LinL:Lost in n-best list

Peng Meng[1,2,3], Yun-Qing Shi[2], Liusheng Huang[1,3], Zhili Chen[1,3], Wei Yang[1,3]
and Abdelrahman Desoky[4]

[1] NHPCC, Depart. of CS. & Tech., USTC, Hefei 230027,China
[2] New Jersey Institute of Technology, Newark, New Jersey, 07102, USA
[3] Suzhou Institute for Advanced Study, USTC, Suzhou, 215123,China
[4] CSEE, University of Maryland, Baltimore County, MD 21250, USA
mengpeng@mail.ustc.edu.cn

Abstract. Translation-based steganography (TBS) is a new kind of text steganographic scheme. However, contemporary TBS methods are vulnerable to statistical attacks. Differently, this paper presents a novel TBS, namely Lost in n-best List, abbreviated as LinL, that is resilient against the current statistical attacks. LinL employs only one Statistical Machine Translator (SMT) in the encoding process which selects one of the n-best list of each cover text sentence in order to camouflage messages in stegotext. The presented theoretical analysis demonstrates that there is a classification accuracy upper bound between normal translated text and the stegotext. When the text size is 1000 sentences, the theoretical maximum classification accuracy is about 60%. The experiment results also show current steganalysis methods cannot detect LinL.

Keywords: LinL, natural language steganography, translation-based steganography (TBS), text steganography, linguistic steganography

1 Introduction

The demand for translating fueled the necessity of machine translation (MT) systems in business, science, World Wide Web, education, news, etc. As a result, the popular use of MT by a wide variety of people creates a high volume of traffic for accessing and generating translation. Such huge traffic allows communicating parties to establish a covert channel to transmit steganographic covers and the adversary is impossible to investigate all of them. This renders translation an attractive steganographic carrier.

The core idea of Translation-Based Steganography (TBS) [1–3] is: “When translating a non-trivial text between a pair of natural languages, there are typically many possible translations. Selecting one of these translations can be used to encode information” [1]. So the methods to generate the various translations for a given sentence are very important for the security of TBS and its embedding rate.

Contemporary TBS methods have used many different machine translators and a post-processing pass to obtain various translations. The translations obtained by these methods are much different from each other, so the stegotexts generated by TBS are also much different from normal translated text. Consequently, Meng *et al.* [4] and Chen *et al.* [5] successfully got their methods (STBS and NFZ-WDA) to detect TBS.

Like the relation between cryptography and cryptanalysis, steganography and steganalysis is a cat-and-mouse game. Although the statistical methods (STBS and NFZ-WDA) seem to be promising on steganalysis of TBS, translated text is still an attractive steganographic carrier due to demand for translation. Because translated texts have been widely used on the Internet, using translated text as a covert channel will draw less attention. For example, the translators of Google [6], Systran [7], Linguatec [8], just name a few, are widely used on the Internet, and in Google's vision, people will be able to translate documents instantly into the world's main languages in the future. So it is attractive to research much securer TBS.

To enhance the security of TBS, the most important work is to obtain various and similar translations for each cover text sentence. We find the n-best list [9] is a promising method to generate the similar translations.

Generally, the machine translator just generates the best translation for a given input. However, the second best translation, third best translation, and so on, can also be generated according to the applications. The first "n" best translations are known as n-best list, which has been widely used for improving the quality of machine translation and automatic speech recognition [9].

The following is an example of n-best list which is generated by Moses [10], and the n-best list is compared with the translations by other on-line machine translators.

Listed below is a German sentence: hierbei handelt es sich nicht nur um einen statistischen fehler oder um glückliche umstände. Translating this sentence to English by Moses, the 5-best list and the translations from Google, Systran, Linguatec are:

1-best: this is no mere statistical error or lucky coincidence .

2-best: this is not mere statistical error or lucky coincidence .

3-best: this is not just statistical error or lucky coincidence .

4-best: this is not only of a statistical error or lucky coincidence .

5-best: this is not only a statistical error or lucky coincidence .

Google: This is not just a statistical error-or lucky circumstances.

Systran: here it does not only concern around a statistic error or happy would stand around itself.

