Data in Brief 38 (2021) 107296

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Nonsemantic word graphs of texts spanning ~ 4500 years, including pre-literate Amerindian oral narratives

Natália Bezerra Mota[a,b], Sylvia Pinheiro[a], Antonio Guerreiro[c], Mauro Copelli[b], Sidarta Ribeiro[a,*]

[a] Instituto do Cérebro, Universidade Federal do Rio Grande do Norte, Natal, Brazil
[b] Departamento de Física, Universidade Federal de Pernambuco, Recife, Brazil
[c] Departamento de Antropologia, Universidade Estadual de Campinas, Campinas, Brazil

ARTICLE INFO

Article history:
Received 13 October 2020
Revised 11 May 2021
Accepted 12 August 2021
Available online 14 August 2021

Keywords:
Graph
Literature
Bronze age
Axial age
Language evolution

ABSTRACT

Non-semantic word graphs obtained from oral reports are useful to describe cognitive decline in psychiatric conditions such as Schizophrenia, as well as education-related gains in discourse structure during typical development. Here we provide non-semantic word graph attributes of texts spanning approximately 4500 years of history, and pre-literate Amerindian oral narratives. The dataset assessed comprises 707 literary texts representative of 9 different Afro-Eurasian traditions (Syro-Mesopotamian, Egyptian, Hinduist, Persian, Judeo-Christian, Greek-Roman, Medieval, Modern and Contemporary), and Amerindian narratives ($N = 39$) obtained from a single ethnic group from South America (Kalapalo, $N = 18$), or from a mixed ethnic group from South, Central and North America (non-Kalapalo, $N = 21$). The present article provides detailed information about each text or narrative, including measurements of four graph attributes of interest: number of nodes (lexical diversity), repeated edges (short-range recurrence), largest strongly connected component (long-range recurrence), and average shortest path (graph length).

DOI of original article: 10.1016/j.tine.2020.100142
* Corresponding author.
E-mail address: sidartaribeiro@neuro.ufrn.br (S. Ribeiro).

https://doi.org/10.1016/j.dib.2021.107296

Specifications Table

Subject area	Psychology
More specific subject area	Developmental and Educational Psychology
Type of data	Tables
How data was acquired	Computational graph analysis from historical texts acquired via internet and Amerindian oral narratives collected from direct interviews or available from the public domain
Data format	Excel tables with four mean graph attributes per text
Experimental factors	Literary texts in English, and Amerindian oral narratives transliterated to Portuguese converted to text files to perform graph analysis.
Experimental features	All word trajectories for sets of 30 words were represented as a graph (each word represented as a node and consecutive words linked by a direct edge). For each text or oral narrative, the mean number of nodes, repeated edges (RE), largest strongly connected component (LSC) or average shortest path (ASP) were calculated.
Data source location	Text/narrative identification included the digital libraries. Oral narratives were obtained from direct interviews, public corpora or publications.
Data accessibility	The data are within this article.
Related research article	Pinheiro et al. (2020) The History of Writing Reflects the Effects of Education on Discourse Structure: Implications for Literacy, Orality, Psychosis and the Axial Age Trends in Neuroscience and Education. Volume 21, 100142. https://dx.doi.org/10.1016/j.tine.2020.100142 [1].

Value of the Data

- Extensive and representative dataset including nonsemantic word graph analysis of texts in English spanning ~4500 years, from the first literate civilizations to contemporary literature, comprising nine different Afro-Eurasian traditions (Syro-Mesopotamian, Egyptian, Hinduist, Persian, Judeo-Christian, Greek-Roman, Medieval, Modern and Contemporary), as well as pre-literate Amerindian oral narratives. The information available here comprises: (1) documentation of the textual corpus, (2) graph representations of words from a representative text of the corpus, and (3) summary graph measures calculated for all texts in the corpus.
- Documentation of the textual corpus: The dataset includes detailed identification of a curated corpus of literary Afro-Eurasian texts (Table S1, $N = 447$); a non-curated corpus of post-medieval literary texts (Table S2, $N = 200$ texts from 10 randomly-chosen sets); a curated corpus of Amerindian oral narratives (Table S3, comprising data from a single ethnic group from South America (Kalapalo, $N = 18$); and data from a mixed non-Kalapalo group from South, Central and North America ($N = 22$), and a curated corpus of post-medieval Poetry (Table S4, $N = 60$ texts, with 20 per medieval, modern or contemporary period).
- Graph representations of a representative text contained in the corpus, the Book of the Dead, chapter 30 (Table S1). The data were analyzed using the same methodology across traditions, which consists in representing each word of a text as a node, and the sequence of words is represented by directed edges. In other words, given a graph $G = (N, E)$, N is the set of nodes composed by the different words in the text, $N = \{w_1, w_2, w_3, ...\}$ and $E = \{(w_i,w_j)\}$ is the set of edges between the nodes w_i in N and w_j in N, that are words in sequence in the text (Fig. 1A).
- Summary graph measures for all texts in the corpus: Table S5 contains graph measures from literary data and Table S6 contains graph measures from Amerindian oral narratives. This graph representation allowed the calculation of four mean graph attributes per text/narrative that correspond to structural features free from subjective evaluation: the

