Journal of Information Hiding and Multimedia Signal Processing ©2010 ISSN 2073-4212
Ubiquitous International Volume **1**, Number **3**, July **2010**

NORMALS: Normal Linguistic Steganography Methodology

Abdelrahman Desoky

Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
abd1@umbc.edu

Received January 2010; revised May 2010

Abstract. *Text-cover of contemporary linguistic steganography approaches has numerous flaws such as incorrect syntax, lexicon, rhetoric, and grammar. Additionally, the content of text-cover is often meaningless and semantically incoherent. Such detectable noise (flaws) by both human and machine examinations can easily raise suspicion. These deficiencies render contemporary approaches highly vulnerable. Unlike all other approaches, the Normal Linguistic Steganography Methodology (NORMALS) neither generates noise nor uses noisy text to camouflage data. NORMALS employs Natural Language Generation (NLG) techniques to generate noiseless (flawless) and legitimate text-cover by manipulating the inputs' parameters of NLG system in order to camouflage data in the generated text. As a result, NORMALS is capable of fooling both human and machine examinations. Unlike Matlist, NORMALS is capable of handling non-random series domains. The implementation, validation, and experimental results of the NORMALS methodology are demonstrated in this paper.*

1. **Introduction.** Linguistic steganography is the scientific art of avoiding the conception of suspicion in covert communications by concealing data in a linguistic-based textual cover. The goal is not to hinder the adversary from decoding the hidden message, but to prevent the arousal of suspicion in covert communications. Fundamentally, when using any steganographic technique if suspicion is raised, the goal of steganography is defeated regardless of whether or not a plaintext is revealed. Contemporary linguistic steganography approaches are not fully capable of passing both computer and human examinations. Particularly, there are no linguistic approaches that fabricate an entire text-cover that are proven to pass both computer and human examinations, as detailed in Section 2. If the contemporary linguistic approaches can fool a computer examination as acclaimed, fooling the human examination may prove to be more difficult, if not impossible. Furthermore, one may argue that if humans can detect a hidden message, then most likely it is feasible to employ artificial intelligence to play the role of human detection. Nonetheless, the inability of contemporary linguistic steganography approaches to pass both computer and human examinations is because these approaches generate numerous detectable flaws (noise), such as incorrect syntax, lexicon, rhetoric, grammar and the content of the linguistic-cover is meaningless and semantically incoherent. Obviously, such flaws can raise suspicion during covert communications. Not enough attention is given to these deficiencies. Additionally, all contemporary efforts are focused on how to hide a message and not on how to hide the transmittal of a hidden message. A successful

linguistic steganography approach must be capable of passing both computer and human examinations.

Recently, the Matlist methodology [1] was introduced to resolve the linguistics flaws of contemporary steganography approaches. This methodology is based on a random series (e.g. random series of binary, decimal, hexadecimal, octal, alphabetic, alphanumeric, or any other form). Unlike Matlist, NORMALS is not based on a random series. For instance, Matlist cannot employ a domain-specific subject that contains inadequate amount of random series such as smoking cessation.

Therefore, Normal Linguistic Steganography Methodology (NORMALS) is presented in this paper. NORMALS is the scientific art of avoiding the conception of suspicion in covert communications by employing Natural Language Generation (NLG) techniques to conceal data. NORMALS overcomes all the steganographic vulnerabilities, linguistic flaws, and limitation issues of all other contemporary linguistic steganography approaches. Contemporary NLG systems generate text that is meaningful, syntactically correct, lexically valid, rhetorically sound, semantically coherent, and legitimate. Therefore, NORMALS takes advantage of NLG techniques to generate the NORMALS Cover (text-cover) by manipulating the inputs' parameters of NLG system in order to camouflage data in the generated text. The vast number of domain-specific subjects that are not based on a random series can be applied by NORMALS. In other words, when communicating parties want to employ a domain-specific subject that is not based on a random series, then Matlist is not an option; instead, NORMALS can be employed.

Some of the main advantages of the NORMALS methodology over all other approaches are as follows. The tremendous amount of text in electronic and non-electronic format makes it impossible for an adversary to investigate them all. This makes it extremely favorable as a steganographic cover in covert communications. NORMALS is resilient against all contemporary attacks including an attack by an adversary who knows all the details of NORMALS (NORMALS is a public methodology). Linguistically, the NORMALS Cover is meaningful, syntactically correct, lexically valid, rhetorically sound, semantically coherent, and looks legitimate. The hidden message is anti-distortion. There is plenty of room to conceal a message using NLG systems, as will be demonstrated in the implementation section.

The remainder of this paper is organized as follows: Section 2 describes the related work; Section 3 introduces the NORMALS methodology; Section 4 demonstrates NORMALS' implementation; Section 5 demonstrates the steganalysis validation of NORMALS and experimental results; Section 6 concludes the paper and highlights directions for future research.

2. **Related Work.**

The output of both linguistic steganographical schemes and Natural Language Generation (NLG) systems is text. However, their goals are totally different. The goal of linguistic steganographical schemes is to conceal information in non-legitimate text to communicate covertly. Text-cover of contemporary steganography approaches has numerous flaws such as incorrect syntax, lexicon, rhetoric, and grammar [2][3][4][5][6][7]. In addition, the content of text-cover is often meaningless and semantically incoherent [2][3][4][5][6][7]. On the other hand, the goal of NLG systems is to represent legitimate text either by an on-line-display or audio speech. The generated text by the contemporary NLG systems is meaningful, syntactically correct, lexically valid, rhetorically sound, semantically coherent, and legitimate [8]. In this section, a brief review of prior work on linguistic steganography and NLG systems that are related to NORMALS is presented.

2.1. **Linguistic Steganography.** Since the twentieth century, the progress and development of linguistic steganography has been minimal. Academically, there were roughly about five major approaches that have been introduced before Nostega-based methodologies were invented [1][17][18][19][20][21]: null cipher, mimic functions, NICETEXT, and noisebased approach e.g. translation-based, confusing approach, SMS-based. These approaches are as follows.

During World War I, the Germans communicated covertly using a series of characters and words known as null cipher [9]. A null cipher is a predetermined protocol of character and word sequence that is read according to a set of rules such as read every seventh word or read every ninth character in a message. Apparently, suspicion is raised because the user is forced to fabricate a text-cover according to a predetermined protocol that is not legitimate. Applying a brute force attack may reveal the entire message.

In 1992, Peter Wayner introduced the mimic functions approach [6][7], which employs the inverse of the Huffman Code by inputting a data stream of randomly distributed bits. The generated text by mimic functions is claimed to be resilient against statistical attacks. Mimic functions can employ the concept of both Context Free Grammars (CFG) and van Wijnaarden grammars to enhance the output. The output from regular mimic functions is gibberish rendering it extremely suspicious. The combination of mimic functions and CFG slightly improved the readability of the text. However, the text is nonsense, full of syntax errors, and semantically erroneous. As a result, the text is discernible by human eyes and detectable by computer. Furthermore, if an adversary were to guess the generation of the text-input, he may reveal the original plaintext [2].

