Searching Texts for Signs of Aphasia

Tatiana Jajcayová, Jozef Kubik, Mária Markošová, and Soňa Senkovičová

Dept. of Applied Informatics
Faculty of Mathematics, Physics and Informatics
Comenius University, Bratislava, Slovakia,
jajcayova@fmph.uniba.sk,
WWW home page: http://dai.fmph.uniba.sk/w/Introduction/en

Abstract: Aphasia is a brain disorder that impairs ability to speak or understand spoken language caused by damage of the brain. It is supposed, that it might also influence the vocabulary and complexity of the written texts of the affected people. In this paper, we analyzed two texts (one written before and the second after the onset of aphasia) for signs of aphasia using two different approaches. The first group of text analyzing methods are string matching algorithms able to find all occurrences of a pattern string in another text string. The second group of methods is from the complex networks theory. In scale free word webs, statistical measures are calculated, such as various types of averages and distributions. More modern approach is to analyze the graphlet structure of the word web. Both studies are applied to the same text data in searching for signs of aphasia.

1 Introduction

Aphasia is a brain disorder that often occurs after the brain damage. It is a partial or total loss of ability to speak or understand spoken language, especially if the brain areas responsible for language are affected. There are many areas in which speech may be impaired, and thus several types of aphasia – some patients know what the object is but cannot say the word for it, others replace words with unrelated ones in what would normally be a well understandable sentence and other ones may have trouble repeating heard words. We will be working with texts of an author with global aphasia.

Paul West was an American writer born in Britain in 1930. Over the course of his life he has written over fifty books in various genres – poetry, novels and essays alike. In the year 2003 he suffered a stroke and as a result global aphasia, he was not able to understand words and not able to speak them as well. However, after speech therapists failed to help him speak more than a few words, his wife, Diane Ackerman, proposed to him to write the first aphasic memoir. After three years the novel was finished, it's name is The Shadow Factory and this work is the subject of our research. To compare how aphasia changes the ability to write, we also analysed one of his previous books, written in 1983, The Rat Man of Paris. We were in communication with the late-writer's wife Diane Ackerman, and hoped to obtain the raw unedited texts of his aphasic novel. Up to this date, unfortunately, we have not received the raw texts. We decided to test our tools on published works instead.

We analyzed both texts from the two different viewpoints. The first group of text analyzing methods are string matching algorithms such as Rabin – Karp algorithm and approximate string matching algorithm. String matching algorithms are able to find all occurrences of a pattern string in another text string. In the experiments we searched for the short words which can be omitted in text due to the aphasia, such as "to be", "above", "below" etc. and also some short phrases. Both books were scaled in length to be comparable for this type of analysis.

The second group of methods is from the complex networks theory. It is known that the positional word web constructed from the English texts is a scale free network [6]. In scale free networks, various statistical measures are analyzed, such as various types of averages and distributions. A more modern approach is to study a graphlet structure of networks to compare them. The graphlet structure is then used to define measures, which are designed to express network structural differences and similarities.

In this paper, we studied the two above mentioned texts from both of this point of views, searching for similarities and differences.

2 String matching analysis

We tested the two texts using several different string matching algorithms. In this section we present the results by two different types of algorithms: the Rabin-Karp exact string matching algorithm and approximate string matching algorithm [3, 7, 11]. We tested the theory [4] that aphasia demonstrates itself in written text by omitting short words such as articles or some verbs, and thus the work written before the injury would have a greater number of short words than the later one. As we will see, the results of a series of our experiments for frequencies of short words in texts do not support this theory unambiguously. Later, we tested also some other characteristics of the texts, and we present those comparisons as well. In general, the string matching refers to the problem of finding all occurrences of a string, a pattern P[1..m] in another string, a text T[1..n]. Large portion of this problem is find-

ing ways to do so most efficiently with respect to time or computer memory. For more details see [3, 11].

