IJARnD
INTERNATIONAL JOURNAL OF ADVANCE RESEARCH AND DEVELOPMENT

(Volume 4, Issue 8)
Available online at: www.ijarnd.com

Text steganography in Letter of Credit (LC) using synonym substitution based algorithm

Aye Aye Chaw
University of Computer Studies, Mandalay, Myanmar

ABSTRACT

Text steganography uses text files as carrier files to hide confidential information. The goal of text steganography is to make the hidden data undetectable except the sender and receiver. This paper focuses on semantic-based text steganography. It hides confidential information by replacing words with their synonyms (same meanings). In international trading processes, products are imported and exported between foreign countries, in this case, the sellers and the buyers do not know each other. Therefore, the seller has a number of risks such as credit risk and legal risk caused by the distance. A letter of credit (LC) has been used to solve these risks. In the modern age, LC information has been sent online. Therefore, the security of LC information has become an important role. Most of the bank's LC information is sent via SWIFT message, email or message formats. Using these types of messages does not secure the LC information. This paper implements synonym substitution-based algorithm to make confidential LC information more secure. This system is implemented using the Java software development kit (J2SE 7).

Keywords— *Synonym, Obfuscation, Deobfuscation, Letter of Credit*

1. INTRODUCTION

In the age of ICT, more and more companies, organizations and individuals send and receive confidential information over the Internet. Therefore, the security of information becomes an essential requirement.

Steganography is a kind of information hiding technique that hides the existence of information by encapsulating it within some carrier-file. There are three main forms of text steganography: Structural, Random and statistical generation, and Linguistic. Structural based text steganography involves manipulating the structure of the text in order to hide data. Random and statistical generation involves generating the cover-text either randomly, or according to some function on an input. Linguistic steganography involves manipulating the syntactic or semantic properties of the existing text. Syntactic text steganography involves altering the structure of the text without significantly altering the meaning. In this paper, we apply Semantic-based steganography which involves replacing words with their synonyms (same meaning) in order to hide data. In this paper, the LC information sending and receiving between the issuing bank and the advising bank is made more secure by using synonym substitution-based algorithm. There are three sub-algorithms in this algorithm: synonym retrieval algorithm, obfuscation algorithm, and deobfuscation algorithm. The synonym retrieval algorithm retrieves synonyms or same meaning of cover-text from WordNet dictionary, obfuscation algorithm encodes the hiding message and deobfuscation algorithm decodes the hidden message. The synonym retrieval algorithm uses the British National Corpus (BNC) to choose the most likely word type and American National Corpus (ANC) to sort the set of synonyms (synset).

2. RELATED WORKS

In the modern age, steganography is used in many different areas such as releasing audio files, terrorists and criminal fields, hiding bank account details, spies, some pieces of software and so on. In releasing an audio file containing a new song to reviewers, the file is watermarked with the identity of the reviewer [1]. Then, steganography has increasingly being used by terrorists and criminals to hide data and exchange information. For example, a child pornography distributor could hide images in the photos on a legitimate eBay listing to distribute them to their customers [11]. Steganography is also used to hide bank account details in an image so when they are required, they can be retrieved, but if someone were to gain access to the computer, they would not be able to identify them [1]. There have been court cases in the US where accused Russian spies were found to have been using steganographic communication channels to communicate with their handlers [10]. There are also a number of pieces of software to perform steganography such as OpenStego [11].

3. MATERIALS AND METHODS

Steganography is concerned with hiding the existence of data by encapsulating it within some cover text. The goal is to make the hidden data undetectable, by man or machine. Modern steganography can be applied to a number of mediums. Text, images, audio and even video are all commonly used as carriers for secret messages [1]. Figure 1 shows the types of steganography.

3.1 Image Steganography

Image steganography uses Image file as cover file to hide message. Steganography is the image domain usually involves hiding data in the least significant bits at certain intervals in the image, for example one bit in each of the red, green blue values. In the transform domain, a transformation is applied to the cover image and then the data is hidden [4].