Chapman and Davida, in 1997, introduced a steganographic scheme consisting of two functions called NICETEXT and SCRAMBLE that uses a large dictionary [10][11]. Suspicion is raised because some synonymous words are not semantically compatible. Furthermore, if the adversary has the original text and semantically analyzes it, he may detect a fingerprint of NICETEXT. This may lead the adversary to know that the template was derived from the original text [2].

Christian Grothoff et al., in 2005, introduced the translation-based steganographic scheme to hide a message in the translation errors (noise) that are naturally generated by a machine translation. Translation-based is a textual steganography approach that can be categorized as a linguistic approach. The major problem with the translation-based scheme is the fact that it cannot stand for a long period of time because of the expected progress and improvement in the machine translation [12][13][14]. More improvement of machine translations will increase the possibility of suspecting a hidden message that is concealed by the steganographic translationbased approach. This improvement of machine translation is feasible and will render translation-based approach obsolete. Another noise-based approach has been proposed by Topkara et al. that employs typos and ungrammatical abbreviations in a text, e.g. emails, blogs, forums, etc., for hiding data [15]. Moreover, Shirali- Shahreza et al. have introduced an abbreviation-based scheme [16] to conceal data using the short message service (SMS) of mobile phones. Due to size constraints of SMS and the use of phone keypad instead of the keyboard, a new language called SMS-Texting was defined to make the approach more practical. However, these approaches are sensitive to the amount of noise (errors) that occurs in a human writing. Such shortcoming not only increases the vulnerability of the approach but also narrows the margin of hiding data. Conversely, NORMALS neither employs errors nor uses noisy text to conceal data.

Recently, a new paradigm in steganography research, namely Noiseless Steganography Paradigm (Nostega) [17][18] is introduced. Nostega conceals messages in a cover as legitimate data rather than noise. A number of linguistic methodologies have been developed

based on the Nostega paradigm. These methodologies are as follows. Summarization-Based Steganography Methodology (Sumstega) [19] exploits automatic summarization techniques to camouflage data in the auto-generated summary-cover (text-cover) that looks like an ordinary and legitimate summary. List-Based Steganography Methodology (Listega) [20] manipulates itemized data to conceal messages in a form of textual list. Notes-Based Steganography Methodology (Notestega) [17][18] takes advantage of the recent advances in automatic notetaking techniques to generate a text-cover. Notestega embeds data in the natural variations among both human-notes and the outputs of automatic-notetaking techniques. Mature Linguistic Steganography Methodology (Matlist) exploits NLG and template techniques along with Random Series values (RS) to camouflage data without generating any suspicious pattern. Matlist employs a particular domain-specific subject such as finance, medicine, science, economics, etc. The qualified domainspecific subject is based on a random series of binary, decimal, hexadecimal, octal, alphabetic, alphanumeric, or any other form.

It is worth noting that the presented NORMALS methodology in this paper follows this new paradigm, Nostega, by exploiting NLG and template techniques that are not based on Random Series values (RS) to camouflage data without generating any suspicious pattern.

2.2. **Natural Language Generation Systems.** NLG is the process of employing a non-linguistic data input to produce an understandable text for both humans and machines. NLG employs knowledge base, artificial intelligence, computational linguistics, and other related techniques to achieve its goal [8][24]. Contemporary NLG techniques employ the knowledge of a domainspecific subject [8] and its linguistics to generate texts in a form of reports, assistance messages, documents, and other desirable text. Contemporary NLG systems generate a mature linguistic text [8][24]. In other words, NLG generates text that is meaningful, syntactically correct, lexically valid, rhetorically sound, semantically coherent, and legitimate. Some examples of NLG systems are WeatherReporter [8][25], FoG [8][25], and StockReporter [24][26]. WeatherReporter and FoG generate a textual weather description. The data input to these schemes is a numerical random series and the domain-specific subject is the weather. This numerical random series represents the numerical weather data and the generated text by these systems describes the changes in weather. However, FoG is more mature than WeatherReporter, and it can generate a textual weather description in two different languages, English and French. Another example of an NLG system is the StockReporter which was formerly known as the Ana scheme. The data input to the StockReporter scheme is a numerical random series and the domain-specific subject is the stock market prices. The numerical random series represents the values of key stocks, and the generated text describes the fluctuations in stock market prices.

The template techniques were formerly known as mail-merge technology [8]. Mail-merge techniques have been employed in software packages such as Microsoft Word and others. The core idea of mail-merge is as simple as "fill in the blank" by employing a predetermined template. Generic mail-merge can generate various text based on its input. Theoretically, NLG and mail-merge systems are equivalent in the sense of the functionality. Practically, any task that can be done by NLG systems can also be achieved by mail-merge systems and vice versa. It is argued that mail-merge techniques are NLG techniques [8]. However, from a complexity point of view, the NLG systems are a step ahead of mail-merge systems. The field of NLG systems has enjoyed significant progress in recent years and is still promising more in the future.

3. **NORMALS Methodology.** Bob and Alice are on a spy mission. Before they went on their mission, which requires them to reside in two different countries, they plot their strategic plan and set the rules for communicating covertly using their professions as a steganographic tool. To make this work, they establish a business relationship as follows. Bob and Alice are smoking cessation consultants working for the same corporation, and they agree to use a steganographic text-cover. When Bob wants to send a covert message to Alice, Bob either posts counseling related documents online for authorized clients and staff to access or he sends counseling-related documents via email to the intended clients and staff. These counseling-related documents conceal a message. Covert messages transmitted in this manner will not look suspicious because after all both Bob and Alice are smoking cessation consultants and everything looks legitimate. Furthermore, Bob and Alice are not the sole recipients. There are other non-spy smoking cessation consultants and clients who send and receive such documents, further warding off suspicion.

However, only Bob and Alice will be able unravel the hidden message because they know the rules of the game. When Alice and Bob communicate, they can use real data from their professions and their established business relationship to make their covert communications legitimate. If real data is not used, then untraceable data can be fabricated to avoid comparison attack if an adversary attempts to trace data and compare it to authenticated data. This will avert the conception of suspicion.

The above scenario demonstrates how NORMALS methodology can be used. NORMALS methodology is demonstrated in the remainder of this section in detail.

3.1. **NORMALS Architecture.** The generated text by Natural Language Generation System (NLGS) is meaningful, syntactically correct, lexically valid, rhetorically sound, semantically coherent, and legitimate [8]. Therefore, NORMALS takes advantage of the NLG techniques to generate the NORMALS Cover (text-cover) by employing NLGS. Briefly, NORMALS employs NLG techniques to generate flawless and legitimate linguistic-cover by manipulating the inputs' parameters of NLG system in order to camouflage data in the generated text. Linguistically, NORMALS Cover inherits the same qualities of the generated text by NLGS rendering NORMALS Cover noiseless and legitimate. As a result, NORMALS is capable of fooling both human and machine examinations. The NLGS has plenty of room to conceal a message and allows the communicating parties to establish a covert channel to transmit the hidden message. NORMALS achieves the steganographical goal through three major components, as shown in Figure 1: the NLGS, the steganographical encoder, and the communication protocol. NORMALS' components are detailed in the following subsections (Section 3.1.1 to 3.1.3).

