VOL. 11, NO. 10, MAY 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

A PERFORMANCE OF TEXT STEGANALYTIC SYSTEM USING GENETIC-BASED METHOD

Roshidi Din[1], Faudziah Ahmad[1], H. S. Hussain[2], Shima Sabri[2], Nik Zulkarnaen Khidzir[3] and Muzaaliff Musa[4]
[1]School of Computing, CAS, Universiti Utara Malaysia, Sintok, Kedah, Malaysia
[2]Kolej Poly-Tech Mara, Cheras, Kuala Lumpur, Malaysia
[3]Faculty of Creative Technology and Heritage, Universiti Malaysia Kelantan, Kelantan, Malaysia
[4]Faculty of Information Science and Engineering, Management and Science University, Selangor, Malaysia
E-Mail: roshidi@uum.edu.my

ABSTRACT

In this paper, a consolidated view of genetic algorithm approach from the perspective of steganalysis system on text based environment is presented. Thus, this paper is tries to measures the detection performance based on genetic algorithm method and statistical method in order to classify the analyzed text as stego text. Three aspects such as time taken, average of cost function and average of mean and standard deviation have been used to measure the performance methods between statistical and proposed GA based. Experiments have shown that proposed of genetic algorithm method gets better performance than statistical method, especially in detecting a short analyzed text. Thus, a finding shows that the proposed genetic algorithm method on analyzed text is promising. For further work, it is suggested that the accuracy rates of detection process on larger sizes of analyzed text through other intelligent methods should be investigated.

Keywords: detection performance, steganalytic system, genetic-based method, text steganalysis.

INTRODUCTION

With the rapid growth of Internet communication, information protection needs has become a main issue in order to keep information itself electronically secure [1]. One of the advance promising researches on information protections for the next generation through untrusted communication channels is a steganology field. Compared to cryptology field, which manipulates the scrambles message of secret writing in cover channel completely meaningless, steganology field try to keeps the cover channel perceptually unchanged after hiding the message of the covered writing. Actually, steganology field is a complementary area of cryptology through the last two decade and has played a role in national or government secret communication [2- 4]. In speaking of steganology, there are two main branches which are steganography and steganalysis. Specifically, steganalysis is concerned with discovering and rendering useless information such as covert messages in given message [5].

In recent years, there has been an increasing interest on steganography and steganalysis environment in order to utilize the implementation of intelligent-based system. This is due to research potentials area particularly in measuring the undetectability of steganography system which is methods of message detection which are still under investigation, and the general detection of the steganalysis has not been devised [6]. Thus, at least three questions of steganalysis on steganography systems that have been found such as [7]:

- How to identify the analyzed message having the hidden information?
- How to find the embedded hidden information before being inserted into cover channels?
- How to justify the analyzed message having information inside them?

In the latest cloud communication, steganalysis can be divided into two domain areas, known as digital media and natural language as presented in Figure-1.

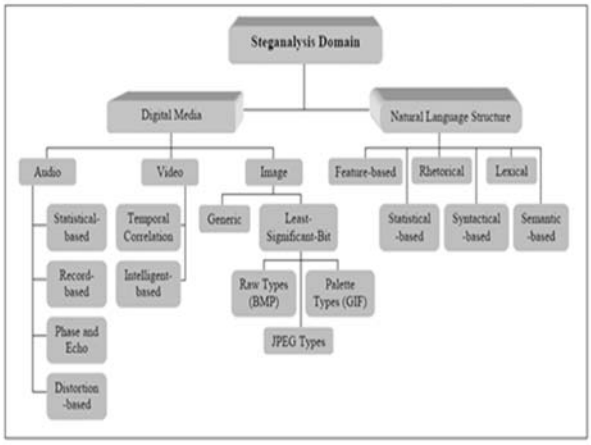

Figure-1. A modern classification of steganalysis areas.

Meanwhile, digital steganalysis are classified into three categories such as;

- audio steganalysis - three types of method involved are statistical-based, record-based, phase and echo, and distortion-based.
- video steganalysis - two kinds of method are identified as temporal correlation, and intelligent-based.

- image steganalysis - two methods are established as generic pixel and least-bit significant pixel types (e.g. .BMP. JPEG, and .GIF).