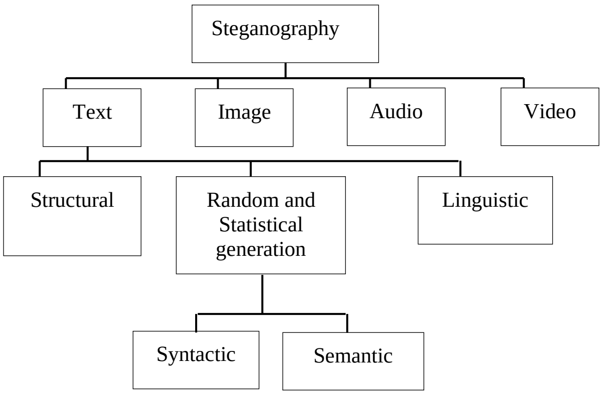


Fig. 1: Types of Steganography

3.2 Audio Steganography

In audio steganography, the data is hidden in such a way that the data is inaudible to the human ear. For example, humans cannot hear a tone that immediately follows a louder tone, so this is often used to hide data as using the most significant bit can be used to help overcome audio compression, without the original file sounding any different to humans. [5].

3.3 Video Steganography

Video can be used to hide data in much of the same ways as with images by hiding data in each individual frame. A famous example of this is a video found on a laptop owned by a suspected al-Qaeda member. The video, at first glance, appeared to be pornography but forensic investigators found that it contained 141 separate text files detailing terrorist operations and plans [6].

3.4 Text Steganography

Confidential data can be hidden in text files by using text steganography. There are three main types of text steganography: Structural, Random and Statistical Generation, and Linguistic.

3.4.1 Structural based text steganography: Structural based text steganography involves manipulating the structure of the text in order to hide data. The structure includes the line spacing, word spacing, font size and anything similar.[13].

3.4.2 Random and statistical generation based text steganography: There are two different ways to generate a cover-text to hide data: randomly that the words chosen are purely based on the data to be hidden, and statistical that the words are chosen to match some statistical criteria.[1].

3.4.3 Linguistic based text steganography: Linguistic Steganography is concerned with hiding information in natural language text. One of the major transformations used in Linguistic Steganography is synonym substitution[9]. Linguistic steganography comes in two forms: syntactic and semantic. Syntactic text steganography involves altering the structure of the text without significantly altering the meaning or tone. Semantic-based steganography is the focus of this paper. This method involves replacing words with their synonyms in order to hide data.

This paper implements semantic-based text steganography by using Synonym Substitution Based Algorithm. There are three sub-algorithms in the Synonym Substitution Based Algorithm: Synonym Retrieval Algorithm, Obfuscation Algorithm, and Deobfuscation Algorithm. The sending side uses Synonym Retrieval Algorithm and Obfuscation Algorithm to encode data, and then the receiving side uses Deobfuscation Algorithm.

3.5 ASCII

ASCII stands for American standard code for information interchange, which was developed by the American National Standards Institute (ANSI). It is the most common code used by computers to translate text (letters, numbers, and symbols) into a form that can be sent to, and understood by, other computers and devices such as modems and printers [14]. Figure 2 shows some of the standard ASCII characters and codes.

Char	Binary	Decimal
...	...	...
A	01000001	065
B	01000010	066
C	01000011	067
D	01000100	068
E	01000101	069
F	01000110	070
...	...	...

Fig. 2: Some of the standard ASCII characters and codes

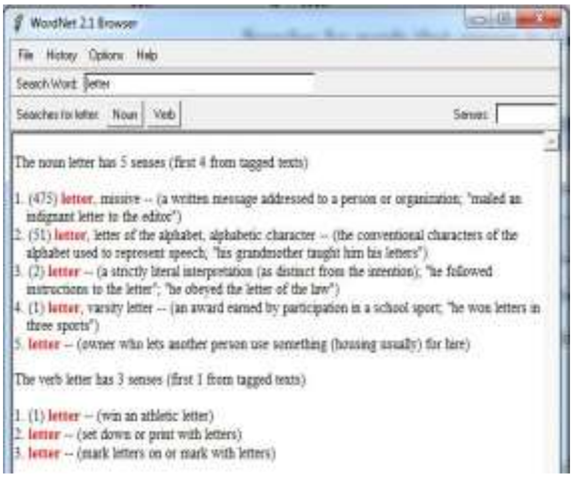
3.6 Synonym Retrieval Algorithm

This algorithm is used to retrieve the same or similar meanings of the cover text from the WordNet dictionary.

```
Data: word
Result: synset
synset = getsynonymsfromdictionary(word);
for i ← 0 to synsetsize do
    synset.addall(getSynonymsfromdictionary(synset[i]);
end
for i ← 0 to synsetsize do
    synset.setfrequency(getfrequency(synset[i-1], synset[i]);
end
synset.sort();
return synset;
```

3.6.1 Dictionary: The dictionary that is used in Synonym Retrieval Algorithm is the WordNet dictionary produced by Princeton University in the US [9]. The dictionary contains around 150,000 words, organized into set of synonyms, called synsets. A search for a word in dictionary will return all synsets, separated for the different word types (noun, verb, adverb, and adjective).