2.1 Experiments with string matching

In the graphs below there are some interesting comparisons we found while searching for short words. Figures 1 through 7 show usage of various words that have a chance to be omitted in speech. We have found Rat Man of Paris is 1.36 times longer when comparing the number of words and has 1.4 times more characters. The following graphs show comparison of two texts where the number of words in The Shadow Factory has been scaled accordingly.

While Figures 1 and 2, where the frequencies of the usage of the forms of the verb 'to be' and of pronouns is tested by the Rabin-Karp algorithm, give mixed results, the Figure 3, which gives the comparison of the usage of prepositions of place, slightly supports the hypothesis that the later post-stroke work contains fewer short words of this kind.

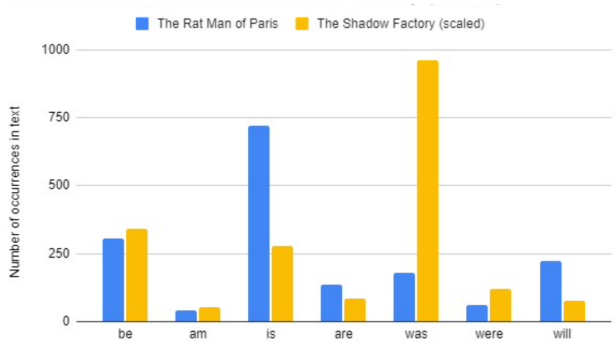

Figure 1: Use of forms of the verb "to be". RMP: The Rat Man of Paris, SF: The Shadow Factory.

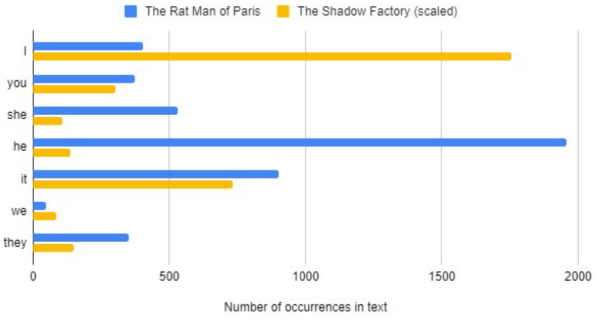

Figure 2: Use of pronouns.

While the Rabin-Karp and other exact string matching algorithms abort a pattern search when its characters stopped matching the text, approximate algorithms will do so only conditionally. We run several test using the approximate algorithm.

Before the search using the approximate algorithm, we specify an edit distance. This will tell the algorithm

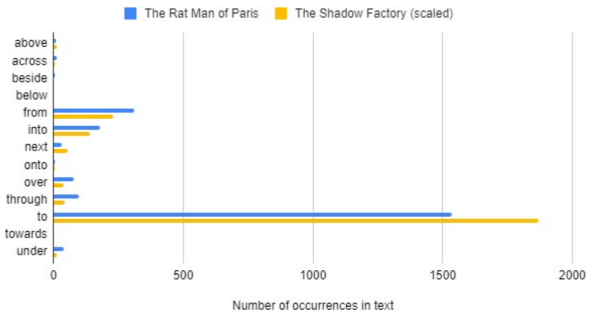

Figure 3: Prepositions of place.

whether it should continue pattern matching despite a mismatch or stop and move on. Edit distance is a number of alterations that can be made to either the pattern or currently searched part of the text in order to make them match. An alteration can be either substitution of a character for a different one, insertion of a character or removal of a character. A brute force approach would simply calculate an edit distance from the pattern P to all substrings of the text T. This is very demanding computationally. However, as it is with many problems, this one can too be solved using dynamic programming. First we make a two dimensional array L where the number of columns and the number of rows correspond to the lengths of compared strings w_1 and w_2. Each new entry in the array will be calculated from the currently compared characters and the values before. For cell $L_{i,j}$, containing the minimum number of alterations needed to match $w_1[1..j]$ with $w_2[1..i]$, we compute the value by comparing the next characters. If the characters are the same no new alterations are needed, and we take the value from $L_{(i-1),(j-1)}$. If they are not the same, the total edit distance will be greater by one. If values $L_{i,(j-1)}$ and $L_{(i-1),j}$ are equal it signalizes that we would change a character. When $L_{i,(j-1)}$ is not equal $L_{(i-1),j}$ means we try inserting or deleting a character. To calculate our $L_{i,j}$ we take the lower value of the two and add one.