Linguatec: this is not only a statistical fault or happy circumstances.

The example shows the sentences of the n-best list are more similar to each other than sentences from different translators. So using n-best list to improve the security of TBS seems to be feasible.

Therefore, this paper presents a novel TBS, namely lost in n-best list (i.e. LinL), which employs the n-best list to resist the current statistical detection. LinL just uses one Statistical Machine Translator (SMT) in the encoding process and selects one of the n-best list of each cover text sentence to encode the secret message. The difference between normal translated text and stegotext is defined by a mathematical model, and finally we give a theoretically maximum classification accuracy between normal translated text and stegotext. A series of experiments also performed to show current steganalysis methods cannot detect LinL.

The organization of this paper is as follows: Section 2 presents an overview of the related work. Section 3 briefly covers the basic operations of the TBS algorithm and some of the steganalysis methods. Section 4 focuses on the Statistical Machine Translation (SMT), and shows why n-best list is suitable for TBS. In Section 5, we use a mathematical model to define the difference between normal translated text and stegotext, and get a formula to compute the classification accuracy upper bound between normal translated text and stegotext. In Section 6 we present the results of using STBS and NFZ-WDA to detect LinL. Possible attacks on LinL are discussed in Section 7. Finally, Section 8 concludes the paper.

2 Related work

Text-based information, like web pages, academic papers, emails, e-books and so on, exchanged or distributed on Internet plays an important role in people's daily life. Because there are a huge number of texts available in which one can hide information, a covert communication known as linguistic steganography [11] has attracted more and more people's attention.

2.1 Linguistic steganography

Linguistic steganography is a text steganography method that specifically considers the linguistic properties when generated and modified text, and in many cases, uses linguistic structure as the space in which messages are hidden [11]. TEXTO [12] is an early linguistic steganography program. It works just like a simple substitution cipher, with each of the 64 ASCII symbols or uuencode from secret data replaced by an English word. Wayner [13] introduced a method which uses precomputed context-free grammars to generate steganographic text without sacrificing syntactic and semantic correctness. Chapman and Davida [14] gave another steganographic method called NICETEXT. The texts generated by NICETEXT not only had syntactic and lexical variation, but whose consistent register and "style" could potentially pass a casual reading by a human observer. Chang and Clark [15] introduced a method to integrate text paraphrasing into a linguistic steganography system.

Non-linguistic approaches to text steganography have also been researched. Liu and Tsai [16] proposed a steganographic method for data hiding in Microsoft Word documents by a change tracking technique. Desoky [17–20] has introduced a series of text steganography methods , which are named as noiseless steganography (Nostega).

2.2 Statistical steganalysis

For detecting the above linguistic steganography, some steganalytic algorithms have been proposed. Taskiran et al. [21] used a universal steganalytic method based on language models and support vector machines to differentiate sentences modified by a lexical steganography algorithm from unmodified sentences. Chen et al. [22] used the statistical characteristics of correlations between the general service words gathered in a dictionary to classify given text segments into stegotexts and normal texts. This method can accurately detect NICETEXT and TEXTO systems. The paper [23] also brought forward a detection method for NICETEXT, which took advantage of distribution of words. Another effective linguistic steganography detection method [24] uses an information entropy-like statistical variable of words together with its variance as two features to classify text segments.

3 Translation-base steganography and steganalysis

This section briefly presents an overview of the translation-based steganography (TBS). To introduce TBS, we focuse on the "Lost in Just the Translation (LiJtT)" [2] which extends the original "Lost in Translation (LiT)" [1] into one which allows the sender to only transmit the stegotext. The encoding processes of both LiT and LiJtT are selecting the translation results by various translators to encoding bits.

Conceptually, TBS works as follows: First, the sender obtain a cover text in the source language. The cover text could be a secret of the sender or could have been obtained from public sources — for example, a news website. Then, the sender translates the sentences in the source language into the target language using multiple different translators. Because a sentence translated by different translators may generate different translation results, the sender essentially creates multiple translations for each sentence and ultimately selects one of these to encode some bits of the hidden message.