Figure 3 shows the result of querying the word “letter” from WordNet dictionary. There are two-word types, noun, and verb. The synonyms of “letter” in noun word type are letter, missive, letter of the alphabet, alphabetic character, varsity letter. There is only one synonym of “letter” for verb word type, which is “letter”.



The noun letter has 5 senses (first 4 from tagged texts)

1. (475) letter, missive -- (a written message addressed to a person or organization; "mailed an indignant letter to the editor")
2. (51) letter, letter of the alphabet, alphabetic character -- (the conventional characters of the alphabet used to represent speech; "his grandmother taught him his letters")
3. (2) letter -- (a strictly literal interpretation (as distinct from the intention); "he followed instructions to the letter"; "he obeyed the letter of the law")
4. (1) letter, varsity letter -- (an award earned by participation in a school sport; "he won letters in three sports")
5. letter -- (owner who lets another person use something (housing usually) for hire)

The verb letter has 3 senses (first 1 from tagged texts)

1. (1) letter -- (win an athletic letter)
2. letter -- (set down or print with letters)
3. letter -- (mark letters on or mark with letters)

Fig. 3: Querying the word “letter” in WordNet dictionary

3.6.2 British National Corpus (BNC): The British National Corpus (BNC) is a 100 million-word collection of samples from a wide range of sources such as regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, school and university essays and many other kinds of text. There are four columns in BNC: No (sort order), frequency, word, word type (noun, verb, adverb, adjective, etc). Figure 4 shows some of the BNC frequency data.

No.	BNC Frequency	Word	Word Type
…	…	…	….
478	21585	early	adv
479	21575	committee	n
480	21504	ground	n
481	21488	letter	n
482	21470	create	v
…	…	…	….

Fig. 4: Some of the BNC Frequency Data

3.7 American National Corpus (ANC)

The American National Corpus (ANC) is a text corpus of American English containing 22 million words written and spoken data. The American National Corpus (ANC) is a massive electronic collection of American English produced from 1990 onward. There are three columns in ANC: word, ANC frequency, and ratio (the frequency ratio for the word) as shown in Figure 5.

Word	ANC Frequency	Ratio
…	…	…
jobs	1739	0.0000784571
reference	1738	0.0000784120
meaning	1736	0.0000782317
saving	1734	0.0000782315
negative	1731	0.0000780962
…	…	…

Fig. 5: American National Corpus

3.8 Obfuscation Algorithm

Obfuscation is the term used to represent hiding data in the text. To obfuscate, two components are required, the synset that is available from Synonym Retrieval Algorithm and the bits to hide (the list of binary data that need to be hidden).

```
Data: Synset, bit to hide
Result: Chosen Synonym
if synset size less than bit then
return synset element at position bit;
End
```

3.9 Deobfuscation Algorithm

This algorithm produces the hidden bits by matching the synset and each word of the stego text.

```
Data: Synset, word
Result: Hidden bit(s)
for $i \leftarrow 0$ to synsetsize do
if synset[i] = word then
return i;
end
end
return null;
```

4. SYSTEM OVERVIEW AND ARCHITECTURE

This paper implements text steganography by applying Synonym Substitution Based Algorithm in letter of credit application. At sending side, the issuing bank uses ASCII table to get equivalent binary data of message, synonym retrieval algorithm to retrieves the synonyms of the message from WordNet dictionary, obfuscation algorithm to encode the LC information. During retrieving the synonyms from WordNet dictionary, synonym retrieval algorithm uses BNC frequency data to choose the possible word type and ANC frequency data to sort the retrieved synonyms. At the receiving side, the advising bank uses the deobfuscation algorithm to decode the LC information. Figure 6 shows the system flow chart.