On the other hand, methods in natural language steganalysis can be divided into five categories such as features-based attack [8 - 11], statistical-based attack [12-15], rhetorical attack [16-21], syntactical-based attack [22 - 23], lexical attack [24], and semantic-based attack [25]. However, digital steganalysis are well established compared to natural language steganalysis which is still under exploration. This is because most of these natural language steganalysis methods are based on the text patterns analysis of natural language form. In addition, several steganalysis methods have been found utilized in natural language based environment [26 - 29]. Besides that, steganalysis methods on text based environment are still trying to find a good pattern combination of suspected hidden messages in the natural language text itself.

In the time being, many studies have been proven the usage of intelligent-based system such as genetic algorithm method on digital steganalysis environment in order to find a hidden message. Despite the different GA based system on digital steganalysis environment that have been proposed, the possibilities of using the GA based system in steganalysis for text based environment are under potential for exploring-utilized. Thus, this study believes that the formalization of GA based model is a major challenge due to developing a strong steganalytic system that can be applied intelligently on text based environment. Hence, a motivation of this study is to utilize a GA based model in order to produce a good model on text based environment. The rest of the paper is organized as follows: Section 2 describes a primary view of text based steganalysis. In Section 3, a cost function formulation on text is described. Discussion results of experimental work are done in Section 4 and finalize the work in Section 5.

A TEXT STEGANOGRAPHY SYSTEM

The basic concept of text steganography system is based on the idea of Prisoner's Problem [30]. It can be assumed that a text sender (known as Alice) and a text receiver (known as Bob) are imprisoned in different prison cell far separately from each other. Their communication is allowed through a gatekeeper [31] known as Wendy who plays the role of the adversary. If Wendy can identify any hidden information of their text communication, she will avoid or disturb all of the text communication. However, Alice and Bob are well alert of these circumstances. Therefore, Alice will try to send a hidden information or hidden message M, within a cover channel which is cover text C, which involves a key K through an embedding process and will produce a stego text S. Since S is a stego text, Wendy can't identify the stego text S and, Bob will be able to reconstruct the hidden information M and cover text C with a stego key K is presented in Figure-2.

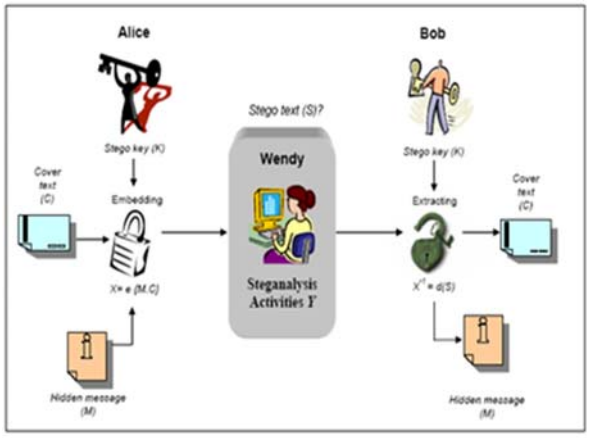

Figure-2. A communication process on steganography system between Alice and Bob.

Thus, Wendy needs a good formulation model for steganalytic system in order to analyze the exchange text of the hidden communication. This is because of one of the possible models is to utilize an intelligent based system such as GA based model on text environment in order to detect a hidden information.

A FORMULATION COST OF TEXT STEGANALYTIC GENETIC-BASED SYSTEM

The first step of genetic-based model is to establish an input text structure based on a genetic-based structure. It can be achieved through the fitness function evaluation of each analyzed text. Thus, a design of the genotype structure for input text sentences is one of the most difficult task and highly dependent on the analyzed text. Assuming that, the text genome, *Gi* from the analyzed text *T* is formulated as;

$$G_i = \text{text genome} = [T_1, T_2, T_3, \ldots, T_{NVar}] \quad (1)$$

Therefore, each chromosome has a cost value which can be found by evaluating the cost function f at analyzed text chromosomes (T_i) so that,

$$\text{cost} = f(\text{analyzed } Gi) = f(T_1, T_2, T_3, \ldots, T_{NVar})$$

Thus,

$$T_i = \sum_{i=1}^{N} G_n . \quad (2)$$

Then, it can be described that the analyzed text for each criterion $T_i = \{t_i, t_{i+1}, t_{i+2}, \ldots, t_{+n}\}$. Each input text structure known as chromosome has a cost value which can be found by evaluating the cost function f at analyzed text chromosomes T_i so that,

$$\text{Cost function} = f(\text{chromosome of analyzed text}) = f(T_i) \quad (3)$$

