Mental lexicon: two sides of hierarchical organization from big data

András Szántó[1], József Venczeli[1], Domicián Kovács[1], László Kovács[2], Csaba Pléh[3], Dániel Czégel[4]

[1]Budapest University of Technology and Economics
[2]University of West Hungary, Szombathely
[3]Central European University, Budapest
[4]Institute for Cross-Disciplinary Physics and Complex Systems, Palma de Mallorca

[4]czegel_d@yahoo.com

Introduction

The issue of general and specific knowledge as mirrored in memory systems has been central in several models of human memory. The now traditional separation of episodic and semantic memory in this regard can be seen as a proposal to contrast time-coded individual and atemporal generic knowledge (Tulving, 1972). In this traditional vision, initially all knowledge is episodic, and they gradually become semantic by repetition and generalization. A new vision entertained e.g., by the natural pedagogy approach of Csibra and Gergely (Csibra & Gergely, 2009) claims that in certain teaching situations children and people in general learn about kinds and they look for generic information.

In this work we present and compare two independent methods of extracting the degree of generality/specificity of words solely from the statistics of text corpora and word association databases. Since these methods $i)$ do not use any a priori knowledge about the meaning of the words, $ii)$ rely on huge databases containing linguistic output of many individuals, these might serve as robust and objective methods to generate or even define generality/specificity scores for a large number and variety of words, which can be then used as a standard for psycholinguistic experiments.

Methods

Burstiness from text corpora

As observed and quantified originally by (Ortuño, Carpena, Bernaola-Galván, Muñoz, & Somoza, 2002), in any signal sequence, units that are occurring only in very specific contexts exhibit a much burstier appearance than units that have a broader, more general meaning. This can be explained by the hierarchical topical organization of signal sequences, more specifically, of text corpora (Altmann, Cristadoro, & Degli Esposti, 2012).

In order to quantify the degree of burstiness β of a given word, we followed a method introduced in (Altmann, Pierrehumbert, & Motter, 2009) by fitting a Weibull distribution to the inter-event time distribution $p_i(\tau)$ of word w_i. The Weibull distribution has the form

$$p\left(\tau;\langle\tau\rangle,\beta\right)\sim\tau^{\beta-1}e^{-a\tau^{\beta}}$$

with $a(\langle\tau\rangle,\beta)=\left[\frac{1}{\langle\tau\rangle}\Gamma\left(\frac{\beta+1}{\beta}\right)\right]^{\beta}$, which interpolates between the exponential distribution ($\beta=1$), corresponding to a homogeneous Poisson process (the given word appears in every slot with the same probability), and a power-law $p(\tau)\sim\tau^{-1}$ $(\beta\to 0)$, corresponding to a scale-invariant appearance of words with the fattest possible tail a normalized probability distribution can have.

In order to minimize fitting error, we first extracted the parameter $\langle\tau_i\rangle=L/f_i$, where L is the total number of words in the text and f_i is the frequency of the word w_i; and then we estimated the single remaining parameter β via a maximum likelihood method. Figure 1 illustrates the locations of words *away* and *aunt* in one of the texts.

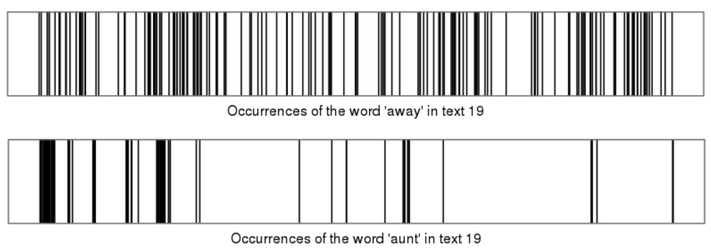

Figure 1: 'Barcodes', illustrating the occurrences of words *away* and *aunt* in the novel *Uncle Tom's Cabin* by *Harriet Beecher Stowe*.

Position in the hierarchical structure of word association network

Although the possible origins and underlying mechanisms producing hierarchical structures are yet to be understood, recently proposed methods quantifying the degree of hierarchy of a given system, modeled as a directed and weighted network, might be able to reveal the role played by different concepts in the mental lexicon.

The general idea behind these measures is composed of two steps: $i)$ one assigns a value p_i to every node i according to how much it is the source of any piece of information circulating in the network, $ii)$ the degree of hierarchy of the network is given by the variability of these p values.

Here we will only use step $i)$ of one such method (Czégel & Palla, 2015), in which the position p_i of a given node i in the hierarchy is determined in the following way.

