# Natural Language Watermarking

Mercan Topkara

Center for Education and Research in Information Assurance (CERIAS)
Purdue University
West Lafayette, Indiana, 47907

Cuneyt M. Taskiran  Edward J. Delp

Video and Image Processing Laboratory (*VIPER*)
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana, 47907

## ABSTRACT

In this paper we discuss natural language watermarking, which uses the structure of the sentence constituents in natural language text in order to insert a watermark. This approach is different from techniques, collectively referred to as "text watermarking," which embed information by modifying the appearance of text elements, such as lines, words, or characters. We provide a survey of the current state of the art in natural language watermarking and introduce terminology, techniques, and tools for text processing. We also examine the parallels and differences of the two watermarking domains and outline how techniques from the image watermarking domain may be applicable to the natural language watermarking domain.

**Keywords:** text watermarking, natural language processing, text steganography

## 1. INTRODUCTION

Even with the proliferation of image and video data in recent years, text data still forms the bulk of Internet traffic and other forms of data we encounter everyday. Most magazines, newspapers, scientific journals, and conferences provide articles in digital format. While this is improving the ways readers can search and access information, it also brings about author concerns about how their work is distributed and re-used. Rights management problems are more serious for text than they are for images and video data since it is much easier for users to download and manipulate copyrighted text. In this paper we review the current state of the art in *natural language (NL) watermarking*, which aims to embed information in text documents by manipulating the semantic and/or syntactic structure of sentences. This approach is different from techniques, collectively referred to as "text watermarking," which modify the appearance of text elements, such as lines, words, or characters.[1] Text watermarking is achieved by altering the text format or fonts, such as modifying inter-word and inter-letter spacing in text. Watermarks inserted by most of these systems are not robust against attacks such as scanning the document and performing optical character recognition or re-formatting of the document file. Although much work has been done in text watermarking, NL watermarking is a relatively new area. In addition to content protection, robust NL watermarking algorithms will enable a wide range of applications such as text auditing, meta-data binding, tamper-proofing, and traitor tracing.

Our goal in this paper is two-fold: First, we review the current state of the art in NL watermarking and introduce the terminology and techniques that may be unfamiliar to the general image watermarking community. We review the set of relevant tools, such as parsers, generators, semantic analyzers, for NL processing and discuss their performance. Second, we discuss how some existing work in image watermarking systems may be applied to NL watermarking. We also examine the differences and parallels between the two areas.

The organization of the paper is as follows: In Section 2 we examine the similarities between image and NL watermarking as well the unique difficulties in NL watermarking caused by the structure of language. Basic concepts of NL processing techniques and available resources, which can be employed to develop NL watermarking systems are introduced in Section 3. Surveys of current state of the art in NL steganography and watermarking are provided in Sections 4 and 5, respectively. Some directions for future work are suggested in Section 7. Finally, conclusions are given in Section 8.

Portions of this work were supported by Grants IIS-0325345, IIS-0219560, IIS-0312357, and IIS-0242421 from the National Science Foundation, Contract N00014-02-1-0364 from the Office of Naval Research, by sponsors of the Center for Education and Research in Information Assurance and Security, and by Purdue Discovery Park's e-enterprise Center.

## 2. NATURAL LANGUAGE WATERMARKING VERSUS IMAGE WATERMARKING

The goals of watermarking in both image and natural language (NL) are the same: The embedding of information by modifying original data in a discreet manner, such that the modifications are imperceptible when the watermarked data is consumed and the embedded information is robust against possible attacks. In image watermarking this goal is achieved by exploiting the redundancy in images and the limitations of the human visual system. Similar approaches are used in other signal-based watermarking domains, such as video and audio. On the other hand, language has a discrete and syntactical nature that makes such techniques more difficult to apply. Specifically, language, and consequently its text representation, has two important properties that differ from image representations.

- Sentences have a combinatorial syntax and semantics. That is, structurally complex (molecular) representations are systematically constructed using structurally simple (atomic) constituents, and the semantic content of a sentence is a function of the semantic content of its atomic constituents together with its syntactic/formal structure.
- The operations on sentences are causally sensitive to the syntactic/formal structure of representations defined by this combinatorial syntax.

Images in general do not lend themselves to a syntactical decomposition similar to the one for language *. The atomic/syntactical nature of language brings about unique challenges for NL watermarking. For example, deriving an analog of least significant bit (LSB) embedding used in image watermarking that modifies text locally, i.e., based on words, without making perceptually significant changes to sentence structure is a hard problem. This is due to the fact that even small local changes in a sentence can change its semantics and/or make it ungrammatical. The only current local modification techniques used are the synonym substitution methods in NL steganography discussed in Section 4.2. These approaches are unsuitable for NL watermarking since they are not robust to attacks.

