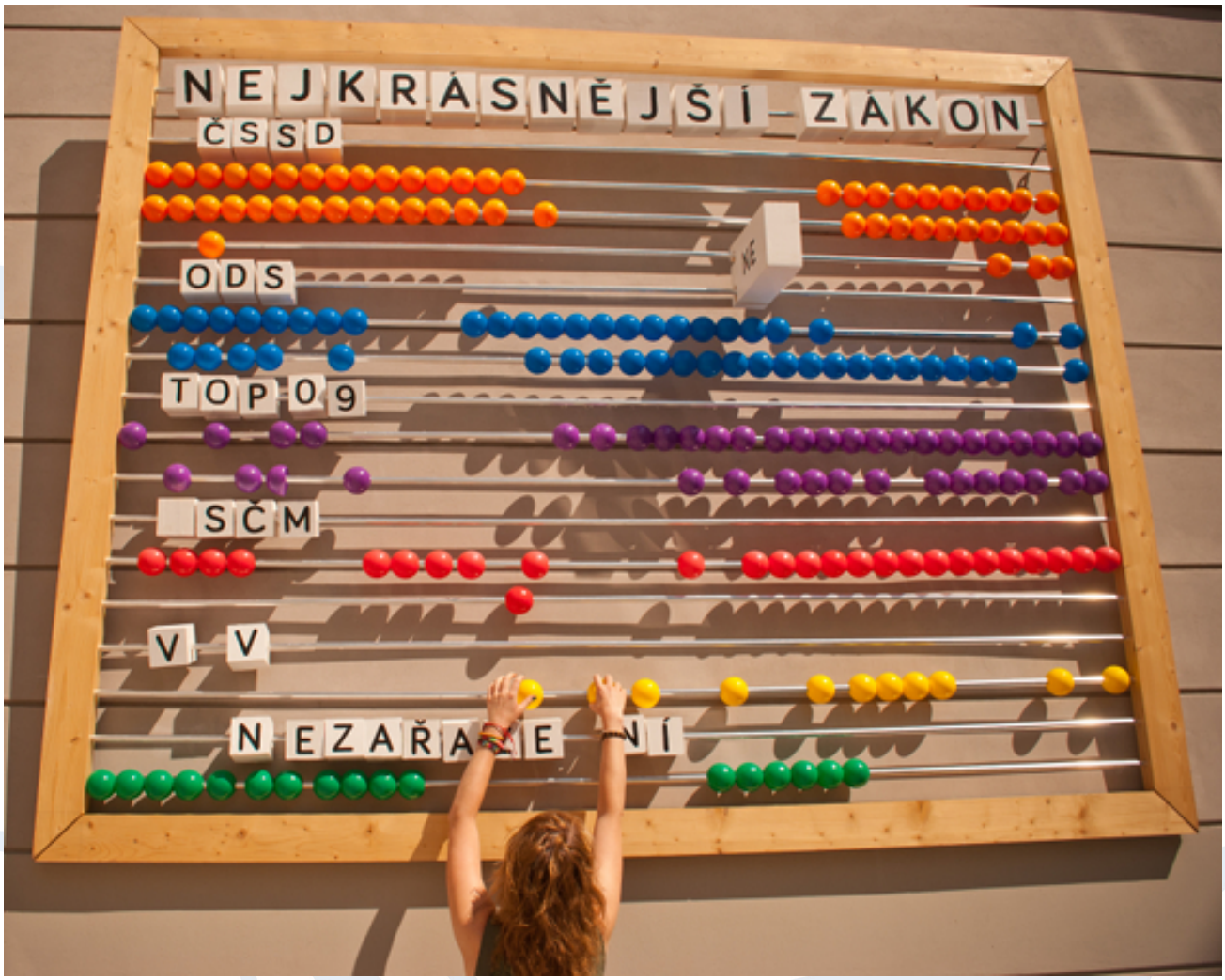
Parole, parole, parole... e statistica



Per distinguere le parole Il cervello potrebbe usare metodi statistici

19 maggio 2016

Distinguere le singole parole nel flusso del parlato non è una cosa semplice e secondo i linguisti per farlo il cervello potrebbe usare dei metodi statistici. Un gruppo di scienziati SISSA ha applicato un metodo di segmentazione delle parole basato sulla statistica e ne ha misurato l'efficacia sul linguaggio naturale, in ben 9 lingue diverse, scoprendo che il ritmo linguistico ha un ruolo importante. La ricerca è stata pubblicata sul *Journal of Developmental Science*.

Vi è mai capitato di rompervi il cervello cercando cogliere anche una sola parola nel flusso ininterrotto di un discorso in una lingua che conoscete a malapena? È ingenuo pensare che quando parliamo esista una anche minima pausa fra una parola e l'altra (come lo spazio che mettiamo per convenzione quando scriviamo): in realtà il parlato è quasi sempre un flusso sonoro continuo. Quando però ascoltiamo la nostra lingua madre questa "segmentazione" delle parole non ci costa quasi nessuno sforzo. Quali sono, si chiedono i linguisti, i meccanismi cognitivi automatici alla base di questa capacità? Un contributo certamente lo dà la conoscenza del vocabolario: la memoria del suono delle singole parole ci aiuta a rilevarle, ma, sostengono molti linguisti, esistono meccanismi automatici e inconsci di "basso livello" che ci aiutano anche quando non riconosciamo le parole, e anche, come nel caso di bambini molto piccoli, quando abbiamo una conoscenza ancora rudimentale della lingua. Questi meccanismi, credono gli scienziati, si basano sull'analisi statistica della frequenza (stimata in base all'esperienza pregressa) delle sillabe in ogni lingua.