In such situation, the cost function is the different values between the desired and occurrences of the analyzed text and also defined as an error of detected text. Consequently, based on equation *(3)*, it can be claimed that:

$$f(T_i) = \sum_{n=1}^{\#words} (\text{desired\_text}_n - \text{occurred\_text}_n).$$

$$= \sum_{n=1}^{\#words} (dt_n - ot_n) \quad (4)$$

where

words number of the words in the text
dt_n word n in the desired text chromosome
ot_n word n in the analyzed text chromosome

Hence, a system tries to get reach at the minimum error of the detected text in order to get the best fit detection result. This value of minimum error detection is known as cost function of detected text, $cost_{min}$. Therefore,

$$\text{cost}_{\min}(f) = \sum_{n=1}^{\#words} \min (dt_n - ot_n) \quad (5)$$

A calculation of cost function values [32] is depending on maximum allowed evolution time and the number of population size. The cost function values are responsible for this evaluation that represents the best solution being the lowest number, or the lower the number the better the solution is.

EXPERIMENTAL RESULT

A development system uses HiddenStegoText.txt [33] dataset as shown in Table-1. Besides, a detection process uses an established Oxford dictionary in order to verify the words or sentences on analyzed text.

Table-1. HiddenStegoText.txt files obtained.

Analyzed Text	StegoText
T_1	theboywiththehorseisinthepark
T_2	thegirlwiththecarisinthepark
T_3	theboywiththehorsewasinthepark
T_4	thegirlwiththecarisintheriver
T_5	theboywiththeboatwasintheriver
T_6	theboysawthatthegirlraninthepark
T_7	thegirlsawthatthehorseraninthepark
T_8	theboysaidthatthegirlwasrunninginthepark
T_9	thegirlsaidthatthedukswamintheriver
T_10	theboyheardthatthegirlwasrunninginthepark

Actually, the system does not know what words or sentences included in each line of the sentences of an analyzed text before the detecting process is done. Once the system receives an analyzed text, it will pass on an analyzed text during detecting process. Three major aspects have been used during experiments study which are;

- time taken - to justify words or sentences on an analyzed text,
- average of $cost_{min}$ function - detected function values on analyzed text ,
- average of mean and standard deviation *SD* (σ) values of each analyzed text.

Thus, this study has done several experimental works on comparing the performance of fitness values between statistical method and GA based method. This comparison is based on the above aspects and indicates the result through Table-2 to Table-4 with Figure-3 to Figure-6.

The time taken of total detected words between statistical method and GA based method has shown in Table-2. It has been identified that 9 words to 11 words are found in the analyzed text. It is also detected that a statistical method is taking a longer time (6 seconds to 10 seconds) than GA based method (4 seconds to 7 seconds) during the detecting process. Figure 3 shows the comparisons of time taken on detecting words count between current statistical method and proposed GA based method.

Table-2. A time taken of word count and average of $cost_{min}$ function values on analyzed text.

Analyzed Text	Detected Word Count (No.)	Time Taken (sec.)		Average of $cost_{min}$	
		Stastical Method	GA Method	Stastical Method	GA Method
T_1	9	6.00	4.00	21.195	17.429
T_2	9	5.00	3.00	20.992	16.616
T_3	9	6.00	4.00	22.045	17.705
T_4	9	6.00	3.00	21.770	17.411
T_5	9	6.00	3.00	22.350	18.379
T_6	10	8.00	4.00	23.929	19.330
T_7	10	8.00	4.00	25.564	19.657
T_8	11	10.00	5.00	29.571	22.566
T_9	10	9.00	5.00	26.561	21.430
T_10	11	10.00	7.00	30.152	23.263

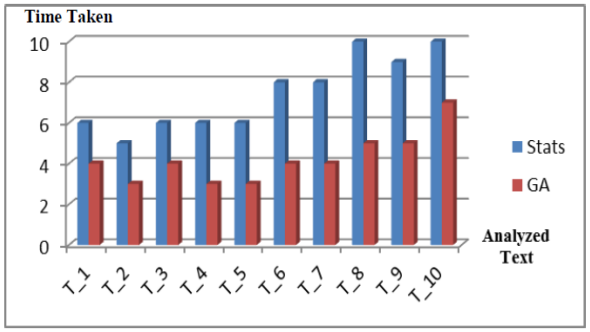

Figure-3. A distribution of time taken on analyzed text.