The encoding process of LiJtT specifically works as follows. After generating multiple translations for a given cover text sentence, the sender uses the secret key (which is shared between the sender and receiver) to hash the individual translated sentences into bit strings. The lowest h bits of the hash strings, referred to as header bits, are interpreted as an integer $b \geq 0$. Then the sentence whose lowest $[h+1, h+b]$ bits corresponds to the bit-sequence that is to be encoded is selected.

When the receiver receives a translation which contains a hidden message, he first breaks the received text into sentences. Then applies a keyed hash to each received sentence. The lowest $[h+1, h+b]$ bits in this hash contain the next b bits of the hidden message. Figure 1 illustrates the protocol.

These methods to generate different translations for data hiding can be detected by statistical methods. Papers [25, 26] present the first steganalysis method on TBS, which needs to know the MT set and the source language of

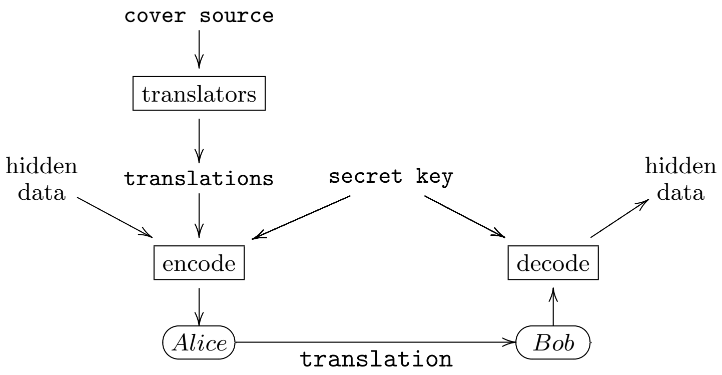

Fig. 1. Illustration of the basic protocol (from [2]). The adversary can observe the message between Alice and Bob containing the selected translation.

the cover text. Due to the source language and the translator set may be part of the private secret of the sender [2], the method cannot be used in general. To blind detection of TBS, Meng *et al.* [4] introduced a statistical steganalysis method which was named STBS. STBS is based on the word and 2-gram frequency difference between normal text and stegotext, the average classifying accuracy is about 80% when the text size is 20K bytes . To accurately detect TBS when the text size is much smaller, Chen *et al.* [5] gave another statistical steganalysis method, which is named natural frequency zoned word distribution analysis (NFZ-WDA). When the text size is 5K bytes, the detection accuracy is above 90%.

The steganalysis methods have demonstrated that the security of TBS is based on the methods to generate various translations. The more similarity between the translations, it is the more difficult to classify normal translated text and stegotext. The contemporary TBS uses different translators and a post-processing pass to generate the various translations for a cover text sentence. Because the translations resulted from different translators are much different to each other, Meng *et al.* [4] and Chen *et al.* [5] successfully introduced their methods to detect TBS. So it becomes clear that generating similar translations for the cover text sentence is pivotal for the security of TBS.

To generate the various and similar translations of a cover text sentence, n-best list of statistical machine translation (SMT) [9] seems to be a good strategy. To thoroughly study the security of using n-best list in TBS encoding process, we introduce the process of statistical machine translation.

4 Statistical Machine Translation

Statistical Machine Translation (SMT) as a research area started in the late 1980s. Lately, most competitive statistical machine translation systems use phrase-based translation [27].

SMT working process can be simply summarized as follows(by translating a different language to English as an example): For all the candidate English sentences of a foreign language sentence, SMT counts a probability cost for each of them and outputs the sentence with the highest probability cost as the translations.

Figure 2 illustrates the process of phrase-based translation.

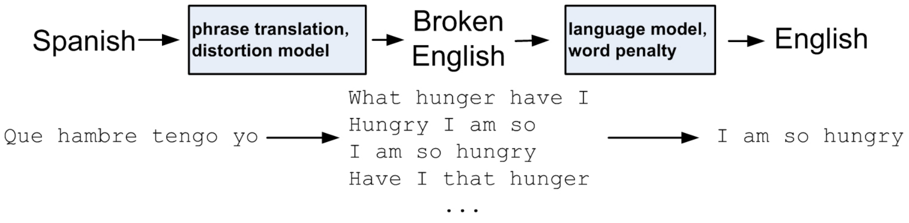

Fig. 2. An illustration of phrase-based translation.