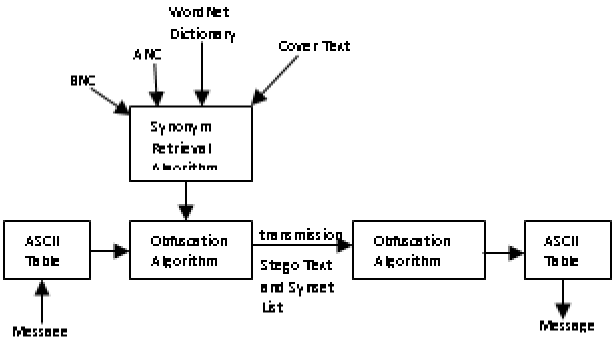


Fig. 6: System flow chart

4.1 Sending Side

There are three steps for sending side (issuing bank): step 1 gets the equivalent binary format of message from ASCII table. Step 2 retrieves list of synonyms (synset) by using Synonym Retrieval Algorithm, step 3 encodes the LC information by using Obfuscation Algorithm. The sending side performs step 2 and steps 3 alternately for each of the cover words. Figure 7 shows the example of the cover text for this paper.

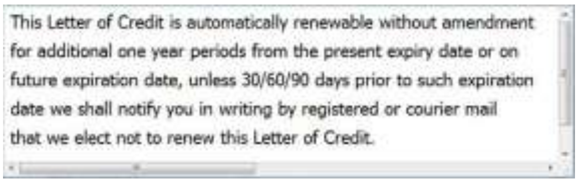
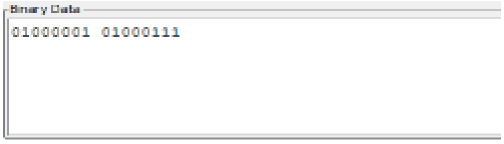

This Letter of Credit is automatically renewable without amendment for additional one year periods from the present expiry date or on future expiration date, unless 30/60/90 days prior to such expiration date we shall notify you in writing by registered or courier mail that we elect not to renew this Letter of Credit.

Fig. 7: The example of cover text

Step 1: Getting the equivalent binary format of message from the ASCII table

Firstly, the issuing bank gets the equivalent binary form of the buyer's LC information by using ASCII table. In this paper, we present the company's name as AG (Aung Gyi Trading Company) for the sending side as LC information. The equivalent binary data of "AG" is obtained as shown in figure 8.



```
01000001 01000111
```

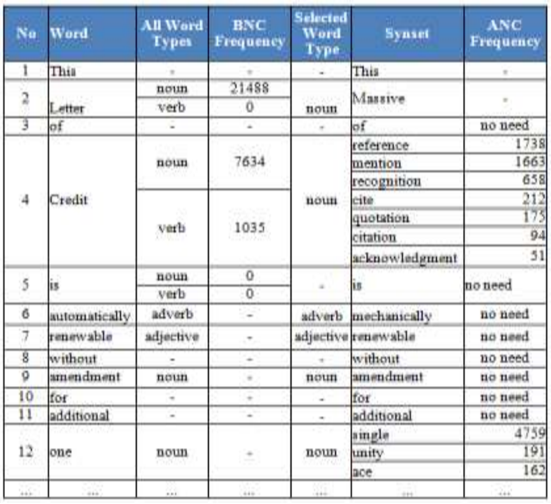
Fig. 8: The binary data of "AG"

Step 2: Synonym Retrieving

Synonym retrieving is processed by synonym retrieval algorithm. In figure 9, the processing the first 12 words of the cover text is shown as the detailed processing of the sending side. Figure 7 shows an example of the cover text.

The first task in retrieving the synonyms from the WordNet dictionary for a cover word is to decide which of the four-word types (noun, verb, adverb or adjective) the word is. In some cases, this is simply because the word can only be of one type

(No 6, 7, and 9, 12 of figure 9) but in many cases, the word can be two or more of the possible types (No 2, 4, and 5 of Figure 9). To choose the most likely word type, the word frequency data from the BNC is used. When a word is searched BNC, the total frequencies for noun, verb, adverb, and adjective are all found, and the greatest frequency is used as the correct word type for the word.

