Proceedings of the International Conference on Intelligent Information Systems
IIS'2013, August 20-23, 2013, Chisinau, Republic of Moldova

# Compactified HVG for the Language Network

D.V.Lande, A.A.Snarskii

Construction of networks with text elements, words, phrases or fragments of natural language as nodes in some cases allows one to detect the structural elements of the text critical for its connected structure and find informationally significant elements, as well as words that are secondary for understanding of the text. Such networks may also be used to identify unconventional text components, such as collocations, supra-phrasal units [1], as well as for finding similar fragments in different texts [2].

There is a multitude of approaches to constructing networks from the texts (so-called language networks) and different ways of interpreting nodes and links, which causes, accordingly, different representation of such networks. Nodes are connected if corresponding words are either adjacent in the text [3, 4], or are in a single sentence [5], or are syntactically [6, 7] or semantically [8, 9] connected.

At the intersection of digital signal processing (DSP) theory and complex network theory there are several ways of constructing networks from the time series, among those are visibility graph construction methods (see survey [10]), namely the horizontal visibility graph (HVG) [11,12]. Based on these approaches, networks can also be constructed from texts in which numeric values are assigned in some manner to each word or phrase. The examples of functions assigning a number to a word are: ordinal number of a unique word in a text, length of the word, "weight" of the word in a text, e.g., generally accepted TFIDF metric (canonically, a product of the term frequency in a text fragment and a binary logarithm of the inverse number of text fragments containing this word– inverse document frequency) or its modifications [13, 14] and other word weight estimates.

In this paper, the standard deviation estimate of word weight is used for constructing word networks [15]. If all the words in the text of $N$ words are numbered in succession (let $n = 1, \ldots N$ be the ordinal number

of the word in a text, the word position), layout of a certain word $A$ can be designated as $A_k(n)$, where $k = 1,2,\dots,K$ denotes the number of occurrence of this word in a text, and $n$ is a position of this word in a text. For example, $A_3(50)$ means that the third occurrence of the word $A$ has position 50 in the text.

The distance between successive occurrences of the word in these terms would be $\Delta A_k = A_{k+1}(m) - A_k(n) = m - n$, where $m$ and $n$ are the positions of the $k+1$-th and $k$-th occurrences of the word $A$ in the text, respectively.

Standard deviation estimate proposed in [15] is calculated as follows:

$$\sigma_A = \frac{\sqrt{\langle \Delta A^2 \rangle - \langle \Delta A \rangle^2}}{\langle \Delta A \rangle}, \quad (1)$$

where $\langle \Delta A \rangle$ is a mean value of the sequence $\Delta A_1, \Delta A_2, \dots, \Delta A_K$, $\langle \Delta A^2 \rangle$ is a mean value of $\Delta A_1^2, \Delta A_2^2, \dots, \Delta A_K^2$, and $K$ is a number of occurrences of the word $A$ in the text.

As opposed to other series examined in DSP theory, the series of numerical values assigned to words are transformed into horizontal visibility graphs (HVG), where each node not only has a corresponding numerical value, but also the corresponding word itself.

The process of constructing the language network using HVG consists of two stages. At the first stage, the traditional HVG is constructed [16]. To do that a series of nodes is put on the horizontal axis, where each node corresponds to a word in order of occurrence in the text, and standard deviation estimates are put on the vertical axis (visually a histogram, see Fig. 1). There is a connection between nodes, if they are in "line of sight" with each other, i.e., if they can be connected by a horizontal line that does not cross any other histogram bar. This (geometric) criterion can be written down as follows, according to [10,11]: the two nodes (words), e.g., $B_3(n)$ and $C_7$ $(m = n + 5)$, are connected if (see Fig. 1)

$$\sigma_n, \sigma_m > \sigma_p, \text{ for all } n < p < m. \quad (2)$$

The process of constructing can be algorithmized. For example, in Figure 1 the word node $A_1(n+2)$ is considered incident (and is connected with edges) to the words $B_3(n)$ and $C_1(n+5)$, $B_3(n)$ being the closest word to the left of $A_1(n+2)$ with a standard deviation estimate $\sigma_n = \sigma_B$ greater than that of the word $A$: $\sigma_{n+2} = \sigma_A$, and $C_7$

$(m = n + 5)$ being the closest word to the right of $A_1(n + 2)$, for which $\sigma_m > \sigma_A$.