A better approach to NL watermarking is to analyze the global semantic/syntactical structure of the sentence to be modified and then apply transformations that preserve its meaning and grammaticality. According to the transformational grammar (TG) theory of Chomsky[3] multiple sentences may be derived from the same underlying form by linguistic transformations. For example, the sentences "Ned loves Jody" and "Jody is loved by Ned" convey the same meaning although one is active and the other is passive. According to the TG theory these two sentences are derived from the same underlying form. The underlying form is known as the *deep structure* and the syntactic structure derived from the deep structure using syntactic transformations is known as the *surface structure* †. A number of example transformations are listed in Section 3.2. These transformations apply not directly to sentences as a whole but to their constituent phrase structures, which can be obtained using sentence parsers, as illustrated in Section 3.3. A rough but useful analogy to the information contained in the D-structure of a sentence is the discrete cosine transform (DCT) of an image. Watermarking techniques that modify perceptually significant portions of an image are more robust against attacks than techniques that modify only perceptually insignificant portions (such as LSB embedding).[5] Similarly, NL watermarking techniques that embed information in the underlying structure of a sentence will be more robust than those that modify the surface representation of the sentence. Such approaches are described in Section 5.2.

NL watermarking bears a close resemblance to the machine translation (MT) task of NL processing. Rather than converting sentences from one language to another, their style and other properties are modified in a single language while embedding information. This makes it possible to adapt numerous MT methodologies and tools to the NL watermarking problem. There are two commonly used approaches from the MT field that are useful to NL watermarking. One is parsing sentences in text into an intermediate representation, transforming this

*Although syntactic approaches to image analysis gained some success in the analysis of some simple, highly structured images, such as electrical circuits and maps, for the most part they have been abandoned since they are not robust for natural images.[2]

†In recent linguistics literature these terms are avoided because of their broad connotations and the terms D-structure and S-structure are preferred.[4]

| Name of the Corpus | Size (app.) | Properties |
|---|---|---|
| Brown | one million words | American English, 15 different categories of text printed in 1961, balanced corpus |
| Lanchester-Oslo-Bergen | one million words | British English counterpart of the Brown corpus |
| Susanne | 130,000 words | Freely available subset of the Brown corpus |
| Wall Street Journal | 40,000,000 words | American English, financial news articles from 1987 to 1993 |
| Reuters | 810,000,000 words | British English,810,000 articles printed from 1996 to 1993 |
| Penn Treebank II | one million words | Parsed sentences of 1989 Wall Street Journal articles |

**Table 1.** Properties of some of the well known corpora available from the Linguistic Data Consortium's website[8]

| Category | Unique Strings | Number of Senses |
|---|---|---|
| Noun | 114648 | 141690 |
| Verb | 11306 | 24632 |
| Adjective | 21436 | 31015 |
| Adverb | 4669 | 5808 |
| Total | 152059 | 203145 |

**Table 2.** Wordnet2.0 Database Statistics

parse structure into a corresponding parse in target style using pre-determined transfer rules, and realizing this parse as a sentence using NL generation methods. This is the method that we have outlined above. Another approach is to use a *interlingua*, a language-neutral canonical form which can represent all sentences that mean the "same" thing in the same way regardless of the stylistic conventions of text.[6] Translation is done by parsing the sentence into the interlingua and later performing generation to the target style using this representation.

## 3. NATURAL LANGUAGE PROCESSING TECHNIQUES AND RESOURCES

Natural Language Processing (NLP) aims to design algorithms that will analyze, understand, and generate natural language automatically. In this section we will briefly introduce NLP techniques and resources that are of interest for information hiding in natural language text. For an in-depth treatment of the NLP field consult references[6] and.[7]

### 3.1. Data Resources

Success of an information hiding system depends on obtaining good models of the cover medium which can only be achieved with large data sets. A statistically representative sample of natural language text is referred to as a *corpus*. Since most of NLP research is based on statistical analysis and machine learning systems, large corpora in machine readable form are essential. Therefore, a number of corpora in electronic form have been created and are commonly used in NLP research. These corpora and information about them are provided in Table 1. In order to make the corpora more useful for NLP research, they are usually annotated with extra information. An example of such annotation is part-of-speech tagging where information about each word's part of speech (such as verb, noun, adjective) is added to the corpus in the form of tags. The Penn Treebank is an example of such a corpus.