Un indice che potrebbe contribuire ai processi di segmentazione è la "probabilità transizionale" (PT), che dà una stima della probabilità che due sillabe stiano all'interno della stessa parola, basandosi sulla frequenza con cui le due sillabe si trovano associate in una data lingua. In pratica, se tutte le volte che sento la sillaba "TA" viene immancabilmente seguita dalla sillaba "DA", allora la probabilità transizionale per "DA", data "TA" è 1 (il massimo). Se invece, ogni volta che sento la sillaba "BU" per metà delle volte capita che la segua la sillaba "DI" e per metà delle volte la sillaba "FI", la probabilità transizionale di "DI" (e di "FI") data "BU" è 0,5, e così via. Il sistema cognitivo potrebbe computare in maniera implicita questo valore sfruttando la memoria linguistica, dalla quale ricaverebbe le frequenze.

Lo studio condotto da Amanda Saksida, ricercatrice della Scuola Internazionale Superiore di Studi Avanzati (SISSA) di Trieste, con la collaborazione di Alan Langus, ricercatore SISSA, sotto la guida di Marina Nespor, professoressa della SISSA, ha usato l'indice PT per segmentare il linguaggio naturale, con due diversi approcci.

A seconda del ritmo

Saksida ha lavorato con i *corpus*, ossia collezioni di testi raccolte appositamente per l'analisi linguistiche. In questo caso specifico si tratta di trascrizioni da registrazioni dell'"ambiente sonoro linguistico" a cui sono esposti bambini molto piccoli. "Volevamo avere un esempio del tipo di stimolo nel quale si sviluppa il linguaggio dei bambini", ha spiegato Saksida, "Ci chiedevamo se un meccanismo di basso livello come la probabilità transizionale funzionasse su stimoli linguistici realistici, molto diversi dagli stimoli costruiti a tavolino che si usano normalmente in laboratorio, che sono più schematici e non contengono fonti di 'rumore'". Saksida e colleghi hanno usato corpus di ben 9 lingue diverse, e vi hanno applicato due diversi modelli basati sulla PT.

Prima di tutto sono stati calcolati i valori di PT in ogni punto del flusso linguistico per tutti i corpus usati, successivamente è stata effettuata la "segmentazione" con due diversi metodi. Il primo si basava sulle soglie assolute: veniva stabilito un certo valore fisso di riferimento per la probabilità

transizionale sotto al quale veniva identificato un bordo. Il secondo metodo invece si basava su soglie relative: i bordi corrispondevano ai minimi locali della funzione della PT.

In tutti i casi, hanno osservato Saksida e colleghi, la probabilità transizionale si è dimostrata uno strumento efficace per la segmentazione (si va dal 49% all' 86% di parole identificate correttamente), indipendentemente dal metodo di segmentazione usato. Questo ne conferma l'efficacia. Da notare che entrambi i modelli si sono dimostrati mediamente efficienti, la cosa singolare però è che quando un modello andava particolarmente bene con una lingua, quello alternativo allora andava sempre significativamente peggio.

"Questo 'incrocio' ci suggerisce che ciascun modello è più adatto per alcune lingue dell'altro e viceversa. Abbiamo quindi condotto delle analisi ulteriori per capire quali caratteristiche della lingua determinano il modello migliore", ha spiegato Saksida. La dimensione cruciale si è rivelata il ritmo linguistico. "Possiamo dividere le lingue europee in due grandi gruppi per quel che riguarda il ritmo: le lingue basate sull'accento (stress-based) e quelle basate sulle sillabe (syllable-based)". Le lingue basate sull'accento hanno meno vocali e parole più corte, come l'inglese, lo sloveno, il tedesco. Quelle basate sulle sillabe sono invece più ricche di vocali e con le parole in media più lunghe, come italiano, spagnolo, finlandese. Il terzo gruppo ritmico delle lingue, che non esiste in Europa e che e bassato su "morae" (un parte della sillaba), come giapponese, si chiama "mora-based" ed e ancora più ricca di vocali delle lingue bassate sulle sillabe.

Il metodo della soglia assoluta è risultato funzionare meglio sulle lingue stress-based, mentre la soglia relativa è migliore per le mora-based. "È possibile dunque che il sistema cognitivo impari a usare il metodo di segmentazione migliore per la lingua madre, che porterebbe difficoltà però con le lingue che non appartengono alla stessa categoria ritmica. Serviranno ovviamente studi sperimentali per verificare quest'ipotesi. Sappiamo dalla letteratura scientifica che i bambini subito dopo dalla nascita usano già l'informazione sul ritmo, e pensiamo che le strategie per scegliere la segmentazione più giusta potrebbero essere una delle aree dove l'informazione sul ritmo è più utile".

Lo studio infatti non può dire se il sistema cognitivo (sia adulto che bambino) usi davvero questo tipo di strategie. "Il nostro studio conferma in maniera chiara che questa strategia funziona attraverso un ampio spettro di lingue", conclude Saksida. "Servirà ora da guida per gli esperimenti in laboratorio."

LINK UTILI:

- **Articolo originale: http://goo.gl/cOk5VD**

IMMAGINI:

- **Crediti: Jev55 (Flickr: https://goo.gl/yVVdJ3)**

Contatti:

Ufficio stampa:
pressoffice@sissa.it
Tel: (+39) 040 3787644 | (+39) 366-3677586

via Bonomea, 265
34136 Trieste

Maggiori informazioni sulla SISSA: www.sissa.it