number of nodes, repeated edges, largest strongly connected component and the average shortest path (Fig. 1A).
- Table S7 summarizes the information presented in Tables S1–S6, and Table S8 lists the column names and provides a brief description of the corresponding variables in Tables S1–S6.
- The present dataset allows for the investigation of nonsemantic discourse structure during the Bronze and Axial Ages, and for the comparison with semantic data, as well as social, economic, and environmental data. as well as with remaining texts from Amerindian civilizations such as the Maya.
- The dataset can also be used for comparison with oral reports produced by people outside the mainstream industrial society, such as urban or rural illiterates, or isolated indigenous groups.

1. Data Description

The dataset here presented is freely available at https://osf.io/x7urg/ and contains graph measures applied to 707 literary texts and to 39 oral Amerindian narratives. Prose and Poetry texts written in English or translated to English were extracted from the public domain of the internet. To cope with computational cost, texts above 50,000 words were trimmed to this maximum; these texts are identified by the term 'cut'. Oral narratives from Amerindian participants were obtained from anthropological interviews. All the data were converted to the .txt file format and edited to perform graph analysis, which parsed the texts into windows of 30 consecutive words represented as a directed graph, assigning to each word a node, and to each consecutive sequence of words a directed edge, jumping 15 words to the next window. Each window had the edges randomized 100 times so as to generate 100 random graphs. For each original or randomized graph, four attributes were calculated to quantify non-semantic language structure (Fig. 1), such as lexical diversity (counted by the number of different nodes - N), short-term recurrence (counted by the repeated edges - RE), long-range recurrence (counted by the largest strongly connected component - LSC) and the graph length (counted by the average shortest path - ASP).

2. Experimental Design, Materials and Methods

This article provides graph measures (Fig. 1) calculated from literary texts, as well as Amerindian oral narratives. The data from this corpus were analyzed and discussed in [1].

Documentation of the textual corpus: The literary data were acquired from public domain virtual libraries such as the Digital Egypt of the University College London (http://www.ucl.ac.uk/museums-static/digitalegypt/), the Electronic Text Corpus of Sumerian Literature of the University of Oxford (http://etcsl.orinst.ox.ac.uk/), Project Gutenberg (www.gutenberg.org), and The Internet Classics Archive of the Massachusetts Institute of Technology (http://classics.mit.edu/). It comprises 447 texts from nine Afro-Eurasian traditions: Syro-Mesopotamian ($N = 62$), Egyptian ($N = 49$), Hinduist ($N = 37$), Persian ($N = 19$), Judeo-Christian ($N = 76$), Greek-Roman ($N = 133$), Medieval ($n = 20$), Modern ($n = 20$) and Contemporary ($N = 31$). Table S1 contains detailed information about this curated corpus, including title (and author if known), tradition, original language, date estimation including source of information, dating method used, date considered and the time interval considered, and the source link for each original text. To control for selection bias for the Medieval, Modern and Contemporary traditions, we analyzed a non-curated corpus of 10 independent sets of 20 randomly selected texts each ($N = 200$). Table S2 displays detailed information about the non-curated corpus regarding author, title, tradition, original language, dating method, estimated date, and set number. We also analyzed a curated corpus of transliterated oral narratives from Amerindian individuals not inserted on literate cultures ($N = 39$). The data were collected from directed interviews (data obtained by author AG under permit FUNAI #1712/09 from the National Indian Foundation), from a public corpus at the University of