Table-2 has also shown the correct words or sentences with the average of $cost_{min}$ function levels after the detecting process of the hidden messages. From the results, both methods are predicting 100% of the hidden messages on analyzed text. However, it has been identified that a statistical method is taking more longer time (between 21 seconds to 32 seconds) compared to GA based method (between 16 seconds to 24 seconds) during finding the cost function average. Figure-4 shows a summary of time taken on finding the cost function average between current statistical method and proposed GA based method

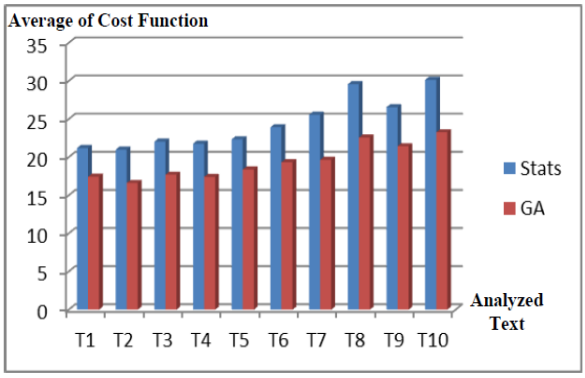

Figure-4. A cost function average between statistical and GA based methods on analyzed text.

The average of mean and standard deviation of the analyzed text has shown in Table-3. From the results, both methods produce good values of mean average. However, a statistical method performed between 0.73 to 0.76 is lower than GA based method compared to when it is performed between 0.99 to 0.999. It is also found that the average standard deviation values between statistical method and GA based method are having a wider range ratio. In comparison with the average of standard deviation of statistical method (between 0.1 to 0.2), a ratio of GA based method standard deviation is only between 0.0001 to 0.002. It shows that the distribution of GA based method is more accurate than the statistical method. In addition, the mean and standard deviation distribution between statistical method and GA based method have shown in Figure-5 and Figure-6.

Table-3. The average of mean and SD on analyzed text.

Analyzed Text	Average of Mean		Average of Standard Deviation - SD (σ)	
	Statistical based	GA based	Statistical based	GA based
T_1	0.7309	0.9988	0.1199	0.0009
T_2	0.7497	0.9986	0.1062	0.0011
T_3	0.7348	0.9980	0.1190	0.0018
T_4	0.7507	0.9987	0.1019	0.0014
T_5	0.7450	0.9982	0.1158	0.0013
T_6	0.7478	0.9979	0.1120	0.0016
T_7	0.7519	0.9973	0.1070	0.0026
T_8	0.7393	0.9977	0.1060	0.0023
T_9	0.7378	0.9981	0.1133	0.0016
T_10	0.7354	0.9985	0.1098	0.0015

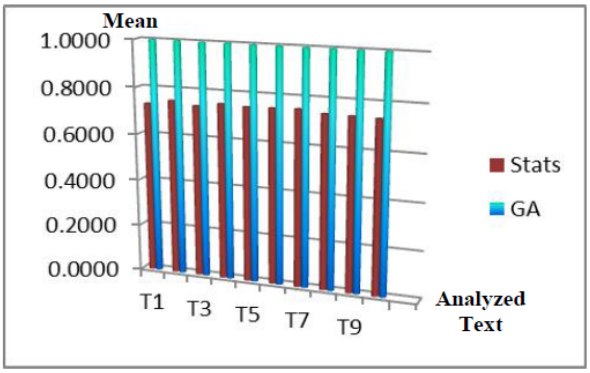

Figure-5. Mean values of analyzed text.