The same results as above for the exact algorithms are supported by testing the use of prepositions of place by the approximate algorithm, as seen on Figure 4.

The work on the novel was hard, lengthy and took three years to finish. Therefore, we also included tests comparing the initial parts of the novel to the concluding parts of the novel. The results (Figure 5) show that the frequency of the short words increased a little by the end of the novel.

We also thought interesting to compare the literary text of the novel with the text of the Preface written by the author for the publication of the book, with the assumption that the preface may be less guarded.

Reading both, pre-stroke as well as the after-stroke, novels of Paul West is a peculiar experience. However, the feeling is very distinct. In the after-stroke work some words stand out as "weird", out of place. As the author himself wrote in the Preface, those words are indeed in-

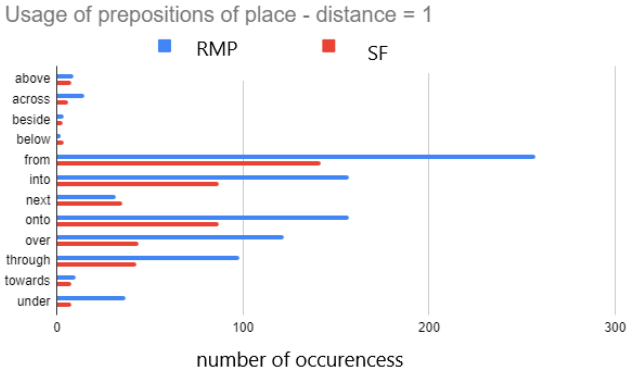

Figure 4: Prepositions of place - approximate algorithm.

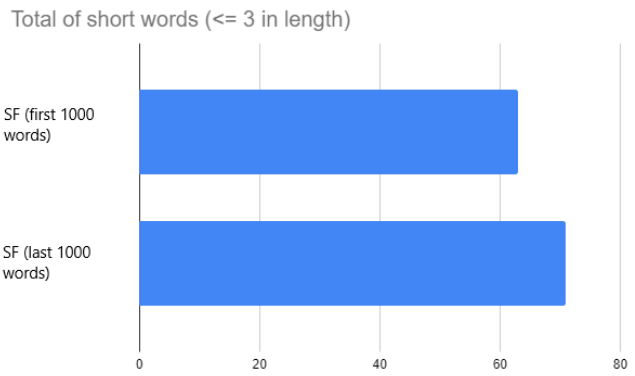

Figure 5: Total short words.

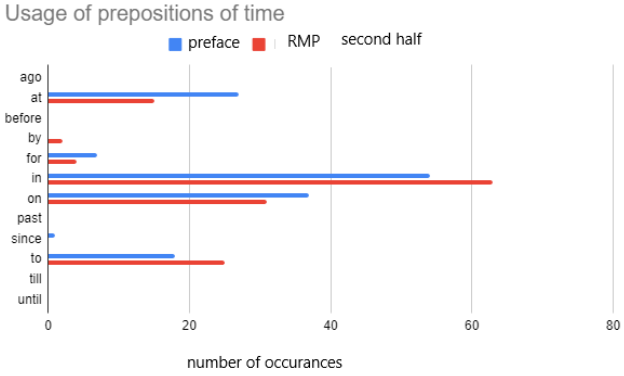

Figure 6: Prepositions of time; Preface vs. novel.