The probability cost that is assigned to a translation is a product of the probability costs of four models: phrase translation table, language model, reordering model, and word penalty.

Each of the four models contributes information over one aspect of the characteristics of a good translation:

“The phrase translation table ensures that the English phrases and the foreign language phrases are good translations of each other.

The language model ensures that the output is fluent English.

The distortion model allows for reordering of the input sentence.

The word penalty provides means to ensure that the translations do not get too long or too short” [27].

Each of the models can be given a weight that sets its importance. Mathematically, the cost of translation is:

$$p(e|f) = \Phi(f|e)^{weight_\Phi} \times LM^{weight_{LM}} \times D(e,f)^{weight_d} \times W(e)^{weight_w}$$

The probability cost of the English translation e given the foreign input f, $p(e|f)$, is broken up into four models, phrase translation $\Phi(f|e)$, language model $LM(e)$, distortion model $D(e,f)$, and word penalty $W(e) = exp(length(e))$. Each of the four model is weighted by a weight [27].

To translate a sentence, the main process of SMT is to search the best translation from hundreds and thousands of candidate translations. An upper bound for the number of candidate English sentences can be estimated by $N \sim 2^{n_f}|V_e|n_f$ [27] where n_f is the number of foreign words of the translated sentence , and $|V_e|$ the size of the English vocabulary. Because the search space is very large, one can imagine that the best translation, the second best translation, the third best translation, and so on, will be very similar to each other.

Thus, the stegotext generated by TBS that is based on n-best-list would be difficult to be differentiated from normal translated text.

To validate the security of using n-best list in TBS, we provide both theory analysis and experiment study. In the next section, we give a theory analysis of using n-best list in TBS.

5 Theoretically analyze the security of LinL

In this section, we estimate the difference between normal translated text and stegotext by establishing a mathematical model, and we finally give a formula to compute the classification accuracy upper bound of LinL.

The translation process of SMT shows each candidate English sentence is associated with a probability cost, i.e., from SMT point of view each candidate English sentence is just treated as a probability, SMT just outputs the sentence with the highest probability as the translations. From the perspective of SMT, the probability cost is considered as the only feature of the translations. So the difference between the n-best list can be defined by the difference between each sentence's probability cost, and the difference between normal translated text and stegotext can be defined by the difference between their probability cost distributions.

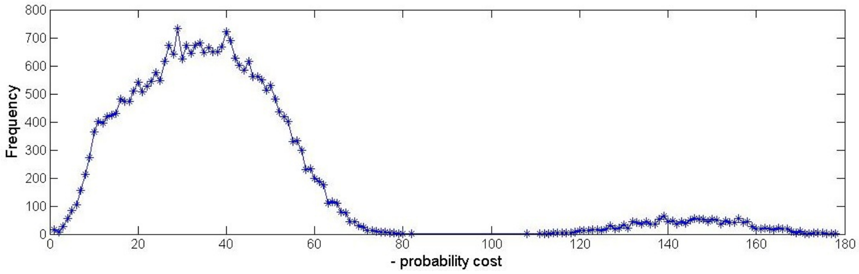

Fig. 3. The distribution of the probability cost of normal translated sentences.

Figure 3 shows the distribution of the probability cost of the normal translated sentences. Except some very high values, the distribution of the probability cost can be approximatively considered as normal distribution. Because the difference of the probability cost of n-best list is very small, he distribution of the probability cost of stegotext sentences can also be approximatively considered as normal distribution.

For a text segment which contains p sentences, there are totaly p probability cost features. Because each probability cost feature can be considered as a normal distribution variable, the vector of the p probability cost features can be considered as p-variate multivariate normal. The p-vector is the only measurement of the text. So the problem of classifying between normal translated text and stegotext is turned to the classification of two multivariate normal distributions.