- p_i is the stationary probability distribution of a random walk on the nodes of the network
- at every step, the transition probability $T(i\to j)$ is given by

$$T(i\to j)=\frac{w(i\to j)}{\sum_k w(i\to k)}\frac{w(i\to j)}{\sum_l w(l\to j)}$$

  where $w(i\to j)$ is the weight of the edge going from i to j.
- at the beginning of every step, we add an amount of $f=4$ random walkers to the nodes, uniformly distributed; then we make the transition using T; and finally we normalize the number of random walkers in the system again to 1. This way the random walkers avoid to get stuck in the sink nodes from which there are no outgoing edges at all.

In order to quantify the position p of words in the associative hierarchy, we applied this method to the South Florida word association network. Although the whole network is way too large to visualize, in Figure 2, the largest connected component of words having first letter *'a'* is shown.

Data

Text corpora. We have collected 100 longer novels written in or translated to English, written after year 1800. First, all the texts have been lemmatized by using the *Natural Language Toolkit* Python package (www.nltk.org). For each of the words (i.e., lemmas) w_i occurring at least once in any text, we computed all the inter-event times, defined as the number of words in the text between two subsequent occurrences of w_i. Then, we selected only those words that appeared in at least half of the texts AND have at least 1000 τ values assigned, in order to have meaningful fitting results. Finally, the histograms (i.e., empirical probability density functions) $p_i(\tau)$ for every word that meets the above criteria are computed.

Word association network. The stationary probabilities p were calculated on a free association network constructed from the South Florida Word Association database (SF) (Nelson, McEvoy, & Schreiber, 2004). The original database contains more than 10,000 words as cues and corresponding answers, gathered from over 6,000 participants, comprising a total of 72177 links, with weights given by the number of participants producing the same association. Thus the database describes a weighted directed graph, in which links point from cue words to replies. In the processing of the database we first excluded those words that occurred exclusively as replies, to prevent walkers from gathering disproportionately in dead-end nodes. This reduced the number of nodes to 5019, and the number of edges to 63621. Weights of remaining edges coming from the same node were divided by the sum of weights coming from the given node. This served to balance the differences between the number of participants each word were presented to, and the biases in weights introduced by removing half of the nodes.

Subjective abstractness scores. We used a subjective abstractness database (Brysbaert, Warriner, & Kuperman, 2014) containing subjective abstractness (SA) ratings for about 40 thousand English lemmas, collected via an internet-based crowd-sourcing method. The subjects are asked to rate the lemmas in a 5-point scale, where 1 is corresponding to the most "abstract (language-based)" items and 5 is corresponding to the most "concrete (experience-based)" items. The SA values of the lemmas are then computed as the mean of ratings given by on average 33.6 participants.

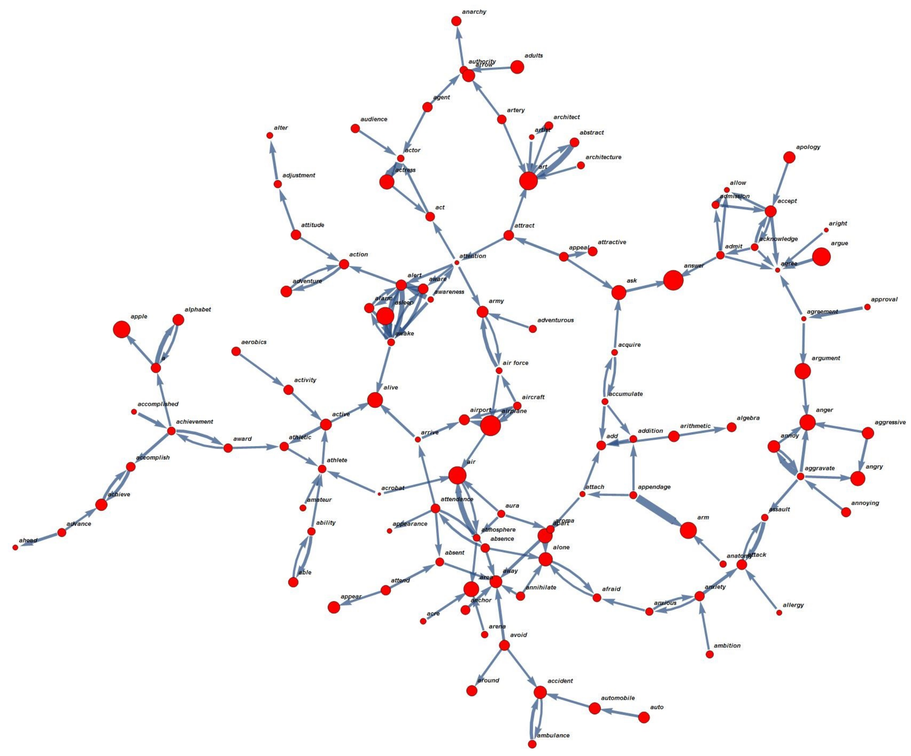

Figure 2: The largest connected component of the network formed by cue words having first letter 'a'. Node sizes are corresponding to the positions p in the associative hierarchy.