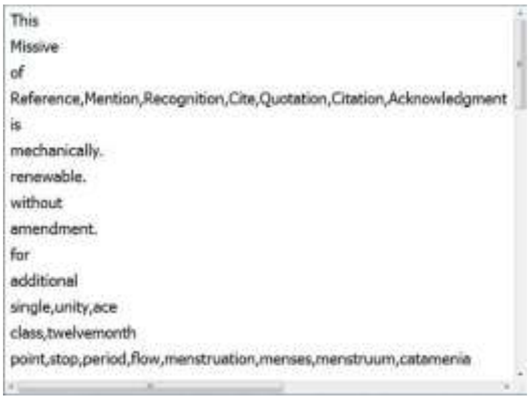


No	Word	All Word Types	BNC Frequency	Selected Word Type	Synset	ANC Frequency
1	This	-	-	-	This	-
2	Letter	noun	21488	noun	Massive	-
		verb	0			
3	of	-	-	-	of	no need
4	Credit	noun	7634	noun	reference	1738
					mention	1663
					recognition	658
		verb	1035		cite	212
					quotation	175
					citation	94
					acknowledgment	51
5	is	noun	0	-	is	no need
		verb	0			
6	automatically	adverb	-	adverb	mechanically	no need
7	renewable	adjective	-	adjective	renewable	no need
8	without	-	-	-	without	no need
9	amendment	noun	-	noun	amendment	no need
10	for	-	-	-	for	no need
11	additional	-	-	-	additional	no need
12	one	noun	-	noun	single	4759
					unity	191
					ace	162
...	...	...	...	...	...	...

Fig. 9: The detailed processing of the synonym retrieval algorithm

In figure 9 there are two-word types for the word "letter", noun and verb (No 2 of figure 9). BNC frequency for "noun" word type is "21488" and for verb word type is "0". Therefore the "noun" word type is selected.

Synonyms are retrieved from WordNet dictionary by using this selected word type. During retrieving process, synonyms that are composed of two or more words are ignored because it is impossible to know that they present one synonym when deobfuscating. In figure 3, synonym of "letter" in "noun" word type is "massive" because all other synonyms are composed of two words. In some cases, there is only one synonym is retrieved, (No 2 of figure 9), and in some cases, there may be more than one synonym (No 4 and 12 of figure 9). The synsets that are more than one synonym are sorted by using ANC frequency data to get the correct position when deobfuscating. The final synset list is shown in figure 10. This list is taken from a synset column of figure 9.



This
Missive
of
Reference,Mention,Recognition,Cite,Quotation,Citation,Acknowledgment
is
mechanically.
renewable.
without
amendment.
for
additional
single,unity,ace
class,twelvemonth
point,stop,period,flow,menstruation,menses,menstruum,catamenia

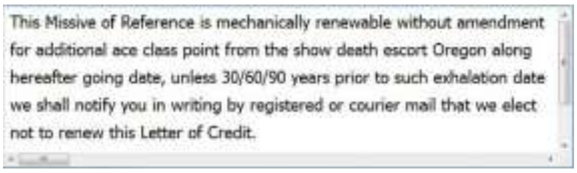
Fig. 10: Synset list that is retrieved from the synonym retrieval algorithm

Step 3: Obfuscation
Obfuscation is the term used to represent hiding data in the text. To obfuscate, two components are required, the synset that is available from Synonym Retrieval Algorithm and the bits to hide (the list of binary data that need to be hidden). Figure 11 shows the detailed processing of the Obfuscation algorithm.

No	Synset	Synset Position	Bits that can hide	Bits to hide (01000001)
1	This	0	-	01000001
2	Massive	0	-	01000001
3	of	0	-	01000001
4	**reference**	**0**	**0**	1000001
	mention	1	1	
	recognition	2	10	
	cite	3	11	
	quotation	4	100	
	citation	5	101	
	acknowledgment	6	110	
5	is	0	-	1000001
6	mechanically	0	-	1000001
	renewable	0	-	1000001
8	without	0	-	1000001
9	amendment	0	-	1000001
10	for	0	-	1000001
11	additional	0	-	1000001
12	single	0	0	00001
	unity	1	1	
	ace	**2**	**10**	

Fig. 11: Synset list that is retrieved from the synonym retrieval algorithm

If there is only one synonym in synset, no bits can be hidden (No 1, 2, 3, 5, 6, 7, 8, 9, 10, 11 of figure 11) and it is added to a list known as stego text. Else,if the first bit in the bit string is a 0, the first element at the first position is returned (No 4 of figure 11), if it is a 1 the element at the second position, if the first two elements are 10 (No 12 of figure 11), the element at the third position is returned, if they are 11 then the element at the fourth position is returned. After all, the returned element is added to stego text. Finally, the synset list retrieved from synonym retrieval algorithm and the stego text available from obfuscation algorithm is sent to the advising bank at the receiving side. Figure 12 shows the stego text.