In addition to corpora, there are also electronic dictionaries available that are designed as large databases of lexical relations between words. The most widely known such dictionary is Wordnet.[9] In Wordnet English nouns, verbs, adjectives, and adverbs are organized into synonym sets, each set representing an underlying lexical concept. The content of Wordnet is summarized in Table 2.[6] VerbNet[10] is another electronic dictionary which is a verb lexicon with syntactic and semantic information for English verbs, using Levin verb classes[11] to systematically construct lexical entries.

| Transformation | Original sentence | | Transformed sentence |
|---|---|---|---|
| *Passivization* | The slobbering dog kissed the big boy. | ⇒ | The big boy was kissed by the slobbering dog. |
| *Topicalization* | I like bagels. | ⇒ | Bagels, I like. |
| *Clefting* | He bought a brand new car. | ⇒ | It was a brand new car that he bought. |
| *Extraposition* | To believe that is difficult. | ⇒ | It is difficult to believe that. |
| *Preposing* | I like big bowls of beans. | ⇒ | Big bowls of beans are what I like. |
| *There-construction* | A unicorn is in the garden. | ⇒ | There is a unicorn in the garden. |
| *Pronominalization* | I put the letter in the mailbox. | ⇒ | I put it there. |
| *Fronting* | "What!" Alice cried. | ⇒ | "What!" cried Alice. |

**Table 3.** Some common syntactic transformations in English.

### 3.2. Linguistic Transformations

In order to embed information in natural language text a systematic method for modifying, or transforming, text is needed. These transformations should preserve the grammaticality of the sentences. Ideally we also require that the differences in sentence meaning caused by the transformations should not be noticeable. Generally three types of transformations are used for modification: synonym substitution, syntactic transformations, and semantic transformations.

*Synonym substitution* is the most widely used linguistic transformation for information hiding systems since it is the simplest transformation. Synonym substitution has to take the sense of the word into consideration. In order to preserve the meaning of the sentence the word should be substituted with a synonym in the same sense. For example the word "bank" has at least three different senses as a financial institution, a river edge, or something to sit on. An electronic dictionary like Wordnet that classifies all words and phrases into synonym sets can be used to search for words that are synonyms for a given word. However, determining the correct sense of a given word, referred to as the word sense disambiguation task in NLP, may present hard problems since it is hard to even derive a general definition for word sense.[12]

A second type of transformation is the class of *syntactic transformations*, such as passivization and clefting, which change the syntactic structure of a sentence with little effect on its meaning. Some of the common syntactic transformations in English are listed in Table 3. In addition to these, there is another group of syntactic transformations that are solely based on the categorization of the main verb of the sentence. Verbs can be classified according to shared meaning and behavior, and different classes of verbs allow different transformations to be performed in the sentence.[11] Examples of a transformation known as the locative alternation are given below.

```
Jack sprayed paint on the wall.           ⇒  Jack sprayed the wall with paint.
Henry cleared the dishes from the table.  ⇒  Henry cleared the table of the dishes.
```

The third type of linguistic transformation is the class of *semantic transformations*. One method to generate meaning-preserving semantic transformations is by using noun phrase coreferences.[13] Two noun phrases are coreferent if they refer to the same entity. Based on the coreference concept different transformations may be introduced. One such transformation is *coreferent pruning*, where repeated information about the coreferences is deleted. The opposite of this operation, *coreferent grafting* may also be performed where information about a coreference is repeated in another sentence, or added to the text using a fact database. Finally, we may perform *coreferent substitution* which may be viewed as a combination of the previous two transformations. As an example of these semantic transformations, consider the following news story.

> Yet Iceland has offered a residency visa to ex-chess champion **Bobby Fischer** in recognition of a 30-year-old match that put the country ''on the map''. **His** historic win over Russian Boris Spassky in Reykjavik in 1972 shone the international spotlight on Iceland as never before. Now Iceland is keen to repay the favour by offering sanctuary to **Mr Fischer**, an American citizen. **He** is being detained in Japan and is wanted in the US for violating international sanctions against the former Yugoslavia by playing there in 1992.

| Parser | Input Format | Output Format | Accuracy |
|---|---|---|---|
| Link, 1995 | Raw sentence | Phrase level parse in PennTreebank Format | Not Available |
| Collins, 2000 | Sentence with part-of-speech tags | Word level parse in PennTreebank Format | 90.1% |
| Charniak, 2000 | Raw sentence | Word level parse in PennTreebank Format | 90.1% |
| XTAG, 2001 | Raw sentence | Word level parse in Tree-Adjoining Grammar Format | 87.7% |

**Table 4.** Properties of commonly used syntactic parsers that are freely available.

The focus of the analysis is the reference item "'Bobby Fischer". Pruning is applied to the first sentence and the extracted information is used to perform a substitution at the second sentence. Similarly, information extracted from the third sentence is used to perform grafting in the fourth sentence. The modified text is given below.

> Yet Iceland has offered a residency visa to **Bobby Fischer** in recognition of a 30-year-old match that put the country ''on the map''. **Ex-chess champion's** historic win over Russian Boris Spassky in Reykjavik in 1972 shone the international spotlight on Iceland as never before. Now Iceland is keen to repay the favour by offering sanctuary to **Mr Fischer**, an American citizen. **He**, an American citizen, is being detained in Japan and is wanted in the US for violating international sanctions against the former Yugoslavia by playing there in 1992.