Once NORMALS scheme is constructed, the steganographical communications will be accomplished in three steps: first, NORMALS generates the required NORMALS Code; second, NORMALS Code will serve as input to the NORMALS NLGS to generate the NORMALS Cover; and third, NORMALS covertly transmits the hidden message through a covert channel.

The following briefly describes an overview of NORMALS components. As stated above, each component will be elaborated on in the following subsections (Section 3.1.1 to 3.1.3).

NORMALS consist of three major components, as shown in Figure 1:

1. **NORMALS NLGS** is responsible for generating NRMALS Cover.
2. **NORMALS Encoder** encodes the message in the form of NLGS' inputs.
3. **NORMALS Communications Protocol (NCP)** responsible for how the intended users will communicate covertly to achieve the steganographical goal.

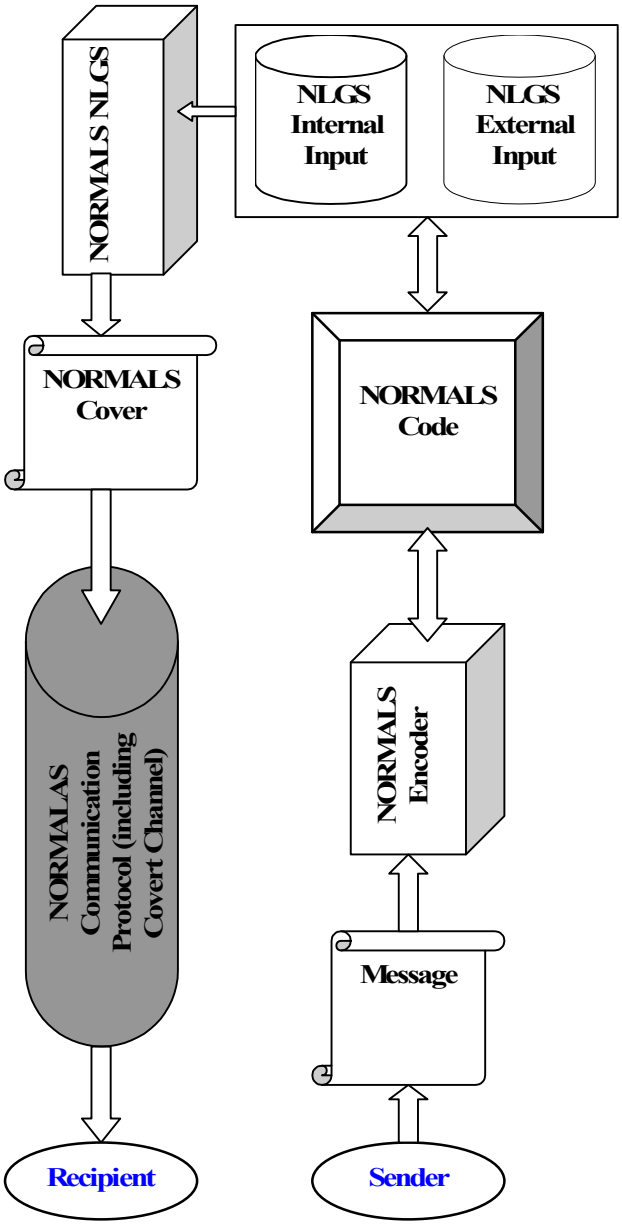


FIGURE 1. Illustrates NORMALS Architecture.

3.1.1. *NORMALS NLGS.* Legitimate users have to predetermine the NORMALS NLGS, as shown in Figure 1. Predetermining a particular NLGS is the initial stage of constructing a NORMALS scheme. From the steganographical point view, one of the major criteria for either selecting or implementing an NLGS, used by NORMALS, is that the output (the generated text) of NLGS must be free of contradictions with itself, with old data of the communicating parties, and with public authenticated data. For example, a weather report or a stock report cannot be used because they depend on public data that may cause a detectable noise. The noise can be contradictions in the content of a single output (the same text-cover contradicts with itself), the output versus the accessible authenticated data, or the output versus old data of the same the communicating parties. Unlike Matlist [1], NORMALS does not depend on a random series. Therefore, if a particular domain-specific subject is not based on a random series, then Matlist cannot be applied. Conversely, NORMALS can be applied on domains that are not based on a random series.

There is no need for keeping NORMALS NLGS secret, and it also can employ a contemporary public NLGS as will be demonstrated in the implementation section. Furthermore, there is no need for altering or modifying the NORMALS NLGS if a contemporary NLGS is employed. This will make an adversary's job horrendous because NORMALS NLGS is well-known as a non-steganographical scheme. Since there is no need for altering the generated text to conceal a message, suspicion will never be raised because the generated text is unaltered and linguistically flawless. To emphasize, a particular generated text for one recipient that is a non-legitimate user means as whatever the content of plaintext states. On the other hand, the identical generated text for a legitimate recipient means there is a hidden message. In such a case, it is infeasible for an adversary to suspect or detect a hidden message. Nonetheless, implementing or modifying a contemporary NLGS is feasible, and as long as the use of the NORMALS NLGS among legitimate users is well planned, the adversary neither will suspect nor detect a hidden message. Modifying a contemporary NLGS to be used by NORMALS should appear as a non-steganographical scheme. This can be accomplished by fabricating a scenario that can avoid the conception of suspicion in covert communications. For example, if NORMALS NLGS is used by both public and legitimate users, such a scenario can convince an adversary that the NORMALS Cover (text-cover) is an innocent text, because it is generated by a non-steganographical scheme. Since the focus of this paper is the linguistic steganography, the details and the approaches of how to build NLGS are not detailed here.

3.1.2. *NORMALS Encoder.* Based on the predetermined NLGS, the legitimate users have to construct the NORMALS Encoder. The initial step of constructing NORMALS Encoder is implementing a NORMALS Code that can be used to encode a message, as shown in Figure 1. Normal Code is implemented by encoding all possible factors and parameters that can generate a text through the predetermined NLGS. In other words, encoding all possible inputs of NORMALS NLGS that can generate text will form NORMALS Code. NORMALS Code will serve as inputs of NORMALS NLGS. In this paper, NORMALS Code shall refer to the encoded message and vice versa. When a legitimate user wants to conceal a message, NORMALS Encoder will encode a message using NORMALS Code then uses the encoded message to feed the NLGS in order to generate the NORMALS Cover (text-cover). There are tremendous ways of constructing NORMALS Encoder and its code making the adversary's job horrendous.

