available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/diin

Digital Investigation

# Detection of substitution-based linguistic steganography by relative frequency analysis

Zhili Chen [a,b,*], Liusheng Huang [a,b], Wei Yang [a,b]

[a] NHPCC, School of CS. & Tech., USTC, Hefei 230027, China
[b] Suzhou Institute for Advanced Study, USTC, Suzhou 215123, China

ARTICLE INFO

*Article history:*
Received 5 August 2010
Received in revised form
17 January 2011
Accepted 16 March 2011

*Keywords:*
Information hiding
Linguistic steganography
Relative frequency analysis
Detection
Substitution-based
Synonym-substitution

ABSTRACT

Linguistic steganography hides information in natural language texts. Because of the increasing in importance and quantity of natural language texts, linguistic steganography plays a more and more important role in Information Security (IS) area today. Substitution-based linguistic steganography is one of the most commonly used linguistic steganography methods, which is of considerable security and favorable simpleness. In this paper, we propose a straightforward method based on Relative Frequency Analysis (RFA), which makes use of the frequency characteristics of the testing texts (the texts being tested), to detect substitution-based linguistic steganography. We formally prove several properties about relative frequency which can be used in the detection process and propose a detection scheme. And then as an example, an existent synonym-substitution system T-Lex is examined and the detection experiment is carried out. In the experiment with pure literature texts, the accuracy, precision and recall of the detection are found to be as high as 98.64%, 97.77% and 99.55%, respectively, when the substitution count is 90, while in the experiment with balanced texts, the highest detection accuracy is 95%, which indicates that the detection scheme is promising.



## 1. Introduction

Nowadays, natural language texts have become one of the largest chunks of digital data that people encounter daily. Numberless articles in digital format come from many newspapers, magazines, scientific journals and conferences every year. Besides, emails, blogs and all kinds of web pages online provide even more textual data. This enhancement in the significance and quantity of digital text on Internet creates increased concerns about the usage of textual data as a covert channel of communication. One of such covert ways of communication is known as linguistic steganography. Linguistic steganography embeds messages into natural language texts in a covert manner such that the presence of the hidden messages cannot be easily discovered by anyone except the intended recipient.

Linguistic steganography methods can be grouped under two categories (Bennett, 2004), Linguistic Driven Generation-based (LDG-based) and Linguistic Driven Modification-based (LDM-based). The first group is based on generating a new stego-text for a given message. For example, NICETEXT (Chapman, 1997; Chapman et al., 2001) and TEXTO (Maher, 1995) are LDG-based. The second group is based on modification of an existent covert text. Coarsely speaking, LDG-based methods, which generate stego-texts looking like natural texts but being lack of coherent sense, have a higher hiding capacity than LDM-based methods. LDG-based methods are mainly used to conceal a great bulk of encrypted

* *Corresponding author*. No.166, Ren'ai Road, Suzhou Industrial Park, Suzhou 215123, PR China. Tel.: +86 512 87161305.
E-mail address: zlchen3@ustc.edu.cn (Z. Chen).
1742-2876/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.diin.2011.03.001

secret information during data transmission. LDM-based methods have a lower hiding capacity and their stego-texts look like natural texts both syntactically and semantically. They are used for hiding more secret information transferred and are widely used in natural language watermarking. One subcategory of LDM-based linguistic steganography is called substitution-based linguistic steganography which substitutes some elements of the cover text with their semantically equivalent ones. In this paper, we mainly focus on the detection of substitution-based linguistic steganography.

This paper presents a method making use of relative frequency analysis to detect substitution-based linguistic steganography. Different from the previous detections (Taskiran et al., 2006; Luo et al., 2008; Yu et al., 2008) which usually need the context information, we consider the detection utilizing frequency information only. We closely look into the frequency characteristics of the substitution elements in the substitution sets in both cases of normal texts and stego-texts and find that the relative frequency satisfies certain properties. We then propose a detection scheme making use of these properties. Finally, we examine the synonym-substitution system implemented by Winstein (1999) as an example, design a detection algorithm under the detection scheme and validate the detection by some experiments. Experimental results show the efficiency of the detection.