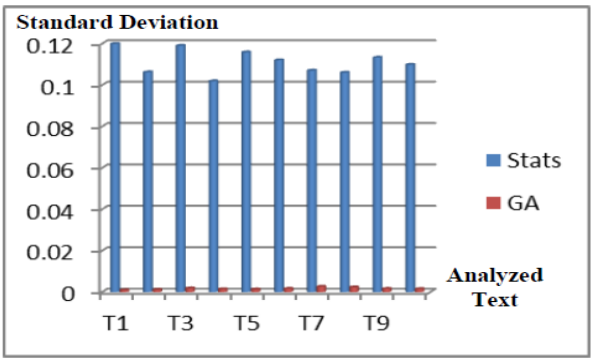

Figure-6. Standard deviation values of analyzed text.

CONCLUSIONS

The primary contribution of this paper is to present a proposed GA based model works which in return would contribute to text environment. The work presented is primarily working - to the best of the author's knowledge; this is among the earlier GA based model of the text steganalysis domain. It makes use of GA based method to justify a better performance compared to statistical method based on cost function values in order to detect a hidden message on an analyzed text. However, it depends on any steganalyst to use and choose the suitable model based on their purposes and its environment. For further work, the use of other intelligent model such as

neural network, fuzzy set selection, ant colony optimization, bee colony etc. should explored and investigated more.

REFERENCES

[1] Beale M. H., Hagan M. T. and Dewuth H. B. 2006. Fundamentals of machine and machine tools. 2nd Ed. Marcel Dekker Inc. New York, USA. pp. 160-168.

[2] R. Guduri, A. V. Rajulu and A. S. Luyt. 2008. Effect of alkali treatment on the flexural properties of hildegardia fabric. Journal of Applied Polymer Science. 10(2): 127-134.

[3] Caldas L. G. and Norford L. K. 1994. Screws, motors and wrenches that cannot be bought in a hardware store. In: Robotics Research: The First International Symposium. M. Brady and R. Paul (Eds.). pp. 679-693.

[4] M. A. Jassim, and M. A. Zulkarnain, Information hiding using LSB technique, International Journal of Computer Science and Network Security, 11(4), (April 2011).

[5] M. M. Amin *et al*., Information hiding using steganography, Proceedings, 4th National Conference on Telecommunication Technology, (2003), 21-25.

[6] S. Huayin, and L. Chang-Tsun, Maintaining information security in E-Government through Steganology, Global E-Government, (2004).

[7] G. Xiuhui, J. Renpu, and W. Jiazhen, Research on information hiding, US-China Education Review, 5(3), (May 2006).

[8] R. Chandramouli, and N. D. Memon, Steganography capacity: A steganalysis perspective, Proc. SPIE Security and Watermarking of Multimedia Contents, (2003).

[9] N. Johnson, and S. Jajodia, Steganalysis of images created using current steganography software, Proc. 2nd Information Hiding Workshop, Springer-Verlag, (1998), 273-289.

[10] D. Roshidi, and S. Azman, Digital steganalysis: computational intelligence approach, International Journal of Computers, (1), (2009), 161-170.

[11] X. Lingyun, S. Xingming, L. Gang, and G. Can, Research on steganalysis for text steganography based on font format, 3rd International Symposium on Information Assurance and Security, (2007), 490-495.

[12] H. Hua-jun, S. Xing-ming, S. Guang, and H. Jun-wei, Detection of hidden information in tags of webpage based on tag-mismatch, 3rd International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 1, (2007).

[13] H. Jun-wei, S. Xingming, H. Hua-jun, and L. Gang, Detection of hidden information in webpages based on randomness, 3rd International Symposium on Information Assurance and Security, (2007), 447-452.

[14] S. Xin-Guang, and L. Hui, A steganalysis method based on the distribution of space characters, Proceedings of Communications, Circuits and Systems International Conference, Guilin, Guangzi, China, 1, (June 2006), 54-56.

[15] I. Nechta and A. Fionov, Applying statistical methods to text steganography, CoRR, (October 2011).

[16] M. Peng, H. Liusheng, Y. Wei, C. Zhili, and Z. Hu, Linguistic steganography detection algorithm using statistical language model, International Conference on Information Technology and Computer Science, (2009), 540-543.

[17] L. Lingjun, H. Liusheng, Z. Xinxin, Y. Wei, and C. Zhili, A statistical attack on a kind of word-shift text-steganography, International Conference on Intelligent Information Hiding and Multimedia Signal Processing, (2008), 1503-1507.