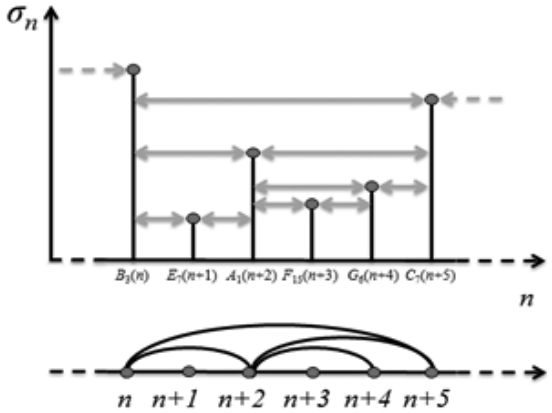


Figure 1. An example of HVG construction

At the second stage, the derived network is compactified. All the nodes corresponding to a single word, e.g., the word $A$, are combined into a single node (naturally, occurrence numbers and positions of the words are lost). The connections of these nodes are also combined. Note that there is no more than one edge left between any pair of nodes, multiple connections are removed (see Fig. 2).

This means, in particular, that the degree (number of connections) of the node $A$ does not exceed the sum of degrees $\sum_k A_k(n)$. As a result, the new network of words – *compactified horizontal visibility graph* (CHVG) – is constructed (Fig. 2).

Texts used for CHVG construction were the novels “The Master and Margarita” by Mikhail Bulgakov and “Moby-Dick; or, The Whale” by Herman Melville, as well as arrays of news information from the Web.

For all CHVG networks of words described here, the degree distribution is close to power law, i.e., these networks are scale free.

For comparison, was studied for the simplest language networks, where during the first stage of the network construction adjacent words were connected, and, at the second stage, the network was compactified. It is obvious that the weight of a node in such network corresponds to the word frequency, and the distribution of these weights follows the Zipf law

[18]. The most connected are the nodes corresponding to the most frequently occurring words – conjunctions, prepositions, etc., which are very important for the text coherence, but are of little interest for the aspect of informational structure.

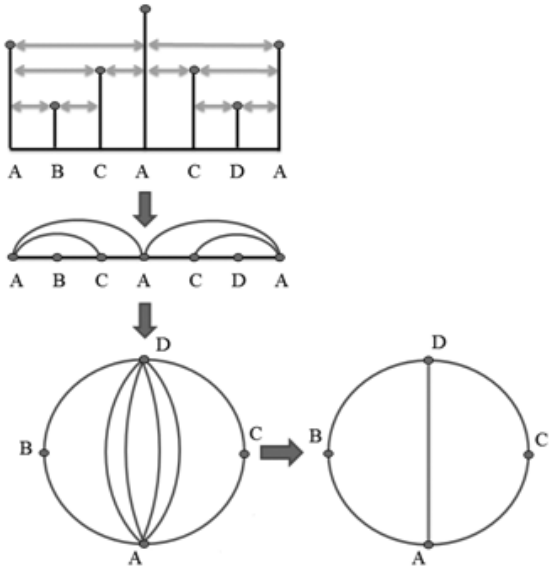


Figure 2. Two stages in construction of CHVG

Among the nodes with largest degrees, alongside with personal pronouns and other function words (particles, prepositions, conjunctions, etc.), there are the words, which determine the informational structure of the text [16, 17].

Let $\Psi$ be a set of $N$ different words (in our case $N$=100) corresponding to the largest-weight nodes of the aforementioned simple language network, and let $\Lambda$ be a set of words corresponding to the largest-weight nodes of the CHVG. Then the set $\Omega = \Lambda \setminus \Psi$ will contain informational words, which are also important for the text coherence. Appendix gives juxtaposition of the top 100 largest-weight nodes for the two types of language networks constructed from the novels “The Master and Margarita” by Michael Bulgakov and “Moby-Dick; or, The Whale” by Herman Melville.

