati su corpora portano a una ridefinizione dell'antonimia, dimostrando come le coppie di opposti co-occorrano nei testi con una frequenza significativamente superiore a quella attesa (Justeson & Katz 1991, 1992): la relazione paradigmatica viene quindi riconsiderata come anche sintagmatica. Questa osservazione dà origine ad indagini sulle sue funzioni testuali e discorsive, in particolare a partire da corpora giornalistici (Jones 2002), generando ulteriori riflessioni teoriche. Spicca, in tal senso, l'inquadramento dell'antonimia all'interno della Grammatica delle Costruzioni (Murphy 2006). La sua rappresentazione all'interno del "Costruzionario" (Masini 2016, 54), tuttavia, rimane un ambito ancora poco esplorato, forse per carenza di una forza teorica decisiva: nuovi dati.

**Riferimenti bibliografici**


Charles, Walter G. & George A. Miller. 1989. Contexts of Antonymous Adjectives. *Applied Psycholinguistics* 10(3). 357–375. https://doi.org/10.1017/S0142716400008675

Croft, William & Alan D. Cruse. 2004. *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Cruse, Alan D. 1986. *Lexical semantics*. Cambridge: Cambridge University Press.

Deese, James E. 1965. *The structure of associations in language and thought*. Baltimore: Johns Hopkins University Press.

Gross, Derek & Katherine J. Miller. 1990. Adjectives in WordNet. In WordNet: an on-line lexical database. *International Journal of Lexicography* 3(4). 265–277. https://doi.org/10.1093/ijl/3.4.265

Jones, Steven. 2002. *Antonymy: a corpus-based perspective*. London: Routledge.

Justeson, John S. & Katz, Slava M. 1991. Co-occurrences of antonymous adjectives and their contexts. *Association for Computational Linguistics* 17(1). 1–19.

Justeson, John S. & Katz, Slava M. 1992. Redefining antonymy: the textual structure of a semantic relation. *Literary and Linguistic Computing* 7(3). 176–184. https://doi.org/10.1093/llc/7.3.176

Langacker, Ronald W. 1987. *Foundations of cognitive grammar*. Stanford: Stanford University Press.

Lyons, John. 1977. *Semantics*. Cambridge: Cambridge University Press.

Muehleisen, Victoria L. 1997. *Antonymy and semantic range in English*. Evaston: Northwestern University. (Tesi di dottorato.)

Masini, Francesca. 2016. *Grammatica delle costruzioni.* Roma: Carrocci.

Murphy, Lynne M. 2006. Antonyms as lexical constructions: or, why paradigmatic construction is not an oxymoron. *Constructions* Special Volume 1–8/2006. https://doi.org/10.24338/cons-449

Paradis, Carita, Willners, Caroline & Jones, Steven. 2009. Good and bad opposites. Using textual and experimental techniques to measure antonym canonicity. *The Mental Lexicon* 4(3). 380–429. https://doi.org/10.1075/ml.4.3.04par

Talmy, Leonard. 2000. *Toward a cognitive semantics*. Cambridge, MA: The MIT Press.

**FLORIANA CARLOTTA SCIUMBATA**

## Dalla parola al dato: metodi e problemi nella creazione di liste lessicali

Il contributo prende in esame tre ambiti lessicali specifici, relativi ai nomi di mestieri e professioni, ai titoli associati al potere (militari, religiosi e nobiliari) e ai luoghi artificiali, già utilizzati nell'estrazione automatica come categorie indicative di trasformazioni sociali e culturali in corpora di narrativa (Sciumbata 2023; Sciumbata, Nadalutti & Tringali 2021). L'obiettivo è mettere in luce alcune difficoltà ricorrenti nella creazione di liste destinate all'estrazione automatica. Una prima criticità riguarda la disponibilità di dati strutturati: dizionari ed enciclopedie cartacei e digitali, pur ricchi di informazioni, presentano definizioni poco sistematiche e dati non strutturati, che ne limitano l'utilizzo diretto in procedure automatiche, soprattutto quando si lavora su domini semantici specifici e su testi letterari diacronicamente distanti.

Anche quando le fonti sono disponibili, esse risultano spesso parziali o non pienamente adeguate agli obiettivi dell'analisi, rendendo necessaria l'integrazione di risorse diverse. In questo contesto, il contributo discute l'uso di basi di dati strutturate e interrogabili come Wikidata, che consentono un'estrazione più controllata e riproducibile delle voci lessicali, ma che possono risultare limitati e che richiedono comunque interventi di selezione, normalizzazione e verifica qualitativa.