Table 1. The means and variances of the probability cost of normal translated texts and stegotexts

Type	Ave	Var
normal	-44.49	42.89
Li2L	-45.16	42.77
Li4L	-46.12	44.99
Li8L	-46.79	47.85

Suppose the distributions of the probability cost of the normal translated texts and stegotexts are denoted by two normal distributions: $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$, where μ_1 and μ_2 are the means, and σ_1 and σ_2 are the variances of the first and second populations, respectively. The means and variances of normal translated texts and stegotexts can be obtained by a statistical method. Table 1 shows the means the variances of different type of texts that we have obtained from more than 10 thousands of sentences of each type. Li2L, Li4L and Li8L represent TBS with 2-best, 4-best and 8-best list to generate the stegotext, respectively.

Assume that the text contains p sentences and the probability cost of all sentences are independent, so the normal translated texts and stegotexts can be denoted by two p-variate multivariate normal distributions: $N(\mu_{p1}, \Sigma_1)$ and $N(\mu_{p2}, \Sigma_2)$, where,

$$\mu_{p1} = \begin{bmatrix} \mu_1 \\ \mu_1 \\ \vdots \\ \mu_1 \end{bmatrix} \qquad \mathit{and} \qquad \mu_{p2} = \begin{bmatrix} \mu_2 \\ \mu_2 \\ \vdots \\ \mu_2 \end{bmatrix}$$

are the mean vectors (each contains p values),

$$\Sigma_1 = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_1 & & \\ & & \ddots & \\ & & & \sigma_1 \end{bmatrix} \qquad \mathit{and} \qquad \Sigma_2 = \begin{bmatrix} \sigma_2 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_2 \end{bmatrix}$$

are the covariance matrices of the first and second populations, respectively.

The problem of classification of two multivariate normal distribution has been thoroughly researched in multivariate statistical analysis. For the two p-variate multivariate normal distributions, as defined above, the maximum classification accuracy can be computed by the following formula [28]:

$$Accuracy = \int_{-\infty}^{\sqrt{p}\frac{|\mu_1-\mu_2|}{\sigma_1+\sigma_2}} (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}t^2} dt$$

Using this formula to compute the maximum classification accuracy of normal translated texts and stegotexts, which only needs to know the means and variances of the probability cost.

With the data of Table 1, the maximum classification accuracy between normal translated text and stegotext can be couputed. Table 2 shows the maximum classification accuracy with the data of Table 1. From the data of Table 2, the following can be concluded:

- The classification accuracy increases with the text size increases.
- The less n-best list used in the TBS encoding process, the more secure for LinL.

Table 2. Maximum classification accuracy of LinL.

Length (Sen.) Type	100	200	300	400	500	600	700	800	900	1000
Li2L	0.53	0.54	0.55	0.56	0.57	0.58	0.58	0.59	0.59	0.60
Li4L	0.57	0.60	0.63	0.65	0.66	0.68	0.69	0.70	0.71	0.72
Li8L	0.60	0.64	0.67	0.70	0.72	0.73	0.75	0.77	0.78	0.79

6 Experiment

A series of experiments were performed to show the security of LinL. The experiments use the steganalysis methods which have successfully detected contemporary TBS to detect LinL.

Moses [10] was used to translate from German to English to generate the n-best list. The WMT08 News Commentary data set [29], about 55k sentences were used to train Moses and as the source text of the experiment. Li2L, Li4L and Li8L were tested. The normal translated texts and stegotexts were split to 10K bytes segment. STBS [4] and NFZ-WDA [5] methods were tested respectively. Table 3 shows the detection results.

The experiment results in Table 3 shows both STBS and NFZ-WDA cannot detect LinL. When using STBS to detect Li2L and Li4L, the detection accuracy is no better than random guess. Even using STBS to detect Li8L, the detection accuracy is still very low. When using NFZ-WDA to detect Li2L, Li4L and Li8L, respectively, it would classify most of the test texts to normal translated text.

Table 3. Experiment results of using STBS and NFZ-WDA to detect LinL

	Type	Train	Test	Non-stego	Stego	Accuracy(%)
STBS	Normal	50	229	155	74	51.02
	Li2L	50	212	142	70	
	Normal	50	229	99	130	48.49
	Li4L	50	169	75	94	
	Normal	50	229	133	96	61.36
	Li8L	50	110	35	75	
NFZ-WDA	Normal	50	229	224	5	51.02
	Li2L	50	212	211	1	
	Normal	50	229	194	35	54.02
	Li4L	50	169	148	21	
	Normal	50	229	178	51	59.29
	Li8L	50	110	87	23	

7 Discussion

This section discusses the various possible attacks on LinL. As one of the serial TBS methods, some the discussions about LiT [1] and LiJtT [2], like future machine translation and repeated sentence problems, are also suitable for LinL. We just discuss the problems that may come out with LinL in this section.