One problem with the above approach is that coreference resolution is one of the hardest tasks in NLP. Furthermore, it may not be appropriate to substitute two coreferent phrases in some circumstances. As a well-known example, consider the following sentences.

> Spiderman just saved us from death.
> Peter Parker just saved us from death.

The phrases Spiderman and *Peter Parker* do in fact refer to the same person but someone who does not know this fact may think the first sentence is true while the second one is not.

### 3.3. Natural Language Parsing

In NLP *parsing* is defined as processing input sentences and producing some sort of structure for them.[6] The output of the parsing may either be the morphological, syntactical, or semantical structure of the sentence or it may be a combination of these. Parsing is essential to get more information about the sentence structure and the roles of the constituent words in this structure. Most parsers use part-of-speech taggers, which categorize words into predetermined classes (such as noun, adjective, or verb), and morphological analyzers, which break up words into their morphemes in pre-processing steps. Properties and accuracies of some of the commonly used parsers are listed in Table 4.

The parser output may be viewed as a transformed representation of the given text. Various transforms used in image data hiding may be used as a simple analogy to parsing. The input text and the tree relationships produced by the parser are conceptually similar to the time and frequency domain representations of an image.

To the best of our knowledge, there is no fully implemented semantic parser available. However, there are many tools that can convert phrase structures generated by syntactic parsers into dependency trees, which illustrate the argument or modifier relation between words in the sentences.[14] The dependency tree generated for a simple sentence above is shown in Figure 1(a).

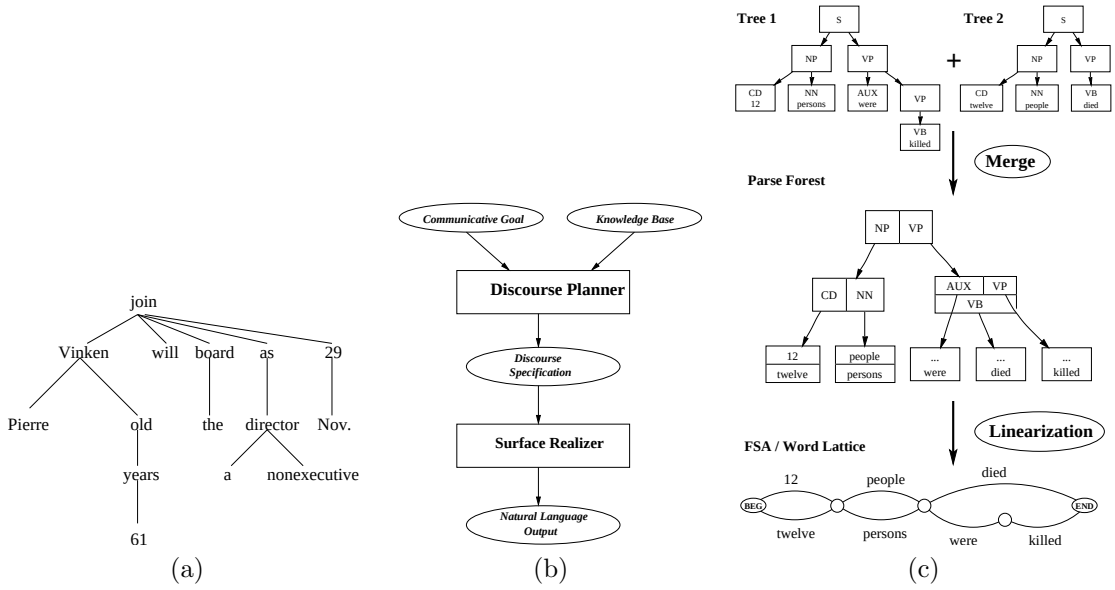


**Figure 1.** (a) Dependency tree for the sentence, "Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29." (b) Components of a typical natural language generation system. (c) An example of text paraphrasing using a finite-state approach

### 3.4. Natural Language Generation

The *natural language generation* (NLG) task is defined as the process of constructing natural language output from non-linguistic information representations according to some communication specifications. The components of a typical NLG system are illustrated in Figure 1(b). A good example of a fully functional NLG system is the Forecast Generator (FOG),[15] a weather forecast system that generates bilingual text in English and French. This system takes raw meteorological data and generates weather forecasts. There are several fully implemented NLG systems freely available for research purposes.[16]

As far as NL information hiding is concerned, NLG is a crucial component. After information is added to a sentence by modifying its structural representation, this altered representation needs to be converted back to natural language using NLG systems. NLG systems also play a crucial part in natural language stegonagraphic systems as cover text generation mechanisms.