Accanto a queste risorse, sono stati sperimentati anche strumenti basati su modelli linguistici di grandi dimensioni, come NotebookLM, utilizzati in fase esplorativa per individuare categorie rilevanti e ampliare liste preliminari. Tali strumenti offrono nuove possibilità di supporto alla costruzione delle risorse, ma mostrano limiti evidenti in termini di sistematicità, tracciabilità delle scelte e riproducibilità dei risultati. Rendono quindi necessario un controllo umano stringente e ne impediscono l'uso come fonte primaria per la creazione delle liste.

Un'ulteriore difficoltà emerge nella fase di estrazione automatica dei dati dai testi, soprattutto in ottica diacronica, ed è rappresentata dalla morfologia dell'italiano. La variazione di genere e numero, le alternanze morfologiche e grafiche e la polisemia rendono problematico il ricorso a liste basate unicamente sui lemmi e impongono scelte metodologiche specifiche, quali il lavoro sulle forme anziché sui soli lemmi, l'uso di testi annotati per parte del discorso e l'eliminazione manuale delle occorrenze ambigue. In questo quadro, anche gli strumenti automatici di annotazione linguistica mostrano limiti nel trattamento di testi letterari storici, per i quali non sono sempre adeguatamente calibrati.

Attraverso l'analisi di casi concreti, il contributo mira a evidenziare alcune buone pratiche nella creazione di liste lessicali per l'estrazione automatica dei dati, con particolare attenzione alle soluzioni operative che si sono dimostrate efficaci nel confronto con corpora letterari storici e con risorse lessicali non sempre adeguate. Le procedure illustrate possono essere riprese e adattate anche ad altri tipi testuali e a liste lessicali riferite a domini semantici differenti.

**Riferimenti bibliografici**

Sciumbata, Floriana Carlotta, Nadalutti, Paolo & Tringali, Luca. 2021. "Trovare lavoro" in un corpus di narrativa del XIX–XX secolo. Procedure, aspetti e problemi di creazione, estrazione e rappresentazione dei dati. *Rivista Internazionale di Tecnica della Traduzione* 23. 235–268. https://doi.org/10.13137/2421-6763/33468

Sciumbata, Floriana Carlotta. 2023. Titles associated with power in a corpus of fictional prose in Italian (XIX–XXI century). In Lopičić, Vesna, Mišić Ilić, Biljana, Mitić, Ivana & Ignjatović, Sanja (a cura di), *Jezik, književnost, moć / Language, Literature, Power*, 279–297. Niš: Filozofski fakultet. https://doi.org/10.46630/jkm.2023.17

**MARTINA VERDELLI**

## Sfide nella raccolta dati di verbi sintagmatici nelle varietà bergamasche: metodologie a confronto

Il presente contributo si pone l'obiettivo di presentare le sfide e problematiche connesse alla raccolta dati incentrata sui verbi sintagmatici (d'ora in avanti VS), una tipologia di espressione multiparola costituita da un verbo (generalmente di moto) e da una particella di originario valore locativo (cfr. in particolare: Cini 2008; Simone 1996 per lo stato dell'arte).

Data la particolarità di queste strutture, la maggior parte degli studi disponibili sul tema si basa su corpora, mentre risultano tuttora assenti contributi di taglio metodologico connessi alle potenziali tecniche da adoperare per raccogliere questo tipo di dati. Le difficoltà legate all'elicitazione di queste strutture sono infatti numerose: da un lato rappresentano una classe potenzialmente aperta, con combinazioni [V+PRT] e significati (anche non composizionali) variabili a seconda della lingua considerata; dall'altro, sono spesso in competizione con forme monosemiche quasi-sinonimiche.

In virtù di ciò, l'intervento intende presentare e confrontare alcune possibili strategie (orali e scritte), prendendo come caso studio le varietà bergamasche, per le quali sono già presenti degli studi sul tema (cfr. in particolare Bernini 2006, 2008, 2012, 2021). Non essendo ancora disponibile un corpus di riferimento per il bergamasco, i dati a disposizione sono però esegui: è infatti possibile raccogliere dati sui VS a partire dalle opere lessicografiche presenti così come dagli atlanti linguistici; tali opere raccolgono però materiale lessicale molto vario, non specificamente incentrato su queste combinazioni. In assenza di un corpus da poter interrogare, si rende dunque necessaria la raccolta di dati sul campo. Per questo motivo sono stati ideati e sperimentati diversi test, la cui sintesi è rappresentata nella tabella che segue.