7.1 Translation quality

Whether the translation quality of stegotext is worse than normal translated text? From SMT point of view, some sentences of stegotext are not the best translation, but the second best translaion, third best translation, and so on, the answer is yes. However, translation quality is difficult to be used as a feature to classify a text to normal translated text and stegotext. First, the translation quality is difficult to count, and the translation quality of different machine translator or the same machine translator with different training database is much different. Second, the best translation given by a MT may not be the best translation from human's perspective. So using translation quality to attack LinL seems impossible.

7.2 Statistical attacks

Statistical attacks have been extremely successful at all area of steganography, such as image [30], video [31] and text [22]. We also cannot preclude the existence of yet-undiscovered statistical methods for defeating LinL. However, a classification accuracy upper bound between normal translated text and stegotet is given, it can be used as a reference when use LinL. For steganography and steganalysis, it is an arm race. Once a statistical steganalysis is known, it is actually easy to modify the steganography method to resist its attacks.

8 Conclusion

This paper introduces a novel translation based steganography, namely LinL, which uses the n-best list of a statistical machine translator (SMT) to encode the secret message. We just use one machine translator in the encoding process, the generated texts (stegotexts) of LinL are very similar to normal translated text, so it is difficult to classify normal translated texts and stegotexts. To show the security of LinL, we have derived a detection accuracy upper bound of LinL, and some steganalysis methods are tested on LinL, the experiment results show current steganalysis methods cannot classify normal translated text and stegotext.

Comparing with contemporary TBS, LinL can resist statistical detection and the embedding rate can be changed easily. Further more, LinL does not need post-processing algorithms either. To enhance the embedding rate, we can select a bigger "n" of the n-best list. To enhance the security of LinL, we just select a smaller "n" of the n-best list. However, if we just select the 1-best translation result, LinL will just be a normal translator.

The security of LinL maybe can continue to improve, for example, according to the sentence length or the probability cost of each translations, to select a different number of "n" for each sentence will be better for the security and embedding rate of LinL. This problem will be investigated in the future work. Although there is still some research work to be done for LinL, the theory analysis and experiment results shown have demonstrated that using n-best list to enhance the security of TBS is promising.

Acknowledgment

This work was partly supported by the Major Research Plan of the National Natural Science Foundation of China (No. 90818005) and the National Natural Science Foundation of China (No. 60903217).

References

1. Grothoff, C., Grothoff, K., Alkhutova, L., Stutsman, R., Atallah, M.: Translation-based steganography. In: Information hiding: 7th international workshop, IH 2005, Barcelona, Spain, June 6-8, 2005: revised selected papers, Springer-Verlag New York Inc (2005) 219–233
2. Stutsman, R., Atallah, M., Grothoff, K.: Lost in just the translation. In: Proceedings of the 2006 ACM symposium on Applied computing, ACM New York, NY, USA (2006) 338–345
3. Grothoff, C., Grothoff, K., Stutsman, R., Alkhutova, L., Atallah, M.: Translation-based steganography. Journal of Computer Security **17**(3) (2009) 269–303
4. Meng, P., Hang, L., Chen, Z., Hu, Y., Yang, W.: STBS: A Statistical Algorithm for Steganalysis of Translation-Based Steganography. In: Information Hiding Conference, 2010. IH'10., Springer-Verlag, Berlin Heidelberg (2010) 208–220