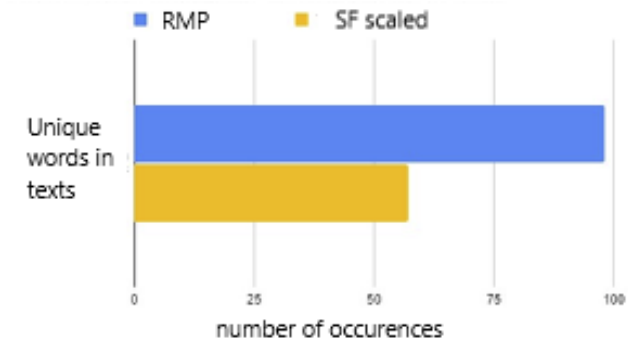

Figure 7: Unique words.

correct. He left them there on purpose, to show readers the state of his mind back then. We would love to quantify these differences in reading experiences. At the moment we do not know how, but our graphs in Figure 7 is a move toward this direction. The hypothesis is that West might have forgotten many short words due to his condition and that is why the number of unique short words in The Shadow Factory was significantly smaller. We observed this behaviour also when comparing the first and the last parts from The Shadow Factory. This time the part with heavier aphasic traits would be the one written first. The later part, indeed contains more unique words.

For the investigating aphasia disorders that impair the ability to use words in the right context, it seems that other more complex tools, like contextual language graphs, which capture the relationships of words in a text, will be needed. We present such experiments and results in the following section.

3 Word web analysis

It has been shown first by Barabási and Albert [1, 2], that the global structure of complex networks is influenced by the local processes of their development. Preferential node attachment leads to the scale free structure of the network, manifested in power law degree distribution

$$P(k) \propto k^{-\gamma},$$

where k is a degree and γ is scaling exponent. Variations of the local processes described above lead to the different final network structure [2, 5]. By scale free we mean that there is no internal scale in the network. For example, the situation is different in random graphs with Poisson degree distribution where there is an internal scale – an average degree. Which means in the random graphs number of nodes having the grater degree or smaller degree than average quickly decreases. This in not the situation in the scale free networks.

As we know, words in a language are not randomly distributed in texts. Their ordering reflects grammatical rules of the language in question. Some of the words are used very often, some of them are rather rare. Word webs enables one to look at the organization of words in a language differently. Positional word web expresses how words are organized in sentences, that means syntactic aspect of the language. Due to the scale free structure in all of the studied languages, it is supposed, that preferential attachment was involved in positional word web development.

Positional word web is created from the text as follows: Unique words (without respect to their grammatical form)

are nodes of the network. If one chooses word w, all words which are placed in the sentence before and after the word w anywhere in the analyzed text are connected by an edge with w. The punctuation marks are not taken into account, they are treated as if they are not included in the text. This process of word web creation has been suggested by Cancho and Solé [6]. The result is a connected graph, because all words in the text has at least one neighbour. It has been shown by Cancho and Solé [6] and by us [8], that such positional word web, created by the above described process and based on the English text is a scale free network and as such, can be analyzed as a graph.

We created two word webs, one from each text, with a help of our own application. In a special cases application is able to recognize shortened versions of words and replace them by a correct ones (such as 'd, which can be had, would, did, depending on the context). Application provides standard graph analysis and calculates number of nodes, edges, occurrences of words in the text etc., and also standard distributions such as degree distribution for example.

Graphlet structure of both networks was analyzed too. Usually one looks for connected graphlets. Graphlets are small nonizomorphic induced subgraphs consisting of from two to five nodes [9]. There are 30 such connected graphlets. Of course, one can take into account graphlets consisting of more nodes, but the number of such graphlets grows up very quickly, making calculations very time consuming. Therefore, the convention has been established, that speaking about graphlets, we have in mind the ones described above [9], [10].

To compare two networks one can calculate "relative graphlet frequency distance" (RGFD). Let $N_i(G)$ be the number of graphlets of the type i in the network G, and let $T(G)$ be the total number of graphlets of G. Thus, relative graphlet frequency distance $D(G,H)$ between the two graphs G and H is defined as: $D(G,H) = \sum_{i=0}^{29} |F_i(G) - F_i(H)|$, where $F_i(G) = -\log \frac{N_i(G)}{T(G)}$ [9]. The result is a positive real number. In general, two nets are similar, if this number is under 50. This is a rule of thumb introduced by Natasha Przulji in [10]. The motivation lies in the observations of distances of protein-protein interaction networks and corresponding model networks. Better comparison one gets using "graphlet degree distribution" (GDD).