NLGS Inputs

There are two types of NLGS inputs, as shown in Figure 1: internal-inputs, such as knowledge base; and external-inputs, such as user-inputs or machine-inputs. Machine-inputs such as an electronic device (e.g. sensors) can feed the NLGS by the required data-inputs. The external-inputs may become internal-inputs for future use. For example, updateable NLGS such as FOG or WeatherReporter collects weather information and saves it in a knowledge base for future use to be compared to its current data-inputs. Rather than abstraction explanations, answering the following question can clarify the picture of NLGS. The question is what and how is the generated text produced? Briefly, the NLG system employs knowledge base, artificial intelligence, computational linguistics, and other related techniques to achieve its goal [8][24]. Contemporary NLG techniques employ the knowledge of a domain-specific subject [8] and its linguistics to generate text in a form of reports, assistance messages, documents, and other desirable text. Contemporary NLG systems generate a mature linguistic text [8][24]. This process of generating text is based on both internal-inputs and external-inputs, as stated earlier. For example,

letter-generator [8], STOP [8], and Chessmaster [27] are NLGS. Chessmaster is a software for playing and teaching chess. It has internal-inputs such as the knowledge base of most professional games and their analyses. If a user runs a particular game and asks Chessmaster about a different move rather than an authenticated move, then Chessmaster will respond with both text and audio-voice. The audio-voice is just reading the generated text. This audio-voice (the generated text) is the Chessmaster's response to the user's question explaining the analysis of a particular move. This example of Chessmaster represents both internalinput which is the authenticated game and external-input which is a different move rather than an authenticated move. The other example, letter-generator also has both internal-inputs and external-inputs similar to the Chessmaster. However, the external-inputs are done through either a questionnaire or selector. A particular user will either answer the questionnaire or select the desirable answer or parameters. Based on the user's response, letter generator will produce the desirable letter (text). For more examples of NLGS refer to [8]. From the steganographical point of view, internal-inputs, external-inputs or both can be encoded to generate NORMALS Code. However, one of the major criteria for implementing NORMALS Code is that the generated text (textcover) by NORMALS Code has to be free of contradictions with itself, with old data (that is used by the communicating parties in the past), and with the public authenticated data.

3.1.3. *NORMALS Communications Protocol.* Steganography is a Greek word which means "cover writing" [1][5]. When defining "Steganography Science"' as the scientific art of hiding a message, suspicion can still be raised and the goal of steganography will be defeated. Covert communications is done through two steps: concealing a message and then transmitting the hidden message. Contemporary steganography approaches are focused on how to hide a message and not on how to hide the transmittal of a hidden message. Concealing the transmittal of a hidden message is as important as concealing a message. Consider the following scenario, a sender when communicating covertly always uses the same steganographic technique and the same steganographic cover type (e.g. Mimic Functions, Translationbased, image-based, or audio-based). Furthermore, the sender always uses email to deliver a hidden message. Covert communications using the same steganographic technique, cover type, and email transmission all the time, will raise suspicion. An adversary overseeing this type of communications will be flagged and suspicion will be raised. Suspicion is raised because the adversary will wonder why the emails always contain one of the following: a fingerprint of Mimic Functions, a translated document, an image, or an audio file. If the sender has no legitimate reason for sending an email containing one of the mentioned items, suspicion can be raised even if the content does not look suspicious and nothing is detected. It is unusual for someone to send such content by email all the time. Suspicion is raised because of the way of delivering the hidden message not because of a vulnerable hiding technique used. However, it is more convincing when a sender has a website and posts a hidden message on it for a recipient to retrieve rather than sending the message through frequent emails. Another example, a sender in the financial industry has a legitimate reason for distributing a price analysis graph. Suspicion will not be raised if a message is concealed in the graph because of the legitimacy of distributing financial graphs. On the other hand, if the graph is a medical report, suspicion will be raised because the sender has no legitimate reason for sending a medical report. To emphasize, the way of delivering the hidden message can raise suspicion even if using a secure hiding technique. NORMALS averts the suspicion that may arise during the transmittal of a hidden message by camouflaging the transmittal of a hidden message. Therefore, NORMALS methodology imposes that the intended users make the

appropriate arrangements, techniques, policy, rules and any other related specifications for achieving the steganographical goal.

NORMALS Communications Protocol (NCP) works in the following way, as shown in Figure 2. A sender and a recipient communicate covertly using NORMALS, and they agree to the following:

A. The particular specifications and configurations of NORMALS scheme and its Decoder.
B. The particular specifications, configurations, policy, arrangements, and techniques of establishing the covert channel for the users to communicate covertly.

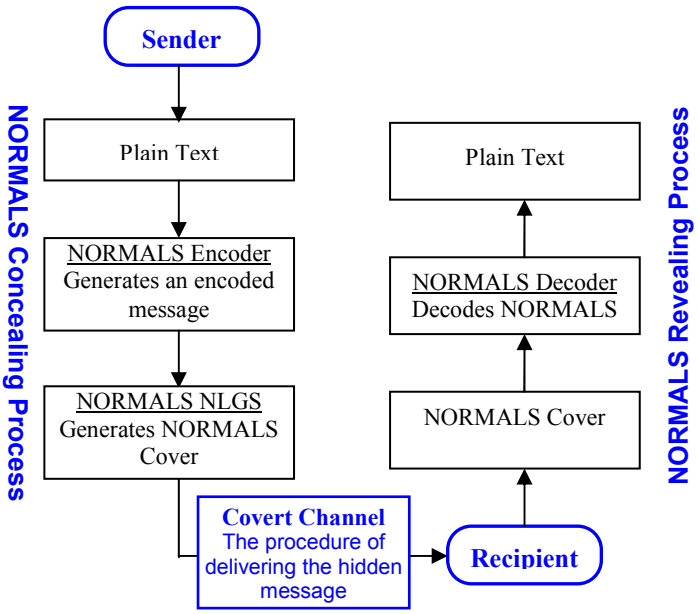


FIGURE 2. Illustrates NORMALS Communications Protocol (NCP) between sender and recipient.

Based on the agreed upon NCP, the intended users are ready to communicate covertly with each other using NORMALS.

4. **NORMALS Implementation.** This section demonstrates an actual implementation of the NORMALS methodology. Note that the techniques presented in this paper are just an example of possible implementation, but NORMALS methodology can be implemented differently. Obviously, NORMALS can easily employ other systems than the NLG system used. In other words, NORMALS can employs other NLG system for different domain other than smoking cessation domain [8]. Nonetheless, the presented demonstration shows the capability and flexibility of achieving the steganographical goal while employing the NORMALS methodology. To emphasize, the purpose of the presented implementation is to show the NORMALS' capability of achieving the steganographic goal rather than making the adversary's task difficult to decode a message. Employing a hard encoding system or cryptosystem to protect a message is feasible and simple using any contemporary encoder or cryptosystem [29][34]. However, this is not the focus of this paper. Therefore, no cryptosystem is used in this paper. The implementation of NORMALS scheme can be achieved through three stages to predetermine the following modules: first, NORMALS NLGS; second, NORMALS Encoder; and third, NORMALS Communications Protocol (NCP). These stages are detailed in following subsections (Section 4.1 to 4.3).