5. Chen, Z., Huang, L., Meng, P., Yang, W., Miao, H.: Blind Linguistic Steganalysis against Translation Based Steganography. In: International Workshop on Digital Watermarking, 2010. IWDW'10., Springer-Verlag, Berlin Heidelberg (2010) 251–265
6. Google: Google translator. (2009) http://translate.google.cn.
7. Systran: Systran translator. (2009) https://www.systransoft.com.
8. Linguatec: Linguatec translation. http://www.linguatec.de.
9. Chen, B., Zhang, M., Aw, A., Li, H.: Exploiting n-best hypotheses for smt self-enhancement. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, Association for Computational Linguistics (2008) 157–160
10. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Association for Computational Linguistics (2007) 177–180
11. Bennett, K.: Linguistic steganography: Survey, analysis, and robustness concerns for hiding information in text. Purdue University, CERIAS Tech. Report (2004)
12. Maker, K.: TEXTO. ftp://ftp.funet.fi/pub/crypt/steganography/texto.tar.gz.
13. Wayner, P.: Disappearing cryptography: information hiding: steganography and watermarking. Morgan Kaufmann Pub (2008)
14. Chapman, M., Davida, D.: Hiding the hidden: A software system for concealing ciphertext as innocuous text. Lecture Notes In Computer Science (1997) 335–345
15. Chang, C., Clark, S.: Linguistic steganography using automatically generated paraphrases. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics (2010) 591–599
16. Liu, T., Tsai, W.: A new steganographic method for data hiding in microsoft word documents by a change tracking technique. IEEE Transactions on Information Forensics and Security **2**(1) (2007) 24–30
17. Desoky, A.: Nostega: a novel noiseless steganography paradigm. Journal of Digital Forensic Practice **2**(3) (2008) 132–139
18. Desoky, A.: Listega: list-based steganography methodology. International Journal of Information Security **8**(4) (2009) 247–261
19. Desoky, A.: NORMALS: normal linguistic steganography methodology. Journal of Information Hiding and Multimedia Signal Processing **1**(3) (2010) 145–171
20. Desoky, A.: Matlist: mature linguistic steganography methodology. Security and Communication Networks
21. Taskiran, C., Topkara, U., Topkara, M., Delp, E.: Attacks on lexical natural language steganography systems. In: Proceedings of SPIE. Volume 6072. (2006) 97–105
22. Zhili, C., Liusheng, H., Zhenshan, Y., Wei, Y., Lingjun, L., Xueling, Z., Xinxin, Z.: Linguistic steganography detection using statistical characteristics of correlations between words. In: Information Hiding: 10th International Workshop, IH 2008, Sana Barbara, CA, USA, May 19-21, 2008, Revised Selected Papers, Springer (2008) 224–234
23. Zhili, C., Liusheng, H., Zhenshan, Y., Lingjun, L., Wei, Y.: A statistical algorithm for linguistic steganography detection based on distribution of words. In: Availability, Reliability and Security, 2008. ARES 08. Third International Conference on. (2008) 558–563

24. Zhili, C., Liusheng, H., Zhenshan, Y., Xinxin, Z.: Effective linguistic steganography detection. In: Computer and Information Technology Workshops, 2008. CIT Workshops 2008. IEEE 8th International Conference on. (2008) 224–229
25. Meng, P., Hang, L., Yang, W., Chen, Z.: Attacks on translation based steganography. In: IEEE Youth Conference on Information, Computing and Telecommunication, 2009. YC-ICT'09., IEEE (2010) 227–230
26. Meng, P., Hang, L., Chen, Z., Yang, W., Yang, M.: Analysis and detection of translation-based steganography. Chinese Journal of Electronics **38**(8) (2010) 1748–1752
27. Koehn, P.: MOSES, Statistical Machine Translation System, User Manual and Code Guide (2010)
28. Anderson, T., Bahadur, R.: Classification into two multivariate normal distributions with different covariance matrices. The Annals of Mathematical Statistics **33**(2) (1962) 420–431
29. WMT08: Wmt08 news commentary (2008) `http://www.statmt.org/wmt08/training-parallel.tar`.
30. Fridrich, J., Goljan, M., Hogea, D.: Steganalysis of JPEG images: Breaking the F5 algorithm. In: Information Hiding, Springer (2003) 310–323
31. Budhia, U., Kundur, D., Zourntos, T.: Digital video steganalysis exploiting statistical visibility in the temporal domain. IEEE Transactions on Information Forensics and Security **1**(4) (2006) 502–516