### 3.5. Text Paraphrasing

The task of *text paraphrasing* entails changing text parameters such as length, readability, and style for a specific purpose without losing the core meaning of the text. Therefore, text paraphrasing is directly related to NL watermarking. Text paraphrasing is also similar to machine translation; however, rather than converting text from one language to another, it is modified from one form to another within the same language. Paraphrasing systems are mainly based on creating or collecting sets or pairs of semantically equivalent words, phrases, and patterns. For example, the sentences

```
After the latest Fed rate cut, stocks rose across the board.
Winners strongly outpaced losers after Greenspan cut interest rates again.
```

form such a semantically related pair. Such training sentence pairs may be located in news stories covering the same event by using multiple sequence alignment techniques.[17] After the system is trained, given a sentence, it is possible to create a paraphrase using the best matching template pair. An example of text paraphrasing using a finite-state approach[18] is shown in Figure 1(c).

| Rule # | Rule | code | prob. |
|---|---|---|---|
| (1) | $S \Rightarrow AB$ | 0 | 0.5 |
| (2) | $S \Rightarrow CB$ | 1 | 0.5 |
| (3) | $A \Rightarrow$ She | 00 | 0.25 |
| (4) | $A \Rightarrow$ He | 01 | 0.25 |
| (5) | $A \Rightarrow$ Susan | 10 | 0.25 |
| (6) | $A \Rightarrow$ Alex | 11 | 0.25 |
| (7) | $B \Rightarrow$ likes $D$ | 0 | 0.5 |
| (8) | $B \Rightarrow$ detests $D$ | 10 | 0.25 |
| (9) | $B \Rightarrow$ wants $D$ | 110 | 0.125 |
| (10) | $B \Rightarrow$ hates $D$ | 111 | 0.125 |
| (11) | $C \Rightarrow$ Everybody | 0 | 0.5 |
| (12) | $C \Rightarrow$ The cleaning lady | 10 | 0.25 |
| (13) | $C \Rightarrow$ A nice kid | 11 | 0.25 |
| (14) | $D \Rightarrow$ milk. | 00 | 0.25 |
| (15) | $D \Rightarrow$ apples. | 01 | 0.25 |
| (16) | $D \Rightarrow$ pumpkin pie. | 10 | 0.25 |
| (17) | $D \Rightarrow$ cookies. | 11 | 0.25 |

(a)

| Position | Prefix | Rule | output string |
|---|---|---|---|
| $\bullet$1011001 | 1 | 2 | $CB$ |
| 1 $\bullet$ 011001 | 0 | 11 | Everybody $B$ |
| 10 $\bullet$ 11001 | 110 | 9 | Everybody wants $D$ |
| 10110 $\bullet$ 01 | 01 | 15 | Everybody wants apples. |

(b)

**Figure 2.** Using a probabilistic context-free grammar to generate cover text for the secret payload 1011001. (a) A very simple probabilistic context-free grammar. The Huffman code corresponding to each rule is also listed. (b) Generation of cover text using the rules determined by the payload.

## 4. PREVIOUS APPROACHES TO NATURAL LANGUAGE STEGANOGRAPHY

Compared to similar work in the image and video domain, work in NL steganography and watermarking has been scarce. The previous work in this area has concentrated on NL steganography. This is probably due to the fact that it is hard to derive robust watermarking methods for text. In this section we review the previous work done in NL steganography.

### 4.1. Using Probabilistic Context-Free Grammars to Generate Cover Text

A *probabilistic context-free grammar* (PCFG) is a commonly used language model where each transformation rule of a context-free grammar has a probability associated with it.[7] A PCFG can be used to generate strings by starting with the root node and recursively rewriting it using randomly chosen rules. Conversely, a string belonging to the language produced by a PCFG can be parsed to reveal the sequence of possible rules that produced it.

In the mimicry text approach described in[19] a cover text is generated using a PCFG that has statistical properties close to normal text. This is achieved by assigning a Huffman code to each grammar rule based on the probability of the rule. The payload string is then embedded by choosing the grammar rule whose code corresponds to the portion of the message being embedded. An example sentence generated by this technique is illustrated in Figure 2. In practice the PCFG and the corresponding rule probabilities are learned using a corpus.

The problem with this method is that even within limited linguistic domains, deriving a PCFG that models natural language is a daunting task. Also, some aspects of language cannot be modeled by context-free grammars at all. Because of these reasons cover texts produced by PCFGs tend to be ungrammatical and nonsensical. This makes it very easy for native speakers to detect such texts, which defeats the steganographic purpose of the method. Therefore, this method can only be used in communication channels where computers act as wardens.