4.1. **Predetermining NORMALS NLGS (Stage 1).** Legitimate users employ a contemporary public NLGS called STOP [8], as shown in Figure 3, 4, A1 (Figure A1 is the entire NLGS in the Appendix A). STOP, an NLG system, was made for helping people stop smoking. The output (the generated text) of STOP is free of contradictions with itself, with old data of the communicating parties, and with the public authenticated data. NORMALS is capable of employing the STOP NLGS although it is not based on a random series.

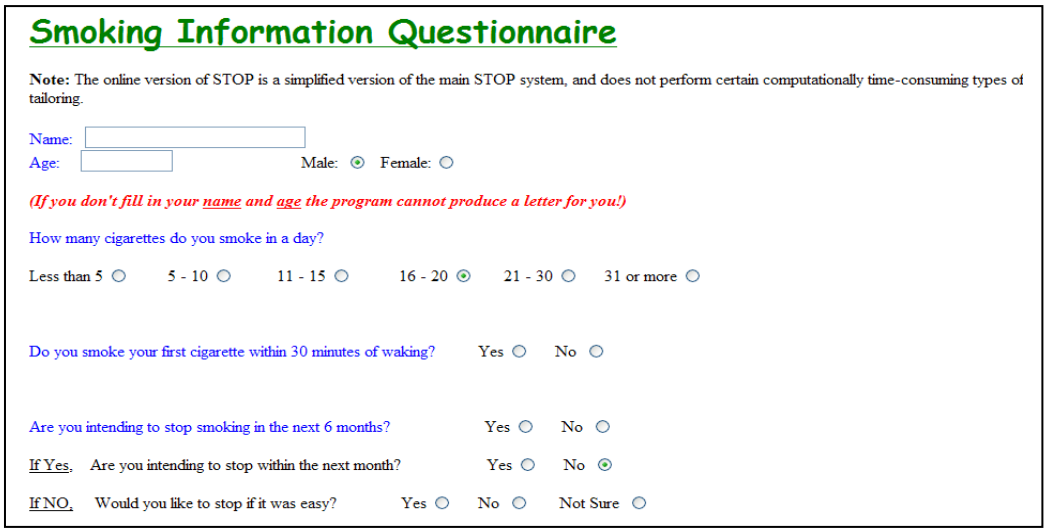


FIGURE 3. Illustrates a piece of NORMALS NLGS. The NLGS is called Stop Smoking and is made public for helping people stop smoking. Neither the Stop Smoking scheme (the NLGS) is modified nor is its output (the generated text) altered. This will render an adversary's job extremely difficult by making it unfeasible to suspect or detect a hidden message.

Not only is it that there is no need for keeping NORMALS NLGS (STOP NLGS) secret, but also it is a contemporary public NLGS as demonstrated. Additionally, the NORMALS NLGS presented in this section is not a steganographical scheme as shown in Figure 3. Furthermore, the NORMALS NLGS is neither altered nor modified. Due to this fact, this will make an adversary's job horrendous because the NORMALS NLGS is well known to the public as a non-steganographical scheme. Moreover, because the generated text by the NORMALS NLGS is unaltered and linguistically flawless, suspicion will never be raised. To emphasize, as stated earlier, a particular generated text for one recipient that is a non-legitimate user means whatever the content of plaintext states. On the other hand, the same generated text for another recipient that is a legitimate user means there is a hidden message. In such a case, it is infeasible for an adversary to suspect or detect a hidden message. As stated earlier, since the focus of this paper is the linguistic steganography, the details and the approaches of how to build NLGS are not detailed here.

4.2. **Predetermining NORMALS Encoder (Stage 2).** Based on the predetermined NLGS (STOP NLGS), the legitimate users have to construct the NORMALS Encoder [29][30][31]. As stated earlier, the initial step of constructing NORMALS Encoder is implementing a NORMALS Code that can be used to encode a message. NORMALS Encoder employs a Normal Code to encode a message simply by encoding all possible factors that can generate a text by the predetermined NLGS. In other words, encoding all possible inputs of NORMALS NLGS that can generate text will form NORMALS Code.

There are two types of inputs in the NORMALS NLGS: internal-inputs and external-inputs. Internalinputs such as knowledge base while external-inputs such as user-inputs or machine-inputs. In this paper, Normal scheme employs only the external-inputs to implement NORMALS Code and to construct NORMALS Encoder, as shown in Figure 4. The external-inputs of NORMALS NLGS, as shown in Figure 4, are done through a questionnaire. Obviously, it is a feasible and trivial task for modifying the STOP NLGS to automate the input process (encoding message) without a questionnaire and to appear as if a user answered the questionnaire.

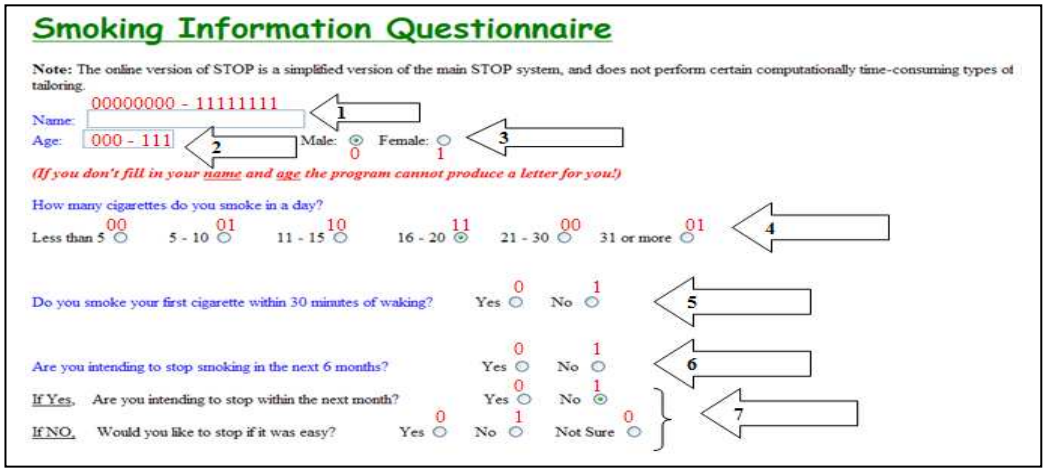


FIGURE 4. Illustrates a piece of NORMALS Encoder where all external inputs are encoded to predetermine the NORMALS Code. Unlike Matlist, NORMALS does not depend on a random series as shown in the figure. The legitimate user generates a NORMALS Cover (text-cover) by answering questions that represent the message, and then the user generates the text based on the encoded message. The entire scheme is demonstrated in Appendix A.

The implementation presented in this paper of constructing both NORMALS Encoder and NORMALS Code has no effect on the generated text. Therefore, the generated text is free of contradictions with itself, old data (that is used by the communicating parties in the past), and the public authenticated data. Additionally, the generated text is identical for both cases regardless if STOP NLGS is used for smokers counseling or if it is used for steganographical proposes by NORMALS. In another words, when STOP NLGS is used to help smokers quit smoking, or when it is used for steganographical proposes, the generated text is identical.