By the graphlets, one can extend the concept of the degree distribution. The node degree $k(m)$ of the node m is a number of edges incident with the node in question and the degree distribution measures how many network nodes have the degree k. From the graphlet point of view, the degree k means that k graphlets of H_0 type (which is an edge) [9] touch (include) certain node. The same way we can look at another types of graphlets. However, in the graphlets it is topologically meaningful to distinguish at which automorphism orbit the node touches them. Two graph nodes belong to the same automorphism orbit, if there exists an automorphism which maps one onto the other. For example if we have a chain of tree nodes, the middle node is in a different automorphism orbit as the end ones, as no automorphism ever maps the middle one onto an end one. The end nodes belong to the same automorphism orbit. 30 different connected graphlets have 73 different automorphism orbits, so the correct analogue to the degree distribution is to measure the number of nodes touching particular graphlet at a node belonging to a particular orbit. Therefore, we get 73 graphlet degree distributions (GDD). Then one can compare two networks G and H calculating a measure called network GDD agreement (GDDA) [9]. There are two types of GDDA - s, namely $A_a(G,H)$ (arithmetic agreement) and $A_g(G,H)$ (geometric agreement). Here the result is a real number between zero and one. The closer to one the agreement is, the more similar the two networks are.

Properties of word webs	RMP	SF
Number of nodes	8422	6582
Number of edges	35054	25478
Maximal degree	2191	1595
Minimal degree	1	1
Average degree	8.324	7.742
Ratio of unique words	0.14305	0.15835
Density	0.00099	0.00118
Av. clustering coeff	0.384	0.378
Network diameter	8	7
Av. shortest dist.	2.92499	2.89893

Table 1: Properties of the two word webs before and after aphasia. RMP: The RatMan of Paris, SF: The Shadow Factory.

For the two networks G and H and the orbit j the j-th agreement $A^j(G,H) = 1 - D^j(G,H)$, where $D^j(G,H)$ is a j-th distance, is defined. The distance is given as follows: for each orbit j and the graph G the j-th GDD $d_G^j(k)$ is measured. If $j = 0$ one has a classical degree distribution. Then $d_G^j(k)$ is scaled as $S_G^j(k) = \frac{d_G^j(k)}{k}$. $S_G^j(k)$ is then normalized by $T_G^j = \sum_{k=1}^{\infty} S_G^j(k)$ giving normalized degree distribution $N_G^j(k)$. The j-th distance of the two networks G and H is given as $D^j(G,H) = \frac{1}{\sqrt{2}} \sum_{k=1}^{\infty} ([N_G^j(k) - N_H^j(k)]^2)^{\frac{1}{2}}$.

Arithmetic agreement is then defined as $A_a(G,H) = \frac{1}{73} \sum_{j=0}^{72} A^j(G,H)$. Geometric agreement is given as $A_g(G,H) = (\prod_{j=0}^{72} A^j(G,H))^{\frac{1}{73}}$.

3.1 Results

First, the statistics of the two word webs is depicted in the Table 1. From the Table 2 it seems that the second word web has more internally connected structure, because the graphlet ratios are systematically slightly higher, but there are no significant differences.

Types of graphlets	RMP	SF
two node	0.0988%	0.1176%
three node	0.00979%	0.01268%
four node	0.001988%	0.002713%
five node	0.000486%	0.000664%

Table 2: Ratios of the connected graphlet numbers, normalized by the number of all graphlets, connected and unconnected of the same size. The number of all graphlets of size n is $\binom{N}{n}$, N is the number of network nodes. RMP: The RatMan of Paris, SF: The Shadow Factory.

Measures	Comp. 1	Comp. 2	Comp. 3
RGFD	4.76126	196.043	187.868
arithm. GDDA	0.870742	0.634399	0.636399
geom. GDDA	0.8372	0.60967	0.616451

Table 3: RGFD a GDDA comparisons of our word webs and random graphs. Comparison 1 compares word webs of the two books in question. Comparison 2 compares word web of Rat man of Paris and random graph having the same number of nodes and edges. Comparison 3 compares word web of The shadow factory and random graph having the same number of nodes and edges.