In particular, the $\Omega$ set of the CHVG built from "The Master and Margarita" contains such words as Иван, Мастер, Варенуха, Берлиоз, Бегемот, Римский, профессор, Левий, Иешуа.

The following results were obtained from studying the language networks:

1. An algorithm for constructing compactified horizontal visibility graph (CHVG) was proposed.
2. Language networks were built from different texts based on series of standard deviation estimates and CHVG.
3. In CHVG obtained from literary works, among the largest-degree nodes there are words responsible not only for the coherence of the text, but also for its informational structure. They reflect the meaning of the mentioned texts.

## References

[1] Dijk van T.A. *Text and Context: Explorations in the Semantics and Pragmatics of Discourse.*–London:Longman. – 357p. (1977).

[2] Broder A. *Identifying and Filtering Near-Duplicate Documents,* COM'00 // Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching. – P. 1-10 (2000).

[3] Ferrer-i-Cancho R., Sole R. V. *The small world of human language* // Proc. R. Soc. Lond. – B 268, 2261 (2001).

[4] Dorogovtsev S.N., Mendes J. F. F. *Language as an evolving word web* // Proc. R. Soc. Lond. – B 268, 2603 (2001).

[5] Caldeira S. M. G., Petit Lobao T. C., Andrade R. F. S., Neme A., Miranda J. G. V. *The network of concepts in written texts* // Preprint physics/0508066 (2005).

[6] Ferrer-i-Cancho R., Sole R.V., Kohler R. *Patterns in syntactic dependency networks* // Phys. Rev. E 69, 051915 (2004).

[7] Ferrer-i-Cancho R. *The variation of Zipf's law in human language.* // Phys. Rev. E 70, 056135 (2005).

[8] Motter A. E., de Moura A. P. S., Lai Y.-C., Dasgupta P. *Topology of the conceptual network of language* // Phys. Rev. E 65, 065102(R) (2002).

[9] Sigman M., Cecchi G. A. *Global Properties of the Wordnet Lexicon* // Proc. Nat. Acad. Sci. USA, 99, 1742 (2002).

[10] Nunez A. M., Lacasa L., Gomez J. P., Luque B. *Visibility algorithms: A short review* // New Frontiers in Graph Theory, Y. G. Zhang, Ed. Intech Press, ch. 6. – P. 119 – 152 ( 2012).

[11] Luque B., Lacasa L., Ballesteros F., Luque J. *Horizontal visibility graphs: Exact results for random time series* // Physical Review E,– P. 046103-1–046103-11 (2009).

[12] Gutin G., Mansour T., Severini S. *A characterization of horizontal visibility graphs and combinatoris on words* // Physica A, – 390 – P. 2421-2428 (2011).

[13] Jones K.S. *A statistical interpretation of term specificity and its application in retrieval* // Journal of Documentation. – 28 (1). – P. 11–21 (1972).

[14] Salton G., McGill M. J. *Introduction to Modern Information Retrieval.* – New York: McGraw-Hill. – 448 p. (1983).

[15] Ortuño M., Carpena P., Bernaola P., Muñoz E., Somoza A.M. *Keyword detection in natural languages and DNA* // Europhys. Lett, – 57(5). – P. 759-764 (2002).

[16] DijkvanT.A. *Issues in Functional Discourse Analysis* / In H. Pinkster (Ed.), Liber Amicorum for Simon Dik. – Dordrecht: Foris. – P. 27) 46. (1990).

[17] Giora R. *Segmentation and Segment Cohesion: On the Thematic Organization of the Text* // Text. An Interdisciplinary Journal for the Study of Discourse Amsterdam. – 3. – № 2. – P. 155-181 (1983).

[18] Zipf G.K. *Human Behavior and the Principle of Least Effort*. – Cambridge, MA: Addison-Wesley Press. – 573 p. (1949).

D.V.Lande[1,2], A.A.Snarskii[1,2]

[1]Institute for information recording, NAS Ukraine, Kiev, Ukraine
[2]National Technical University “Kiev Polytechnic Institute”, Kiev, Ukraine
E-mails: dwlande@gmail.com, asnarskii@gmail.com