USING DON QUIJOTE TO TRACK IDEAS IN A COMPLEX WORLD

E. Alvarez-Lacalle[1], B. Dorow[2], J.-P. Eckmann[3], and E.Moses [1]

(1) Dept. of Complex Systems, Weizmann Institute of Science, Rehovot, Israel.
(2) Dept. de Physique Théorique et Section de Mathématiques, Université de Genève, Switzerland.
(3) Institute for Natural Language Processing, University of Stuttgart, Germany.
(e-mail: enric@ecm.ub.es)

Thoughts and ideas are multi-dimensional and often concurrent, yet they can be expressed surprisingly well in a linear form by the translation into language. This fundamental transformation requires memory, and implies the existence of correlations, e.g. in written text. However, correlations in word appearance decay quickly, while previous observations of long ranged correlations using random walk approaches yield little insight on memory or on semantic context. We aim to uncover these correlations and to understand how complex ideas are transmitted and processed.
We look at generalized combinations of words that a reader is exposed to within a “window of attention” spanning about a hundred words. We define a vector space of such word combinations, and analyze its structure by looking at words that co-occur within the window of attention. Singular value decomposition of the co-occurrence matrix identifies a basis whose vectors point at specific topics, or “concepts” that are relevant to the text. As the reader follows a text, the “vector of attention” traces out a trajectory in the “concept space”. We find that memory of the direction is retained over long times, forming power law correlations. The appearance of power laws hints at the existence of an underlying hierarchical network. Indeed, imposing a hierarchy similar to that defined by volumes, chapter, paragraphs etc., succeeds in creating correlations in a surrogate random text that are identical to those of the original text. We conclude that hierarchical structures in text serve to create long range correlations, and to utilize the reader’s memory in re-enacting some of the multi-dimensionality of the thoughts being expressed.

[1] E. Alvarez-Lacalle, B. Dorow, J.-P. Eckmann, and E. Moses PNAS **103**: 7956-7961 (2006).