RGFD of the word webs $D(G_1,G_2) = 4.76126$ indicates high similarity of the graphlet structure of both texts. No influence of aphasia is seen. But GDDA analysis is better for comparison, because it provides numbers in the (0,1) interval. We calculated both of the agreements. Geometric agreement is $A_g(G_1,G_2) = 0.8372$ and arithmetic agreement is $A_a(G_1,G_2) = 0.870742$. Here G_1 is a word web of the Rat Man of Paris, and G_2 the one of The Shadow Factory. Both agreements are far more closer to 1 then to 0 and indicate great similarity of the graphlet structure of both networks. Therefore we can state, that we did not find a significant influence of aphasia on the graphlet structure of our word webs. Also from the syntactic point of view, both texts are written by the same style, with the same basic vocabulary.

We also made a comparisons of both networks to the random graphs having the same number of nodes and edges. The results are in Table 3. We can see, that both word webs are far more similar to each other then to random graphs with respect to the graphlet structure.

Degree distributions of both networks show, that they are both scale free. Both degree distributions are linear in log-log plot, indicating power law (3) with the scaling exponent $\gamma = 2.1$. Aphasia does not influence the degree distribution of the word web at all.

4 Conclusion

4.1 String matching analysis

Results of experiments presented above are a part of the broader research project into the diagnosis of brain disorders that are demonstrated in language. Results of our tests using several exact and approximate algorithms do not decisively support the tested logopedical hypotheses. It is clear that use of not edited texts of authors with aphasia would be more revealing. Lajos Grendel, a Slovak writer and publicist, representative of Hungarian literature in Slovakia, is another author known to suffer aphasia. We hope to gain an access to his "raw" texts written before and after the injury and analyse them.

4.2 Word web analysis

As has been stated above, word web analysis does not show significant differences in networks due to aphasia. The reason might be twofold: first, it is known, that The Shadow Factory has been edited by writer's wife before publication. We made an effort to get the raw text, before edition, but without success. The second possibility lies in a fact, stated in the Introduction, that despite speech therapist's failure to teach writer to speak again, he gained a lot of his writing abilities in these three years of writing The Shadow Factory. Our analysis might confirm this, which could be an important result, after more detailed studies. It shows, that language can be totally lost in one respect due to aphasia, but can be possibly gained back by specialized targeted training.

References

[1] A. L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 3616

[2] R. Albert, A. L. Barabási, Statistical mechanics of complex networks, Rev. Modern Phys. 74 (2002) 47

[3] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, Introduction to algorithms, 2nd edition (2001), MIT press

[4] Z. Cséfalvay. Súčasný pohľad na diagnostiku a terapiu afázie, Česká a slovenská neurologie a neurochirurgie. – Roč. 70/103, č. 2 (2007), s. 118–128

[5] S. N. Dorogovtsev, J. F. F. Mendes, Evolution of networks, Adv. Phys. 51 (2002) 1079

[6] R. Ferrer I Cancho, R. Solé, The small world of human language, Proceedings of the Royal Society of London 268 (2001), 2261

[7] T. Jajcayova, Z. Majerikova, Note on Some Interesting Test Results for the Rabin-Karp String Matching Algorithm, ITAT 2018 : Information Technologies – Applications and Theory

[8] M. Markošová, Network model of human language, Physica A: Statistical Mechanics and its App. 387 (2008), 661

[9] N. Przulj, Biological network comparison using graphlet degree, Bioinformatics 23 (2007) 177

[10] N. Przulj, D. G. Corneil, I. Jurisica, Efficient estimation of graphlet frequency distribution, Bioinformatics 22 (2006) 974

[11] S. Shabnam Hasan, F. Ahmed, R. S. Khan, Approximate String Matching Algorithms: A Brief Survey and Comparison, International Journal of Computer Applications 120(8):26–31, June 2015.