NORMALS Cover

NORMALS' concealment process is as follows. NORMALS Encoder will encode a message using NORMALS Code then uses the encoded message (NORMALS Code) to feed the NLGS in order to generate the NORMALS Cover (text-cover), as shown in Figure 5. To emphasize, NORMALS Encoder implements NORMALS Code by encoding each answer (input) in the STOP NLGS questionnaire as shown in Figure 4. Encoding each answer (input) can be in any encoding technique. In this paper, encoding each answer (input) in the STOP NLGS questionnaire is encoded in binary to construct NORMALS

Code as shown in Figure 4. When a legitimate user wants to conceal a message, NORMALS Encoder will convert the message (plaintext) in a concatenated binary string of the ASCII representation of message. Then, NORMALS Encoder will divide this binary message on each answer. As a result, each answer will conceal a part of the message as shown in Figure 4. Simply, NORMALS Encoder selects the answers that represent the same binary string of the message. Finally, NORMALS NLGS, which in this paper as an implementation example the STOP NLGS, will generate a textcover (NORMALS Cover) based on the selected answers (inputs of STOP NLGS), as shown in Figure 4, 5. Note that there are numerous ways of constructing NORMALS Encoder and its code, which makes the adversary's job extremely difficult.

Example of NORMALS Cover

- The plain text is: *"{Come 8pm}"*
- In this paper NORMALS' Encoder converts the plaintext in a concatenated binary string of the ASCII representation of the message as follows: "01111011010000110110111101101101011001010010000000111000011100000110110101111101"
- NORMALS Encoder will divide the above binary message on each answer. As a result, each answer will conceal a part of the message, as shown in Figure 4, simply by selecting the answers that represent the same binary string of message, as shown in Figure 4 and in table 1 and 2.
- Finally, STOP NLGS will generate a text-cover (NORMALS Cover) based on the selected answers (inputs of STOP NLGS), as shown in Figure 4.

TABLE 1. Details the steganographic code of the message *"{Come 8pm}"* which also is shown in Figures 4 and A1 (Figure A1 is the entire NLGS in the Appendix A)

Inputs order number as appeared in the NLGS used in the Appendix Section.	**Number of digit that can be concealed in a single input in the NLG system used, as shown in Appendix Section.**	**Total of digits that can conceal data, as shown in Appendix Section.**	**The concatenated binary string of the ASCII representation of the message.**
1	8	8	0111 1011 (from table 2)
2	3	3	010 (2^{nd} # e.g. age 62, # 2 is the 010)
3	1	1	0
4	2	2	00
5-8	1 (for each input)	4	1101
9-12	2 (for each input)	8	10111101
13-40	1 (for each input)	28	1011010110010100100000001110
41-44	2 (for each input)	8	00011100
45-49	1 (for each input)	5	00011
50-51	2 (for each input)	4	0110
52-60	1 (for each input)	9	101111101
61	1	1	None
Total	**81**	**81**	**80 digit**

Smoking Information for Hayman Latham

Dear Hayman Latham
Thank you for taking the trouble to return the smoking questionnaire that we sent you. In it you said that you're not planning to stop smoking in the next six months. However, you would like to be a non-smoker if it was easy to stop. Many people like you have been able to stop and you could too if you really wanted to. We hope this information will be of interest to you.

It's easier to stop if you WANT to...

You like to smoke because:

- It helps to break up your working time.
- It stops you putting on weight.
- It helps you to relax.
- It helps you cope with stress.
- It is something you do when you are bored.
- It is something to do with friends and family.
- It stops you getting withdrawal symptoms.

You don't like smoking because:

- It is bad for your health.
- It makes you less fit.
- It is a bad example for children.
- It is expensive.
- It is bad for the health of those near you.
- It is unpleasant for people near you.
- It makes your clothes and breath smell.

You said you don't like smoking because it is *bad for your health*. You are right to think this.
You are less likely to have another stroke if you stop smoking. Stopping smoking for the sake of your health really does make sense. Stopping smoking would prevent your lungs getting any worse. Ex-smokers notice an improvement in their health and fitness when their lungs begin to recover. This may take a few weeks.
If you stop smoking you are less likely to get circulation problems in the future. There is no safe number of cigarettes to smoke. Even if you only smoke occasionally it is still worth giving up. You also dislike smoking because it *affects your fitness*. Giving up smoking improves your physical and mental fitness. It also increases your stamina. Another bad thing about smoking for you is that it is a *bad example to children*. This is true. Children are far more likely to smoke if those around them smoke. Stopping smoking sets a good example.
You could do it... You are right to think that if you tried to stop smoking you would have a good chance of succeeding. You have several things in your favour.

- You have stopped before for more than a month.
- You are a light smoker.
- You have good reasons for stopping smoking.
- You expect support from your workmates.

We know that all of these make it more likely that you will be able to stop. Most people who stop smoking for good have more than one attempt.
It's often easier than you think... You said in your questionnaire that you might find it difficult to stop because you would *put on weight*. A few people do put on some weight. If you did stop smoking, your appetite would improve and you would taste your food much better. Because of this it would be wise to plan in advance so that you're not reaching for the biscuit tin all the time. Remember that putting on weight is an overeating problem, not a no-smoking one. You can tackle it later with diet and exercise. You also said that you might find it difficult to stop because *your partner and a lot of your friends smoke*. When lots of people around you are smoking it can be more difficult to stop, but not impossible. Many people have managed. If you decide to stop, tell your family and friends. Some of them might want to stop as well and you can help each other. If not, think what they could do to help and ask for their support. They might decide to stop when they see that you have succeeded. For you, another difficulty with stopping is that smoking helps you cope with *stress*. Many people think that cigarettes help them cope with stress. Taking a cigarette only makes you feel better for a short while. Most ex-smokers feel calmer and more in control than they did when they were smoking.
And finally... We hope this letter will help you to think more about the benefits of stopping smoking tobacco. Many people who feel like you do now, do eventually stop smoking. Although it might be hard, if you really want to stop you will be able to do it.
With best wishes,
The SToP Team.

SMOKELINE is the Scottish helpline for stopping smoking. Calls are free and there is someone to speak to between 12 noon and 12 midnight. The phone number is: **0800 84 84 84**

FIGURE 5. Illustrates NORMALS Cover after camouflaging the message *"{Come 8pm}"*, as shown, it is meaningful, syntactically correct, lexically valid, rhetorically sound, semantically coherent, and looks legitimate. These are some of NORMALS advantages making it capable of fooling both human and computer examinations.

TABLE 2. Details the steganographic code for the first letter of the first or last name. The highlighted rows represent the encoded message.