### 4.2. Information Embedding Through Synonym Substitutions

The simplest method of modifying text for embedding of a payload is to replace selected words by their synonyms so that the truth values of the modified sentences are preserved, as described in Section 3.2. A dictionary, such

| Type | Code | Word |
|---|---|---|
| name–male | 0 | ned |
| name–male | 1 | tom |
| name–female | 0 | jody |
| name–female | 1 | tracy |

(a)

| Style | Payload | Output string |
|---|---|---|
| name–male name–male name–male | 011 | ned tom tom |
| name–male name–male name–female | 011 | ned tom tracy |
| name–male name–female name–male | 011 | ned tracy tom |
| name–male name–female name–female | 011 | ned tracy tracy |
| name–female name–male name–male | 011 | jody tom tom |
| name–female name–male name–female | 011 | jody tom tracy |
| name–female name–female name–male | 011 | jody tracy tom |
| name–female name–female name–female | 011 | jody tracy tracy |

(b)

**Figure 3.** Example of a simple dictionary and how the style template affects the output for the *NICETEXT* system.[22] (a) A simple dictionary with two types, name–male and name–female. (b) Using a style and the dictionary in (a) to generate text corresponding to a payload string.

as Wordnet, may be used to find, for each selected word $w$ in text, the *synonym set*, which is defined as a set of words that are synonymous with $w$.

In the method described in[20] words in a synonym set are indexed according to their alphabetical order. During embedding the selection process picks a subset of words from the text for replacement. A simplified example of this embedding is given in[21] as follows: Suppose we have the sentence

$$\texttt{Midshire is a} \left\{ \begin{array}{cc} 0 & wonderful \\ 1 & decent \\ 2 & fine \\ 3 & great \\ 4 & nice \end{array} \right\} \texttt{little} \left\{ \begin{array}{cc} 0 & city \\ 1 & town \end{array} \right\},$$

where the words in the braces are the synonym sets. If the current string to be embedded is $(101)_2 = 5$, it is first represented in mixed radix form as

$$\left( \begin{array}{cc} a_1 & a_0 \\ 5 & 2 \end{array} \right) = 2a_1 + a_0 = 5,$$

with the constraints that $0 \leq a_1 < 5$ and $0 \leq a_0 < 2$. Thus, we obtain the values $a_1 = 2$ and $a_0 = 1$ which indicates that we should use the words *fine* and *town*.

### 4.3. Generating Cover Text Using Hybrid Techniques

The *NICETEXT* system[22, 23] for the generation of natural-like cover text according to a given payload uses a mixture of both of the methods discussed above. The system has two components: a dictionary table and a style template. The dictionary table is a large list of (*type,word*) pairs where the *type* may be based on the part-of-speech[22] of *word* or its synonym set.[23] Such tables may be generated using a part-of-speech tagger or Wordnet. The dictionary is used to randomly generate sequences of words. The style template, which is conceptually similar to the PCFG of Section 4.1, improves the quality of the cover text by selecting natural sequences of parts-of-speech while controlling the word generation, capitalization, punctuation, and white space. An example of a simple dictionary and how the style template affects the generated text is illustrated in Figure 3. A dictionary containing more than 200,000 words categorized into more than 6,000 types was used in.[22] Different style templates, such as Federal Reserve Board meeting minutes or Aesop's Fables, may be learned using online corpora and employed in the system.

## 5. PREVIOUS WORK ON NATURAL LANGUAGE WATERMARKING

As pointed out in Section 4, work on NL watermarking is more scarce than work on NL steganography. To the best of our knowledge, the only NL watermarking systems are those proposed by Atallah et al..[13, 24, 25]

### 5.1. Synonym Substitution Based on Quadratic Residues

The idea of employing the semantics and syntax of text for inserting watermarks was first proposed by Atallah et al.[24] in 2000, where ASCII values of the words were used for embedding information into text by performing lexical substitution in synonym sets.

Let $m_{i \bmod k}$ be the bit of watermark message that is to be embedded and $w_i$ be the current word being considered in the cover text with ASCII value $A(w_i)$. If $m_{i \bmod k} = 1$ and $A(w_i) + r_{i \ mod \ k}$ is a quadratic residue modulo $p$, then $w_i$ is kept same. Otherwise it is modified. Here $p$ is a 20 digit prime key, $k$ is the number of bits in the watermark message, and $r_0, r_1, ..., r_{k-1}$ is a sequence of pseudo-random numbers generated using $p$ as seed.

### 5.2. Embedding Information in the Tree Structures of Sentences

In later work[13, 25] Atallah et al. have proposed two algorithms that embed information in the tree structure of the text rather than using lexical substitution. These techniques aim to modify the structural properties of intermediate representations of sentences, built using NL processing tools. In other words, the watermark is not directly embedded to the text, as is done in lexical substitution, but to the parsed representation of sentences. Utilizing the intermediate representation makes these algorithms more robust to attacks compared with lexical substitution systems.

