Italian Crowdsourcing Project: Tempi di riconoscimento visivo, accuratezza e prevalenza per 130.495 parole italiane #699

Author:

Simona Amenta

Co-authors:

Andrea Gregor de Varda, Pawel Mandera, Emmanuel Keuleers, Marc Brysbaert, Marco Marelli

Introduzione: Nonostante l'italiano sia oggetto di numerosi studi psicolinguistici, mancano ancora risorse su larga scala per l'analisi del linguaggio. Negli ultimi anni, il *crowdsourcing* si è affermato come metodo efficace per raccogliere dati sul riconoscimento delle parole, coinvolgendo ampi gruppi di partecipanti con caratteristiche demografiche diverse, al di fuori dei tradizionali contesti di laboratorio (Keuleers & Balota, 2015). Crowdsourcing Vocabulary Projects sono già stati sviluppati per varie lingue europee (Amenta, 2024).

Obiettivi: Questo studio presenta l'Italian Crowdsourcing Project (ICP), un ampio dataset contenente tempi di risposta e accuratezza nel riconoscimento di 130.495 parole, rendendolo il più esteso del suo genere per numero di item.

Metodi: I dati sono stati raccolti tramite un compito online di conoscenza delle parole (*word knowledge task*), a cui hanno partecipato 156.625 madrelingua italiani, per un totale di 15.906.229 datapoint (in media 85,65 osservazioni per parola).

Risultati: I tempi di reazione dell'ICP mostrano una forte correlazione (r = .78) con quelli ottenuti in un tradizionale esperimento di laboratorio con la stessa procedura. Inoltre, gli effetti delle principali variabili psicolinguistiche (es. frequenza, lunghezza) risultano replicabili in questo dataset. Infine, per la prima volta in italiano, è stata anche calcolata e studiata la *prevalenza* delle parole (cioè la percentuale di persone che conoscono una determinata parola), confermando risultati osservati in altre lingue.

Conclusioni: L'ICP è disponibile pubblicamente e rappresenta una risorsa fondamentale per lo studio del linguaggio, utile sia per la validazione cross-linguistica sia per l'analisi delle differenze individuali.