First/Last letter of the first or last name		**NORMALS Code for the first NLGS's Input.** This encoding is done by selecting names that starts or ends by a particular letter according to this table.
A	Q	0000 (Note, A or Q takes same value which is "0000")
B	R	0001 (Note, B or R takes same value which is "0001")
C	S	0010 (Note, C or S takes same value which is "0010")
D	T	0011 (Note, D or T takes same value which is "0011")
E	U	0100 (Note, E or U takes same value which is "0100")
F	V	0101 (Note, F or V takes same value which is "0101")
G	W	0110 (Note, G or W takes same value which is "0110")
H	X	0111 (Note, H or X takes same value which is "0111")
I	Y	1000 (Note, I or Y takes same value which is "1000")
J	Z	1001 (Note, J or Z takes same value which is "1001")
K		1010 (Unique letter, which means that the value of "1010" assigned to only "K")
L		1011 (Unique letter, which means that the value of "1011" assigned to only "L")
M		1100 (Unique letter, which means that the value of "1100" assigned to only "M")
N		1101 (Unique letter, which means that the value of "1101" assigned to only "N")
O		1110 (Unique letter, which means that the value of "1110" assigned to only "O")
P		1111 (Unique letter, which means that the value of "1111" assigned to only "P")

4.3. **Predetermining An Actual NORMALS Communications Protocol (Stage 3).** Legitimate users prearranged and plotted the required scenario to avert suspicion in covert communications. Simply, the legitimate users are a counseling group helping smokers quit. The clients are required to answer an online questionnaire and to submit it. Consequently, based on the answers of each client, the NLG system will generate a letter for each client. Obviously, the legitimate users are required to submit the entire records for each client. This will legitimize the procedure of sending a group of records to headquarters. Transmitting a hidden message in this manner will avoid raising suspicion. Briefly, a sender and a recipient communicate covertly using NORMALS, and they agreed to the following: the particular specifications, configurations, policy, arrangements, and techniques of NORMALS scheme and its covert channel for delivering the hidden message.

4.4. **NORMALS Performance.** The translation-based approach [12][13][14] was revised and published in April 2006. The highest bitrate of this revised version of translation-based approach is roughly 0.33%. It was also acclaimed that the bitrate may increase in the future which is not true because the expected improvements in machine translation will not only decrease the bitrate, but also will ban the use of the translation-based approach in the future. In addition, translation-based approach, as pointed out by Grothoff et al., cannot be applied to all languages because of the fundamental structures are radically different. This generates severely incoherent and unreadable text [12][13][14]. On the contrary, NORMALS can be applied to all known languages without any exceptions while the generated list-cover is linguistically legitimate. In regards to mimic functions, its bitrate was investigated for experimental comparison and an average of 0.97% bitrate was observed using Spam Mimic scheme [28]. The text-cover generated by mimic functions has serious flaws as detailed in Section 2. Unfortunately, the bitrate of NICETEXT was never published and there is no public scheme to allow the calculation of its bitrate. However, based on the samples given in the cited papers [10][11], the bitrate calculated

by Nostega experiment [18] is approximately 0.29%.

Improving NORMALS' Bitrate:

As stated above, by encoding both internal and external inputs of the NORMALS NLGS, the bitrate definitely will be increased. In the presented NORMALS scheme, the bitrate achieved was based on the use of encoding only the external inputs. Encoding the internal-inputs and increasing the amount of linguistics used definitely will increase the bitrate. For instance, a technique such as text substitution can be employed by NORMALS where words or a combination of words can be substituted with other words or combinations of words. Maximizing the amount of linguistics and using text substitution (e.g. words, sentences, etc.) will obviously increase the bitrate.

The technique of text substitution is similar to the semantic substitution of NICETEXT [10][11]. However, semantic substitution has been used by other steganographical approaches, but unfortunately they cause detectable noise which makes their approaches fail [12][13][14]. Unlike all other steganographical approaches where the use of text substitution causes semantic errors, NORMALS does not cause any errors when it employs any text substitution techniques. NORMALS methodology is based on natural language generation techniques where these techniques ensure the production of legitimate text. NORMALS Text Substitution (NTS) is a feature in NORMALS that gives it the advantage of being flexible in generating the NORMALS Cover and it increases the NORMALS' bitrate.

Message Size:

Generally, the size of a message is a concern for most if not all steganography approaches. However, in the presented NORMALS scheme, NORMALS camouflages a long message. When a message is long, then NORMALS generates a longer text-cover e.g. distributes it in a set of client-records (the generated text-covers) using either single or multiple transmission(s). In the presented implementation example, sending a set of client-records is a common procedure in the counseling profession. On the other hand, short messages can be a bit tricky. However, NORMALS supports the camouflage of short messages by using any end of message symbols such as delimiters, null character, etc.

5. Steganalysis Validation.

The aim of this section is to show the resilience of NORMALS to all possible attacks. NORMALS is a public methodology; however, the word "public"' does not imply in this paper that the adversary has the same or entire NORMALS scheme. However, it is assumed that an adversary has the entire knowledge about the methodology of NORMALS, but he does not have the specifications of the actual implementations of the NORMALS scheme used.

5.1. Traffic Attack.

Traffic attack is the procedure of investigating and cracking steganographic communications by investigating only the communications' traffic without investigating a particular steganographic cover. If the steganographical users are communicating with each other in a visible manner by sending, accessing, or obtaining such materials when the users have no legitimate reason to do so, then suspicion can be raised without any further investigation. For example, a medical doctor communicates using weather analysis report documents with one of his patients or vice versa. This can easily raise suspicion because a medical doctor should send medical documents not weather analysis report documents. Furthermore, if the patient has no legitimate reason for receiving or sending such documents, then suspicion can also be easily raised. Traffic attack can be applied to any contemporary steganographic technique regardless of the steganographic

cover type (e.g. image-cover, audio-cover, text-cover, etc) and can achieve successful results with relatively low costs. Further investigations can be applied once suspicion is raised during a traffic attack.

The NORMALS ensures that the communicating parties establish a secure covert channel for transmitting the hidden message covertly. In other words, NORMALS naturally camouflages the delivery of a hidden message in such a way as to appear legitimate and innocent. Thus, suspicion is averted during the transmittal of a hidden message. The scenario in Section 3 demonstrates how Bob and Alice communicated in a natural way that can avert suspicion. This scenario shows how NORMALS can be effective for camouflaging the transmittal of a hidden message. When a particular text under investigation is accessed by people who have a legitimate reason to obtain such information then suspicion is averted. This is because the professions of the intended users play the role of camouflaging the delivery of hidden messages between the intended users such as the example of Bob and Alice. On the other hand, if Bob sends information that is not related to his profession (such as a "weather report") to Alice, suspicion will be raised without any further investigation. As long as there is a legitimate reason for sending and accessing this material, suspicion can be averted. As a result, the NORMALS steganographic communications will remain unseen to the adversary because, by establishing a covert channel, the delivery of a hidden message is also hidden to achieve unseen delivery of the unseen.

Investigating all similar traffics are impossible because there is an astronomical amount of these traffics to suspect, rendering NORMALS favorable as a steganography methodology to be adopted.

5.2. **Contrast Attack.** One of the intuitive sources of noise that may alert an adversary is the presence of contradictions in the text such as finding, in consumer prices index report, the value of a product edging up while saying that it has decreased. It is worth noting that the traffic analysis, discussed earlier, can also be pursued as a base for launching contrast attacks in case the data are not publicly accessible. In the later case, comparing current data against a record of old data searching for any inconsistency over some period of time can be tracked. Countering against such an attack is always a challenge because it requires consistency with data previously used over an extended period of time. Contradictions would surely raise suspicion about the existence of a hidden message. The NORMALS scheme, as demonstrated through the example in Section 4, is simply made contrast-aware in order to avert such attacks.