The difference between the two proposed algorithms in[13, 25] is that the first one modifies syntactic parse trees of the cover text sentences for embedding while the second one uses semantic tree representations. A *syntactic tree* is a representation of various parts of a sentence that has been syntactically parsed. Examples of syntactic trees for two sentences are given below.

```
I took the book.
(S (NP I) (VP took (NP the book)) (. .))

The book was taken by me.
(S1 (S(NP (DT The) (NN book))(VP (VBD was) (VP (VBN taken) (PP (IN by) (NP (PRP me))))) (. .)))
```

By contrast a *semantic tree* is a tree-structured representation that is imposed over the flat text meaning representation of a sentence.[13] Such representations of sentences may be generated by using ontological semantics resources.[26] A sentence and its semantic tree are given below.

```
The EU ministers will tax aviation fuel as a way of curbing the environmental impact of air travel.

author-event-1--|--author--unknown
                |--theme--levy-tax-1--|--agent--set-4--|--member-type--geopolitical-entity
                                      |                |--cardinality--unknown
                                      |                |--members--(set| "EU nations")
                                      |--theme--kerosene-1
                                      |--purpose--regulate-1--|--agent--unknown-1
                                                              |--theme--effect-1--|--caused-by--flight
```

Selection of sentences that will carry the watermark information depends only on the tree structure and proceeds as follows: The nodes of the tree $T_i$ for sentence $s_i$ of text are labeled in pre-order traversal of $T_i$. Then, a node label $j$ is converted to 1 if $j + H(p)$ is a quadratic residue modulo $p$, and to 0 otherwise, were $p$ is a secret key and $H()$ is a one-way hash function. A node label sequence, $B_i$, is then generated by traversing $T_i$ according in post-order. A rank, $d_i$, is then derived for each sentence for $s_i$ using $d_i = H(B_i) \,\mathrm{XOR}\, H(p)$ and the sentences are sorted by rank. Starting from the least-ranked sentence $s_j$, the watermark is inserted to $s_j$'s successor in the text. The sentence $s_j$ is referred as a *marker* sentence, since it points to a watermark carrying

sentence. Watermark insertion continues with the next sentence in the rank ordered list. Once the sentences to embed watermark bits are selected, the bits are stored by applying either *syntactic* or *semantic transformations*, which were are explained in detail in Section 3.2.

## 6. EVALUATION OF NATURAL LANGUAGE WATERMARKING SYSTEMS

Evaluation of NL watermarking algorithms present unique and difficult challenges compared to the evaluation of image or video watermarking algorithms. The genre of the text that is being modified for watermarking has an important effect on the process of evaluation. For example, when watermarking a magazine article or a novel, the emphasis may be on the preservation of the author's style. On the other hand, when watermarking a cooking recipe or a user manual, preserving the preciseness and jargon would be more important.

Most of the state of the art NL evaluation tools that were developed for evaluating the grammar and fluency of machine translation systems may be adapted to evaluate watermarking systems. One example of such an evaluation approach is the BLEU system[27] used in machine translation evaluations that uses a weighted average of variable length phrase matches against reference translations. For previous research on this topic refer to.[28]

## 7. FUTURE DIRECTIONS

Despite some advances, NL watermarking is still in its infancy compared to image watermarking. We believe that significant synergy will emerge when NL and image watermarking communities work more closely. For some aspects of NL watermarking many ideas from image watermarking can easily be adapted; for other aspects totally new approaches that can handle the discrete and recursive nature of language needs to be developed.

We believe approaches that rely on embedding information in the syntactic structure of sentences are the most promising ones for NL watermarking. Tools for syntactic analysis of text are readily available and have been tested according to very well-known benchmarks. Future NL watermarking systems should pay attention to both coherent semantics and rhetorical structure of the output text.

Evaluation of NL watermarking systems presents a much greater difficulty than that for image watermarking systems, since such evaluations need to face the thorny issues of meaning, grammaticality, and text style. Currently neither objective assessments of human perception of NL watermarked text using various algorithms nor studies on the robustness of NL watermarking schemes under attacks are available. Much work needs to be done in this area. Availability of watermarking evaluation testbeds for NL watermarking, similar to those for image watermarking[29] is a necessity in this respect.

## 8. CONCLUSIONS

Natural language watermarking using linguistics techniques is a new field of research with large potential for many applications. Currently, there are no fully functional systems beyond the proof-of-concept level, although the interest in this field has grown rapidly in recent years. There would be rapid improvements in natural language watermarking if the knowledge and expertise in image and audio watermarking can be employed with the help of collaboration with researchers in these fields.

## ACKNOWLEDGMENTS

The authors would like to thank Eugene Lin from Purdue University for his helpful comments. We also profited from valuable discussions with Prof. Mikhail J. Atallah, Prof. Victor Raskin, and Umut Topkara from Purdue University; Guiseppe Riccardi, Dilek Z. Hakkani-Tür, and Srinivas Bangalore from AT&T Labs − Research; and Owen Rambow from Columbia University for helpful discussions and feedback on NL watermarking.