[18] H. Huajun, Z. Shaohong, and S. Xingming, Steganalysis of information hidden in webpage based on higher-order statistics, International Symposium on Electronic Commerce and Security, (2008).

[19] Marcu D. Building up rhetorical structure trees. Proceedings of the 13th National Conference on Artificial Intelligence, Portland, Oregon, 2, (August 1996), 1069-1074.

[20] D. Marcu, The rhetorical parsing of natural language texts, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain, (July 1997), 96-103.

[21] D. Marcu, A decision-based approach to rhetorical parsing, Proceeding of the 37th Annual Meeting of the

Association for Computational Linguistics, Maryland, (June 1999), 365-372.

[22] U. Hermjakob, and R. Mooney, Learning parse and translation decisions from examples with rich context, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, (1997), 482-489.

[23] I. A. Bolshakov, and A. Gelbukh, Text segmentation into paragraph based on local text cohesion, Proceedings of the 4th International Conference on Text, Speech and Dialogue, Lecture Notes In Computer Science, Springer-Verlag, London, 2166, (2001), 158-166.

[24] F. Fukumoto, and Y. Suzuki, Extracting key paragraph based on topic and event detection - towards multi-document summarization, ANLP/NAACL Workshops, NAACL-ANLP 2000 Workshop on Automatic Summarization, Seattle, Washington, 4, (2000), 31-39.

[25] H. Yang, and X. Cao, Linguistic steganalysis based on meta features and immune mechanism, Chinese Journal of Electronics, 19 (4), (2010), 661-666.

[26] Z. Xinxin, H. Liusheng, L. Lingjun, Y. Wei, C. Zhili, and Y. Zhenshan, Steganalysis on character substitution using support vector machine, 2nd International Workshop on Knowledge Discovery and Data Mining, (2009), 84-88.

[27] C. M. Taskiran, U. Topkara, M. Topkara, and E. J. Delp, Attacks on lexical natural language steganography systems, Proceeding of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents, San Jose, (2006), 15-19.

[28] M. Alfonso, C. Justo, and A. A. Irina, Measuring the security of linguistic steganography in Spanish based on synonymous paraphrasing with WSD, 10th IEEE International Conference on Computer and Information Technology (CIT 2010), (2010), 965-970.

[29] Z. Chen, L. Huang, Z. Yu, L. Li, and W. Yang, A statistical algorithm for linguistic steganography detection based on distribution of words, 3rd International Conference on Availability, Reliability and Security, (2008), 558-563.

[30] Z. Chen, L. Huang, Z. Yu, W. Yang, L. Li, X. Zheng, and X. Zhao, Linguistic steganography detection using statistical characteristics of correlations between words, Information Hiding: 10th International Workshop, (2008), 217-220.

[31] M. Peng, H. Liusheng, C. Zhili, Y. Wei, and L. Dong, Linguistic steganography detection based on perplexity, International Conference on Multimedia and Information Technology, (2008), 217-220.

[32] Z. Chen, L. Huang, Z. Yu, X. Zhao, and X. Zheng, Effective linguistic steganography detection, IEEE 8th International Conference on Computer and Information Technology Workshops, (2008), 224-229.

[33] G. J. Simmons, Prison's Problem and the Subliminal Channel, Advances in Cryptology: Proceeding of CRYPTO 83. D. Chaum, ed. Plenum, New York, (1983), 51 - 67.

[34] D. Roshidi, and S. Azman, A conceptual framework for natural language steganalysis, Proceeding of 2011 4th IEEE International Conference on Computer Science and Information Technology (ICCSIT2011), Chengdu, China, 3(7), (June 2011), 264-268, ISSN 978-1-61284-833-4.

[35] D. Roshidi, S. Azman, T Zalizam T Muda, L. Puriwat, A. Amphawan and O. Nizam, "Fitness Value Based Evolution Algorithm Approach for Text Steganalysis Model", NAUN International Journal of Mathematical Models and Methods in Applied Sciences, 5, 7, (January 2013), 551-558.

[36] C. Lyon, The representation of natural language to enable neural networks to detect syntactic structures, PhD Thesis, University of Hertfordshire, (1994).