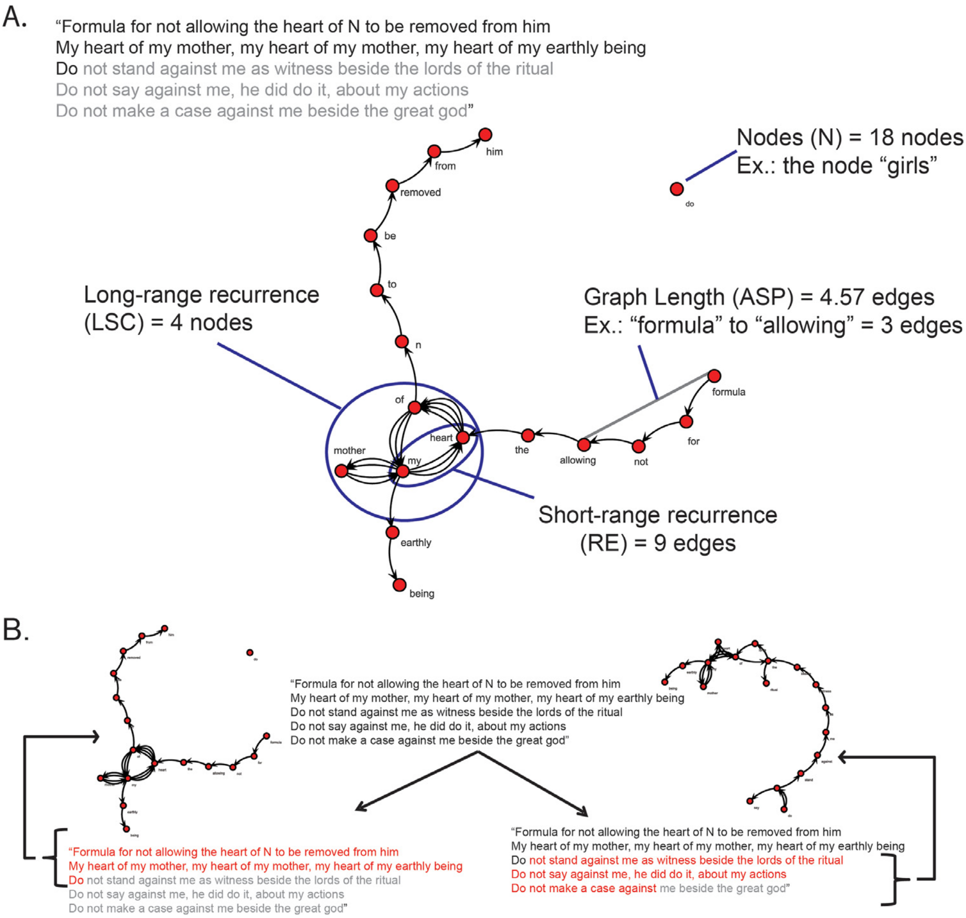

Fig. 1. *Nonsemantic word graph analysis of literature texts*. (A) From original text to graph, as previously implemented for historical texts and oral reports [1,2]. Example from the Book of the Dead, chapter 30 (**Table S1**). The graph attributes investigated comprised lexical diversity estimates by the number of nodes (N), long-range recurrence estimated by the largest strongly connected component (LSC), short-range recurrence estimated by the repeated edges (RE), and graph length estimated by the average shortest path (ASP). Red circles indicate nodes, black arrows indicate edges. **(B)** Moving windows of 30 words each, with 50% overlap between consecutive windows, were used to calculate mean values per graph for each attribute (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)**.**

Campinas ([3]), or from publications ([4-8]). The data were obtained from a single ethnic group (Kalapalo, recorded by author AG), and from a mixed group comprising participants from South America [3–6], Central America [7] and North America [8]. Table S3 displays information on the narrative content, ethnic group, language, region, researcher responsible for the data recording, source, participant's sex and age, and recording date. Finally, to control for structural profiles related to the difference between prose and poetry, we also analyzed a curated corpus of 60 poetic texts from Medieval, Modern and Contemporary traditions ($N = 20$ per tradition). Table S4 contains information regarding author, title, tradition, original language, date estimation including source of information, dating method used, date and the time interval considered, and the source link for each original text.