## REFERENCES


1. J. T. Brassil, S. Low, and N. F. Maxemchuk, "Copyright protection for the electronic distribution of text documents," *Proceedings of the IEEE*, vol. 87, no. 1, pp. 1181– 1196, July 1999.
2. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons Inc., 2001.
3. N. Chomsky, *The Minimalist Program*. MIT Press, 1995.
4. S. Pinker, *The Language Instinct*. William Morrow and Company, 1994.
5. I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1673–1687, December 1997.
6. D. Jurafsky and J. Martin, *Speech and Language Processing*. Upper Saddle River, New Jersey: Prentice-Hall, Inc, 2000.
7. C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.
8. The Linguistic Data Consortium, "http://wave.ldc.upenn.edu/cgi-bin/ldc/agree?text."
9. C. Fellbaum, *WordNet an electronic lexical database*. MIT Press, 1998.
10. K. Kipper, H. T. Dang, and M. Palmer, "Class-based construction of a verb lexicon," *AAAI-2000 Seventeenth National Conference on Artificial Intelligence*, July 30 - August 3, 2000, Austin, TX.
11. B. Levin, *English Verb Classes and Alternations: A preliminary investigation*. Chicago, IL: University of Chicago Press, 1993.
12. N. Ide and J. Vronis, "Word sense disambiguation: The current state of the art," *Computational Linguistics*, vol. 24, no. 1, 1998.
13. M. Atallah, V. Raskin, C. F. Hempelmann, M. Karahan, R. Sion, U. Topkara, and K. E. Triezenberg, "Natural language watermarking and tamperproofing," *Proceedings of the Fifth Information Hiding Workshop*, vol. LNCS 2578, 7-9 October 2002, Noordwijkerhout, The Netherlands.
14. F. Xia and M. Palmer, "Converting dependency structures to phrase structures," *Proceedings of the Human Language Technology Conference*, 18–21 March 2001, San Diego, CA.
15. L. Bourbeau, D. Carcagno, E. Goldberg, R. Kittredge, and A. Polguère, "Bilingual generation of weather forecasts in an operations environment," *Proceedings of the Thirteenth International Conference on Computational Linguistics*, 1990, Helsinki, Finland, pp. 318–320.
16. "http://www.fb10.uni-bremen.de/anglistik/langpro/nlg-table/nlg-table-root.htm."
17. R. Barzilay and L. Lee, "Learning to paraphrase: An unsupervised approach using multiple-sequence alignment," *Proceedings of NAACL Human Language Technology Conference*, 2003, Edmonton, Canada.
18. B. Pang, K. Knight, and D. Marcu, "Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences," *Proceedings of NAACL Human Language Technology Conference*, 2003, Edmonton, Canada.
19. P. Wayner, "Mimic functions," *CRYPTOLOGIA*, vol. XVI, no. 3, pp. 193–214, July 1992.
20. "The tyrannosaurus lex system available at http://www.fb10.uni-bremen.de/anglistik/langpro/nlg-table/nlg-table-root.htm."
21. R. Bergmair, "Towards linguistic steganography: A systematic investigation of approaches, systems, and issues.," tech. rep., University of Derby, August 2004.
22. M. Chapman and G. Davida, "Hiding the hidden: A software system for concealing ciphertext in innocuous text," *Proceedings of the International Conference on Information and Communications Security*, vol. LNCS 1334, 1997, Beijing, China.
23. M. Chapman and G. Davida, "Plausible deniability using automated linguistic stegonagraphy," *roceedings of the International Conference on Infrastructure Security*, October 1-3 2002, Bristol, UK, pp. 276–287.
24. M. Atallah, C. McDonough, S. Nirenburg, and V. Raskin, "Natural Language Processing for Information Assurance and Security: An Overview and Implementations," *Proceedings 9th ACM/SIGSAC New Security Paradigms Workshop*, September, 2000, Cork, Ireland, pp. 51–65.
25. M. Atallah, V. Raskin, M. C. Crogan, C. F. Hempelmann, F. Kerschbaum, D. Mohamed, and S. Naik, "Natural language watermarking: Design, analysis, and a proof-of-concept implementation," *Proceedings of the Fourth Information Hiding Workshop*, vol. LNCS 2137, 25-27 April 2001, Pittsburgh, PA.
26. S. Nirenburg and V. Raskin, *Ontological Semantics*. MIT Press, 2004.

27. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," *Proceedings of 40th Annual Meeting of the ACL*, July 2002, Philedelphia.
28. E. Hovy, M. King, and A. Popescu-Belis, "Principles of context-based machine translation evaluation," *Machine Translation*, vol. 16, pp. 1–33, 2002.
29. H. C. Kim, H. Ogunleye, O. Guitart, and E. J. Delp, "Watermarking evaluation testbed (wet) at purdue university," *Proceedings of the SPIE Conference on Security, Steganography, and Watermarking of Multimedia Contents VI*, vol. SPIE 5306, 1822 January 2004, San Jose, CA.