## 2. Related work

Substitution-based linguistic steganography can be classified into synonym substitution-based methods (Winstein, 1999; Atallah et al., 2000; Bergmair, 2004; Bolshakov and Gelbukh, 2004; Bolshakov, 2004; Calvo and Bolshakov, 2004; Topkara et al., 2006; Liu et al., 2007), semantically equivalent rule substitution-based methods (Hugg, 1999), synonymous sentence substitution-based methods (Murphy, 2001), translation-based methods (Grothoff et al., 2005a,b; Stutsman et al., 2006) and so on, according to the substitution element. Among those, synonym substitution-based linguistic steganography is most widely used. In synonym substitution system, the hidden message is embedded by substituting a word with one of its synonyms. The stego-text keeps the same sense before and after substitution. In this section, we introduce T-Lex system as an example of substitution-based linguistic steganography and discuss its drawbacks that make the accurate detection possible. After this, we analyze the previous attacks against this system, and discuss their weakness.

### 2.1. T-Lex system

The most important problem that synonym substitution faces is how to define the synonym set. In natural language, words often have many senses in different contexts. How to determine the exact sense in a certain context is a hard problem known as word sense disambiguation in Natural Language Processing (NLP). The definition of synonym set must guarantee that all synonym sets are mutually disjoint in order not to cause word sense disambiguation problem. However, the multi-sense property makes definition of synonym set difficult. For example, word $A$ and $B$ are synonyms in a context, word $B$ and $C$ are also synonyms in anther context, but word $A$ and $C$ can have different senses in any context.

Winstein (1999) proposed a solution for synonym set definition in T-Lex system. He used WordNet (WordNet) to select synonyms with correct senses. In WordNet, Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synset), each of which expressing a distinct concept. Synsets are interlinked both conceptual-semantically and lexically. In T-Lex system, not all synonyms in WordNet database are included in the synonym database. Only those words completely in the same synsets are grouped in the same synonym set. For example, assume that words $a$, $b$, $c$ only belong to the synsets $S1{:}\{a, b\}$, $S2{:}\{a, b, c\}$. In this case, even though both words $a$ and $b$ have more than one sense, they still can be interchanged semantically in all contexts. Applying the criteria described above, Keith obtained synsets containing about 30% of 70,803 single word entries in WordNet as the synonym database of T-Lex system. The average synset size is 2.56 while the maximum is 13 and the size is 2 (Winstein, 1999).

T-Lex system currently only hides text messages in the cover texts, but modification of hiding any kind of messages is easy. A given text message is embedded into the cover text using the synset database as follows. First, the letters of the message text are Huffman coded according to English letter frequencies. Then, the Huffman code binary string is represented in mixed-base form. As a simple example, suppose that the binary string to be embedded is $(010)_2$ and that currently the sentences following are being considered.

$$\text{"... A bicycle was lying upon the } \left\{ \begin{array}{l} \textit{roadside} \\ 0 : \textit{wayside} \\ 1 : \textit{roadside} \end{array} \right\} \textit{grass ...}\text{"}$$

$$\text{"... and he had a pair of } \left\{ \begin{array}{l} \textit{shrewdly} \\ 0 : \textit{shrewdly} \\ 1 : \textit{astutely} \\ 2 : \textit{sagaciously} \\ 3 : \textit{sapiently} \end{array} \right\} \textit{careless boyish eves...}\text{"}$$

In the example, the first words with no number leading in the brace are the original words to be replaced and the numbered words constitute their corresponding synonym set. In mixed-base form, each digit has a different base. For $(010)_2 = 2$, we have

$$4a_1 + a_0 = 2$$

With the constraints $0 \le a_0 < 4$ and $0 \le a_1 < 2$. Thus, we get $a_0 = 2$ and $a_1 = 0$. This indicates that "roadside" and "shrewdly" should be replaced by "wayside" and "sagaciously".

### 2.2. Drawbacks of T-Lex system

Two shortcomings of T-Lex system have been pointed out (Taskiran et al., 2006). One is that it sometimes substitutes words with their synonyms that do not agree with the correct English usage; the other is that the words after substitution do not agree with the genre and the author style of the cover text.