| Task di riferimento | Descrizione dell'attività |
|---|---|
| Kit stimoli | Descrizione di una serie di immagini raffiguranti entità animate che compiono o subiscono azioni |
| Traduzioni e retroformazioni | Traduzioni di frasi dall'italiano al dialetto (e viceversa) |
| Frasi da completare | Completamento di una serie di frasi scegliendo autonomamente o tra le opzioni proposte un costituente verbale o un VS |
| Frasi da creare | Creazione di una serie di frasi a partire da un dato costituente verbale o VS |
| Discorso libero | Intervento libero su un tema proposto dall'intervistatore |
| Gioco di ruolo | Dialogo tra due o più informatori a partire da un tema o scena proposta dall'intervistatore |

Tabella 1. Task sperimentati per la raccolta di VS

Facendo riferimento a quanto raccolto sul campo nel corso degli anni 2023-2025, l'intervento illustrerà non solo l'articolazione e somministrazione di ciascun *task*, ma ne metterà in luce vantaggi e svantaggi. L'obiettivo è quello di mettere in luce quali possano essere i *task* più rilevanti e utili per raccogliere questo tipo di dato, al punto da poter essere adottati anche per altre varietà prive di un corpus.

**Riferimenti bibliografici**


Bernini, Giuliano. 2006. Strategie di lessicalizzazione: tipologia e apprendimento. Il caso dei verbi di moto. *Zeitschrift fur Literaturwissenschaft und Linguistik* 36. 95–118. https://doi.org/10.1007/BF03379631

Bernini, Giuliano. 2008. Per una definizione di verbi sintagmatici: la prospettiva dialettale, in Cini Monica (a cura di), *I verbi sintagmatici in italiano e nelle varietà dialettali. Stato dell'arte e prospettive di ricerca. Atti delle giornate di studio. Torino, 19–20 febbraio 2007*, 141–157. Frankfurt am Main: Peter Lang.

Bernini, Giuliano. 2012. Lexicalization and grammaticalization in the area of multi-word verbs: A case-study from Italo-Romance. In Bambini, Valentina, Ricci Irene & Bertinetto, Pier Massimo (a cura di), *Linguaggio e cervello – Semantica. Language and the brain – Semantics*, *Atti del XLII congresso internazionale di Studi della Società di Linguistica Italiana (SLI)*, *Pisa, 25-27 settembre 2008,* 131–154. Roma: Bulzoni.

Bernini, Giuliano. 2021. Different sources of convergent patterns in the Alps. *STUF - Language Typology and Universa*ls. *Sprachtypologie und Universalienforschung* 74(1). 39–71. https://doi.org/10.1515/stuf-2021-1023

Cini, Monica (a cura di). 2008. *I verbi sintagmatici in italiano e nelle varietà dialettali. Stato dell'arte e prospettive di ricerca. Atti delle giornate di studio. Torino, 19-20 febbraio 2007*. Frankfurt am Main: Peter Lang.

Simone, Raffaele. 1996. Esistono verbi sintagmatici in italiano? *Cuadernos de Filología Italiana* 3. 47–61. https://revistas.ucm.es/index.php/CFIT/article/view/CFIT9696110047A

**PAOLA VERNILLO**

## Dalla distinzione teorica alla variabile osservabile: la specificità lessicale nel giudizio metaforico

Nel quadro della *Conceptual Metaphor Theory* (Lakoff & Johnson 1980), la metafora è stata definita nei termini di una mappatura sistematica tra un dominio *source*, concreto e percettivamente accessibile, e un dominio *target*, astratto. Sul piano lessicale, tale asimmetria è stata indagata principalmente tramite scale di *rating* (Brysbaert, Warriner & Kuperman 2014) volte a quantificare la concretezza, intesa come grado di riferibilità percettiva dell'espressione linguistica. A livello sperimentale, la concretezza è stata inoltre collegata a variabili centrali per il processamento metaforico (sensatezza, facilità, familiarità, appropriatezza, convenzionalità) ed è stata impiegata come predittore della qualità percepita delle espressioni metaforiche (Blasko & Connine 1993; Thibodeau, Sikos & Durgin 2018).

