Morfologia e DSS
(Denoising Source Separation)

Basilio Calderone
(*work in progress*)

Obiettivo primario del lavoro è l'indagine di possibili meccanismi di organizzazione del dato linguistico che strutturano, in linea teorica, il lessico mentale di un parlante. Tali meccanismi devono rispondere sia ad esigenze di natura cognitiva (si pensi, ad esempio, ai principi di economia cognitiva per la rappresentazione e il trattamento del dato), sia a vincoli strutturali e composizionali del dato linguistico stesso (frequenza, analogie formali) [2]. Lo studio ha visto l'impiego di algoritmi per l'identificazione di fattori e componenti di dati statisticamente multidimensionali (Denoising Source Separation, DSS [6]). Tale metodologia, ancora in fase sperimentale, si è dimostrata promettente, ai fini simulativi, per la modellizzazione delle dinamiche che regolano l'apprendimento morfologico di dati verbali. Considerato come un vettore multidimensionale, il dato morfologico è inteso come risultato finale di un processo di combinazione in cui diversi componenti interagiscono tra loro in misura diversa. Tali componenti, individuati dall'algoritmo ad apprendimento ultimato, vengono identificati con gli attributi lessicali e grammaticali caratterizzanti una determinata forma verbale. In particolare, attributi P, paradigmatici (roots), e attributi G, morfosintattico-grammaticali (endings), emergono come "ingredienti primi" della forma flessa, vale a dire come i costituenti basici che definiscono la natura lessicale e grammaticale del verbo [5]. L'eliminazione della supervisione durante la fase di training (e quindi l'assenza di un "supervisore" che informi il sistema sul corretto output da esibire) garantisce un principio di elaborazione del dato che, prescindendo dal tipico controllo esogeno imposto dallo sperimentatore, si affida a dinamiche computazionali locali per cogliere e categorizzare le ricorrenze statistiche presenti nel corpus di apprendimento.

Gli attributi lessicali e grammaticali scoperti dal sistema si configurano

come una efficiente rete associativa [3] attraverso la quale le forme flesse memorizzate nel lessico mentale sono messe in relazione tra loro, sulla base di proprietà e attributi morfologici, in modo tale da garantire procedure di accesso all'informazione lessicale efficienti e robuste, e al contempo, interpretare e produrre parole nuove sulla base delle parole già acquisite. I fasci di attributi morfologici (paradigmatici e grammaticali), di natura emergente, appaiono inoltre sensibili ad effetti di frequenza.

In particolare si registrano effetti di *type* e di *token*. Si osserva infatti che il numero di attributi grammaticali appare proporzionale alla type frequency della classe e che tokens ad alta frequenza subiscono un maggior dettaglio di attributi, quasi una capillare focalizzazione, da parte del sistema, su quelle forme che appaiono più frequentemente. In un paradigma metodologico come quello adottato, finalizzato all'estrazione di features morfologiche, la nozione di "morfema" come unità linguistica autonoma, minima e composizionale si perde, a vantaggio di una rappresentazione distribuita della componente morfologica [1]. Recenti studi confermano la scarsa plausibilità psicolinguistica del morfema durante processi di apprendimento, optando per un componente morfologica fuzzy e poco discretizzata, definita da gradienti più che da confini di morfema [4].

I possibili sviluppi di questo lavoro sono molteplici. In primo luogo, intendiamo studiare in modo più analitico il tipo di dinamica che qui abbiamo appena delineato. A questo proposito, una selezione più accurata del corpus di addestramento, con particolare riferimento al rapporto numerico tra forme flesse ed esponenti lessicali, ci dovrebbe consentire di definire una vera e propria curva di apprendimento, espressa nei termini del numero di attributi (P e G) definiti dal sistema per ogni classe paradigmatica e grammaticale.

In secondo luogo è nostro intendimento raffinare la codifica numerica delle forme in input. Si è già avuto modo di osservare che alcuni attributi morfologici innaturali possono essere ricondotti a un effetto di falsa analogia indotta dalla casuale similarità tra caratteri distinti all'interno della parola-verbo. É ragionevole supporre che una codifica numerica più motivata sul piano linguistico (con le vocali che risultino più vicine tra loro nella codifica di quanto non lo siano rispetto alle consonanti) possa aiutare a risolvere questo problema.

References

[1] Anderson, Stephen R. 1992. *A-morphous morphology.* Cambridge University Press, Cambridge.

[2] Bertram, Raymond, Robert Schreuder & Harald Baayen. 2000. The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology: Memory, Learning, and Cognition,* 26:419-511.

[3] Bybee, Joan L. & Paul Hopper (eds.). 2000. *Frequency and the emergence of linguistic structure.* John Benjamins, Amsterdam.

[4] Hay, Jennifer & Harald Baayen. 2005. Shifting paradigms: gradient structure in morphology. *Trends in Cognitive Sciences,* 9(7):342-348.

[5] Laudanna, Alessandro, Simone Gazzellini & Maria De Martino. 2004. Representation of Grammatical Properties of Italian Verbs in the Mental Lexicon. *Brain and Language,* 90:95-105.

[6] Särelä, Jakko & Harri Valpola 2005. Denoising source separation. *Journal of Machine Learning Research*, 6:233-272.