Graph representation and the summary graph measures for all texts in the corpus: All the literary texts were translated to English, converted to txt files and edited to remove notes, comments,

line breaks, prefaces, pages or tablet numbering or any publisher information. Oral narratives were transcribed and analyzed in their original language. Paragraphs were preserved. Given the txt files, the texts were represented as non-semantic graphs as illustrated in Fig. 1B, using the software *SpeechGraphs* available at https://neuro.ufrn.br/softwares/speechgraphs [9]. To normalize for differences in the number of words, and to represent graphs that would be comparable to a short oral report, moving windows of 30 consecutive words (with overlap of 15 words) were used. For each window we calculated the number of nodes (N) to estimate lexical diversity, the amount of repeated edges (RE) linking the same pair of nodes to estimate short-range recurrence, the number of nodes contained in the largest strongly connected component (LSC) to estimate long-range recurrence, and the average shortest path (ASP) linking all pairs of nodes to estimate graph length (Fig. 1A). Next, we averaged the graph measures from all the 30-word windows contained in each text or oral narrative to generate mean values (Tables S5 and S6). In addition, for each 30-word graph, 100 surrogated random graphs were calculated using the same words and the same number of edges, but a shuffled word sequence. Mean graph attributes were calculated for the random graphs as well (RE random, LSC random and ASP random). Table S5 contains graph measures from literary data and Table S6 contains graph measures from Amerindian oral narratives. A comprehensive description of similarities and differences across these datasets is available in [1].

Ethics Statement

Not applicable because all the data are text filed available in public domain.

CRediT Author Statement

S. Ribeiro: Conceptualization, Writing – reviewing & editing, Supervision. **N.B. Mota:** Visualization, Investigation, Writing – original draft preparation; **S. Pinheiro:** Data curation, Investigation, Writing – original draft; **A. Guerreiro:** Methodology, Data curation; **M. Copelli:** Methodology, Validation, Writing – reviewing & editing.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgments

We thank M Laub and JE Agualusa for source material; PPC Maia and S Morais for IT support; D Koshiyama, I Pereira and V Tollendal for documentation support, and the Instituto Metrópole Digital and the High-Performance Computing Center at UFRN (NPAD/UFRN).

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107296.

References

[1] S. Pinheiro, N.B.M. Mota, M. Sigman, D. Fernández-Slezak, A. Guerreiro, L.F. Tófoli, G. Cecchi, M. Copell, S. Ribeiro, The history of writing reflects the effects of education on discourse structure: implications for literacy, orality, psychosis and the axial age, Trends Neurosci. Educ. 21 (2020) 100142.
[2] N.B. Mota, M. Sigman, G. Cecchi, M. Copelli, S. Ribeiro, The maturation of speech structure in psychosis is resistant to formal education, NPJ Schizophr. 4 (2018) 25.
[3] C. Galves, A.L.d. Andrade, P. Faria, Tycho Brahe Parsed Corpus of Historical Portuguese, University of Campinas: Campinas, 2017.
[4] C. Albisetti, A.J. Venturelli, Enciclopédia Bororo, Volume II, (Lendas e Antropônimos) Ed. Faculdade Dom Aquino de Filosofia, Ciências e Letras, Instituto de Pesquisas Etnográficas, Campo Grande, 1969.
[5] A.P. Kamaiwrá, Uma análise linguístico-antropológica de exemplares de dois gêneros discursivos Kamaiurá [A linguistic-anthropological analysis of copies of two Kamayura discursive genres], Department of Linguistics, Portuguese and Classical Languages, University of Brasilia: Brasilia, 2010.
[6] E.P. Rosse, Dinamismo de objetos musicais ameríndios: notas a partir de cantos yãmĩy entre os maxakali (tikmũ'ũn). [On the dynamism of Amerindian musical objects: notes from yãmĩy chants among the Brazilian maxakali (tikmũ'ũn)], Per Musi 32 (2015) 53–96.
[7] J. Gómez de García, M. Axelrod, M.L. García, Sociopragmatic influences on the development and use of the discourse marker vet in Ixil Maya, in: J.M. Andrea L. Berez, Daisy Rosenblum (Eds.), Fieldwork and Linguistic Analysis in Indigenous Languages of the Americas, University of Hawaii Press., Honolulu, 2010, pp. 9–31.
[8] O.C. Lovick, Studying Dena'ina discourse markers: evidence from elicitation and narrative, in: J.M. Andrea L. Berez, D. Rosenblum (Eds.), Fieldwork and Linguistic Analysis in Indigenous Languages of the Americas, University of Hawaii Press, Honolulu, 2010, pp. 173–202.
[9] N.B. Mota, R. Furtado, P.P. Maia, M. Copelli, S. Ribeiro, Graph analysis of dream reports is especially informative about psychosis, Sci. Rep. 4 (2014) 3691.