Un aspetto distinto è la specificità lessicale, ossia il grado di dettaglio semantico di un termine all'interno di una gerarchia categoriale (es. *camminare* vs. *zoppicare*). Sebbene discussa teoricamente dalla linguistica cognitiva (Dancygier & Sweetser 2014), la specificità non è stata formalizzata sistematicamente come predittore del processamento metaforico (Bolognesi, Burgers & Caselli 2020). Il disallineamento tra concretezza e specificità è, però, evidente: termini ugualmente concreti possono differire per granularità semantica, con possibili effetti sull'interpretazione e sulla resa estetica della metafora. Se formulazioni più generali facilitano l'interpretazione grazie a *mapping* consolidati, formulazioni più specifiche arricchiscono il dominio *source* ampliandone il potenziale inferenziale.

Il presente contributo indaga sperimentalmente il ruolo della specificità del predicato nella qualità percepita delle metafore in italiano. Si ipotizza che la qualità metaforica, valutata in termini di efficacia, appropriatezza e valore estetico, dipenda non solo dalla distanza concettuale tra domini, ma anche dalla granularità semantica del *source*. A parità di frequenza, predicati meno specifici dovrebbero favorire valutazioni più elevate, mentre predicati più specifici, sebbene più vividi, potrebbero compromettere la qualità della metafora. L'effetto dovrebbe attenuarsi o scomparire in contesti letterali.

Per testare questa ipotesi, è stato costruito un *dataset* controllato di metafore predicative italiane secondo un disegno misto con quattro condizioni: metaforico-generale, metaforico-specifico, letterale-generale e letterale-specifico. La raccolta comprende 188 frasi sintatticamente comparabili (94 coppie), differenziate per specificità e concretezza e adattate dal *Figurative Archive* (Bressler et

al. 2026), bilanciate per frequenza e struttura sintattica e validate preliminarmente. Un *pilot* ha verificato la plausibilità della distinzione tra predicati generali e specifici; un successivo *rating* di concretezza (scala 1-7; 55 partecipanti) ha escluso confondimenti sistematici, evidenziando un effetto della metaforicità ma non della specificità né della loro interazione. Gli esperimenti in corso estenderanno l'analisi alla valutazione multidimensionale della qualità metaforica, integrando *rating* di comprensibilità, facilità di interpretazione, ricchezza dell'immaginario e numero di interpretazioni (Katz et al. 1988). I dati saranno analizzati con modelli lineari misti, usando specificità e concretezza come variabili indipendenti e la frequenza come covariata.

In conclusione, il lavoro qui descritto intende mostrare come la qualità metaforica sia sensibile non solo alla concretezza ma anche alla specificità lessicale. Intende, altresì, proporre una riflessione su come una distinzione teorica possa essere tradotta in dato osservabile e variabile empiricamente manipolabile nella ricerca linguistica contemporanea.

**Riferimenti bibliografici**


Blasko, Dawn G. & Connine, Cynthia M. 1993. Effects of familiarity and aptness on metaphor processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19(2). 295–308. https://doi.org/10.1037/0278-7393.19.2.295

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

Bressler, Maddalena, Mangiaterra, Veronica, Canal, Paolo, Frau, Federico, Luciani, Fabrizio, Scalingi, Biagio, Barattieri di San Pietro, Chiara, Battaglini, Chiara, Pompei, Chiara, Romeo, Fortunata, Bischetti, Luca & Bambini, Valentina. 2026. Figurative Archive: An open dataset and web-based application for the study of metaphor. *Scientific Data* 13. 151. https://doi.org/10.1038/s41597-025-06459-7

Brysbaert, Marc, Warriner, Amy Beth & Kuperman, Victor. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46. 904–911. https://doi.org/10.3758/s13428-013-0403-5

Dancygier, Barbara & Sweetser, Eve. 2014. *Figurative Language*. Cambridge: Cambridge University Press.

Katz, Albert. N., Paivio, Allan, Marschark, Mark & Clark, Jim. M. 1988. Norms for 204 literary and 260 nonliterary metaphors on 10 psychological dimensions. *Metaphor and Symbolic Activity* 3(4). 191–214. https://doi.org/10.1207/s15327868ms0304_1

Lakoff, George & Johnson, Mark. 1980. *Metaphors we live by*. Chicago: University of Chicago Press.

Thibodeau, Paul H., Sikos, Les & Durgin, Frank H. 2018. Are subjective ratings of metaphors a red herring? The big two dimensions of metaphoric sentences. *Behavior Research Methods* 50(2). 759–772. https://doi.org/10.3758/s13428-017-0903-9