Results

We included in our analysis the 1049 lemmas that appear in all of the databases, i.e., have each of the β, p, f and SA values assigned.

The relationship between the lemmas' burstiness β, position in the associative hierarchy p, corpus frequency f and subjective abstractness score SA is shown in Figure 3, 4, and in Table 1.

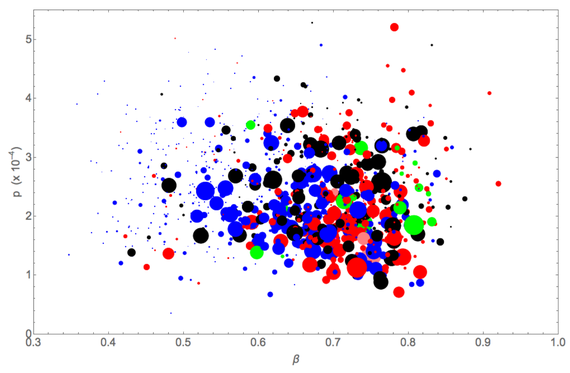

Figure 3: The words' burstiness β and position in the associative hierarchy p. Colors indicate their part of speech; the subjective abstractness scores SA are visualized as point diameters d via the relation $d\sim e^{-\mathrm{SA}}$, meaning that the larger the point, the more *abstract* the word is.

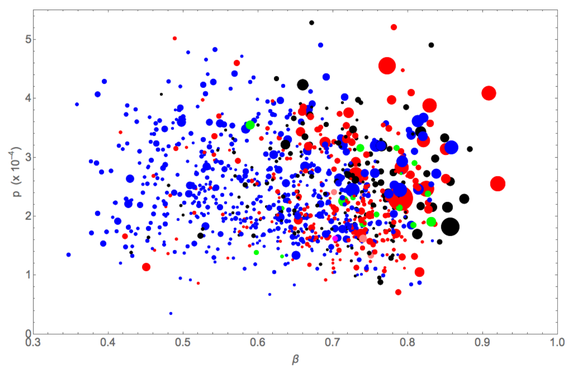

Figure 4: The words' burstiness β and position in the associative hierarchy p. Colors indicate their part of speech; the area of the points are proportional to the corpus frequency f of the corresponding words.

	β	p	$\log_{10} f$	SA
β		−0.12	0.25	−0.44
p			0.25	0.31
$\log_{10} f$				0.13
SA				

Table 1. Pearson correlations between the words' burstiness β, position in the associative hierarchy p, logarithm of the corpus frequency f and subjective abstractness score SA.

Discussion

- Burstiness β as a measure of *context-independence in texts* and position in the associative hierarchy p reflecting *how central the words are in the mental lexicon* are two nearly independent quantities.
- *Abstract words are not central* in the associative hierarchy. This quantitatively justifies an interesting phenomena: although abstract words usually have multiple or broader meaning $(r(\mathrm{SA},\beta)=-0.44)$ and have strong text-organizing role, they are rarely activated in word association tasks.

  This fact might also serve as an explanation for the lack of correlation between β and p: this relationship is a result of two competing effects, the broadness of contexts a given word can appear in, and its concreteness.
- Frequent verbs appear homogeneously in texts, and rare verbs appear inhomogeneously. On the other hand, for nouns, there is no such correspondence between frequency and burstiness. This suggests that frequent verbs (do, make, take, etc.) are *always* playing context-independent role, as opposed to frequent nouns, that can be context-independent and context-dependent equally well.

References

Altmann, E. G., Cristadoro, G., & Degli Esposti, M. (2012). On the origin of long-range correlations in texts. *Proceedings of the National Academy of Sciences*, *109*(29), 11582–11587.

Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS One*, *4*(11), e7678.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, *46*(3), 904–911.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in cognitive sciences*, *13*(4), 148–153.

Czégel, D., & Palla, G. (2015). Random walk hierarchy measure: What is more hierarchical, a chain, a tree or a star? *Scientific reports*, *5*.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407.

Ortuño, M., Carpena, P., Bernaola-Galván, P., Muñoz, E., & Somoza, A. M. (2002). Keyword detection in natural languages and dna. *EPL (Europhysics Letters)*, *57*(5), 759.

Tulving, E. (1972). Episodic and semantic memory 1. *Organization of Memory. London: Academic*, *381*(4).