This Missive of Reference is mechanically renewable without amendment for additional ace class point from the show death escort Oregon along hereafter going date, unless 30/60/90 years prior to such exhalation date we shall notify you in writing by registered or courier mail that we elect not to renew this Letter of Credit.

Fig. 12: Stego Text

4.2 Receiving Site

There are two steps for receiving side (advising bank): step 1 decodes the LC information by using deobfuscation algorithm and step 2 gets the equivalent string format of the available bit string from the obfuscation algorithm by using ASCII table.

Step 1: Deobfuscation
Deobfuscation refers to extracting hidden data from the text. To deobfuscate, two components are required: the stego text and the synset list. Figure 13 shows the detailed processing of deobfuscation.

No	Synset	Bits that can hide	Stego Text	Selected Binary	Current Binary Data
1	This	-	This	-	0
2	Massive	-	Massive	-	
3	of	-	of	-	
4	**reference**	**0**	**reference**	0	

	mention	1			
	recognition	10			
	cite	11			
	quotation	100			
	citation	101			
	acknowledgment	110			
5	is	-	is	-	010
6	mechanically	-	mechanically	-	
7	renewable	-	renewable	-	
8	without	-	without	-	
9	amendment	-	amendment	-	
10	for	-	for	-	
11	additional	-	additional	-	
12	single	0	**ace**	10	
	unity	1			
	ace	**10**			

Fig. 13: The detailed processing of deobfuscation.

The algorithm matches each word of the stego text and the synset list. If the stego word is the same with the synset in synset list then the position of the synset is returned (0 for the first position, 1 for the second position, 10 for the third position, and 11 for the fourth position) and so on. The algorithm goes through each synset list and stego text. Finally, a list of the binary string is returned.

Step 2: Getting the equivalent string format of message from the ASCII table

Finally, the equivalent string format of the binary data that is output deobfuscation algorithm is obtained from the ASCII table. Then, the message “AG” is obtained.

5. CONCLUSIONS

Text is still one of the major forms of communication in the world, both in digital and printed form. So the security of text over the internet has been important. This paper implements for the users who want to send text message safely from issuing bank to advising bank on communication channel by using synonym substitution-based algorithm. The three algorithms based on the synonym substitution algorithm can support to safe Letter of Credit (LC) information between issuing banks and advising banks. This system can be performed not only on low power machines but also online. By using this system, the hidden message cannot be suspected by another one who found the message because of hiding message is a normal text message. So the buyer and the seller can safely connect to send LC information by using this system.

6. REFERENCES

[1] A Synonym substitution-based Algorithm for text steganography.pdf.

[2] American Prosecutors Research Institute, ed. Steganography: Implications for the Prosecutor and Computer Forensics Examiner. Child Sexual Exploitation Program Update 1.1 (2004).

[3] United States Department of Justice. The criminal complaint by Special Agent Ricci against alleged Russian agents. 2010.

[4] Tayana Morkel, Jan HP Eloff, and Martin S Olivier. “An Overview of Image Steganography”. In: Proceedings of the Fifth Annual Information Security South Africa Conference ISSA2005).

[5] Jeff England. Audio Steganography. Echo DataHiding.

[6] Sean Gallagher. Steganography: how al-Qaeda hid secret documents in a porn video. arstechnica. May 2012.

[7] T. Moerland. Steganography and Steganalysis. Leiden Institute of Advanced Computing Science. 2003.

[8] Letters of credit for importers and exporters

[9] Practical Linguistic Steganography using Contextual Synonym Substitution and Vertex Colour Coding.pdf

[10] United States Department of Justice. The criminal complaint by Special Agent Ricci against alleged Russian agents. 2010.

[11] OpenStego. url:http://openstego.sourceforge. net/.

[12] T. Moerland. Steganography and Steganalysis. Leiden Institute of Advanced Computing Science. 2003.

[13] http://www.businessdictionary.com/definition/ASCII.html