5.3. **Comparison Attack.** Noise, in the context of comparison attacks, reflects an alteration of authenticated data. The goal is to find any incorrect or altered data that may imply the presence of a hidden message. When NORMALS employs a public NLGS such as STOP NLGS, then an adversary can access the same NLGS, and he can generate all possible text to compare it with NORMALS Cover attempting to detect any alteration to the generated text by the STOP NLGS. Countering such an attack is vain because NORMALS methodology does not encode any message by altering authenticated data. To emphasize, whether the STOP NLGS is used by NORMALS or for nonsteganographical purposes (helping smokers to quit smoking), the generated text in both cases is identical. Therefore, there is no noise can be detected by comparison attack. Definitely, suspicion is averted during such attacks. As long as an attack is known, it is feasible to be avoided simply by constructing the NORMALS scheme to be aware of contemporary attacks [12][13][14]. For example, if the communicating parties are concerned about comparison attacks then NORMALS scheme should be made comparison-aware in order to avoid such an attack, as demonstrated in the above examples in Section 4.

5.4. **Linguistic Attack.** Linguistics examination distinguishes the text that is under attack from normal human language. Distinguishing the text from normal human language can be done through the examination of meaning, syntax, lexicon, rhetoric, semantic, coherence, and any other issues that can help to detect or suspect the existence of a hidden message. These examinations are used to determine whether or not the text that is under attack is abnormal. The generated text by the contemporary NLG systems is meaningful, syntactically correct, lexically valid, rhetorically sound, semantically coherent, grammatically correct, and legitimate [8]. Since NORMALS is based on NLG techniques, the generated text (NORMALS Cover) is free of linguistic errors. Generally, the linguistic limitation that is imposed by employing the domain-specific subject makes it possible for any NLG system to be linguistically error free. Furthermore, if there are engine errors, it should not be a concern for two reasons; first, it is feasible to resolve any implementation problem; second, nothing is concealed in errors. Therefore, it is obvious that NORMALS is capable of passing any linguistic attack by both human and machine examinations.

5.5. **Statistical Signature.** In this paper, the statistical signature (profile) of a text refers to the frequency of words and characters used. An adversary may use the statistical profile of normal text that contains no hidden message and compare it against a statistical profile of the suspected text to detect any differences. An alteration in the statistical signature of a normal text can be a possible way of detecting a noise that an adversary would watch for. Tracking statistical signatures may be an effective means for attack since it can be easily automated and combined with traffic analysis. However, NORMALS is resiliently resistant to statistical attacks as demonstrated by the experimental results below.

5.5.1. *Word Frequency.* Human language in general, and the English language in particular, have been statistically investigated [32][33] to discover their statistical properties. The most notable study on the frequency of words was done by George Kingsley Zipf [32][33]. Zipf investigated the statistical occurrences of words in the human language and in particular the English language. Based on the statistical experimental research, Zipf concluded his observation which is known as Zipf's law [32][33]. Zipf's law [32][33] states that the word frequency is inversely proportional to its rank in an overall words frequency table, which lists all words used in a text sorted in a descending order of their number of appearances. Mathematically, Zipf's law implies that $W_n \sim 1/n^a$, where W_n is the frequency of occurrence of the n^{th} ranked word and "a" is a constant that is close to 1. Based on such a mathematical relationship, a logarithmic scale plot of the number of words' appearance and their rank will yield a straight line with a slope "-a" that is close to –1. The value of "a" is found to depend on the sample size and mix. Zipf's law was originally observed on a huge bundle of textual collections containing numerous different domain-specific subjects by different authors, writing-styles, writing-fingerprints, etc. Consequently, this huge bundle of textual collections is fairly blended which causes the occurrence of approaching or reaching Zipfian of –1.

The NORMALS' experiment applied Zipf's law directly on NORMALS Cover considering the worse case scenario that an adversary knows NORMALS methodology and knows if there is a hidden message, where the hidden message is concealed. Unlike Zipf's experiment, the NORMALS experiment applied Zipf's law on a short piece of text with a unique domain-specific subject. Based on the experimental observation, as shown in Figure 6, NORMALS Cover (that contains a hidden message) holds a Zipfian slope of –0.8374. On the other The NORMALS' experiment applied Zipf's law directly on NORMALS Cover considering the worse case scenario that an adversary knows NORMALS methodology and knows if there is a hidden message, where the hidden message is concealed. Unlike

Zipf's experiment, the NORMALS experiment applied Zipf's law on a short piece of text with a unique domain-specific subject. Based on the experimental observation, as shown in Figure 6, NORMALS Cover (that contains a hidden message) holds a Zipfian slope of –0.8374. On the other hand, the unaltered authenticated data of the same domain, without a hidden message, holds a Zipfian slope with an average of –0.71518, as shown in Table 3. Furthermore, it is observed that there are two Zipfian regions, as shown in Table 3: the highest Zipfian region holds a Zipfian slope in the range of –0.8118 to –0.8993; and the lowest Zipfian region holds a Zipfian slope in the range of –0.5745 to –0.6942. In this experiment, the highest Zipfian region is in the range of –0.8118 to –0.8993 and is the closest to the ideal Zipfian of –1. Zipfian of the presented NORMALS Cover is –0.8374, which falls in the highest Zipfian region. As a result, NORMALS Cover is in a safe side of both the ideal Zipfian of –1 and the Zipfian of the same domain. Similarly, the above observation was also observed, as shown in Table 3, in a different domain-specific subject such as the Consumer Prices Index (unaltered authenticated data without hidden message), where it holds a Zipfian slope with an average of –0.74835 [1], the highest Zipfian region in the range of –0.8245 to –0.9557 [1], and the lowest Zipfian region in the range of –0.6052 to –0.7493 [1].

The conclusion of NORMALS' experiment of word frequency is as follows. Since NLGS is based on a domain-specific subject, then when applying Zipf's law, NORMALS Cover should be similar to a Zipfian slope of its domain-specific subject (the unaltered authenticated data of the same domain that contains no hidden message), and it is not required to fully obey Zipf's law (Zipfian of –1). To emphasize, if the Zipfian slope of the NORMALS' domain-specific subject (the unaltered authenticated data of the same domain that contains no hidden message) is equal to N value, then NORMALS Cover should be either equal or close to that N value. Generally, it is feasible to fool any attack as long as the attack model is known [12][13][14], simply by constructing the steganographic scheme as attack-aware [12][13][14]. Furthermore, it is feasible to alter a natural language in a way that can fool Zipf's law if it is required. Simply, NORMALS can be designed as Zipf-aware [12][13][14] since the statistical model is already known. Obviously, Normals Cover (contains hidden message) and the generated text that has no hidden message from the same NLGS will always hold identical Zipfian's value.

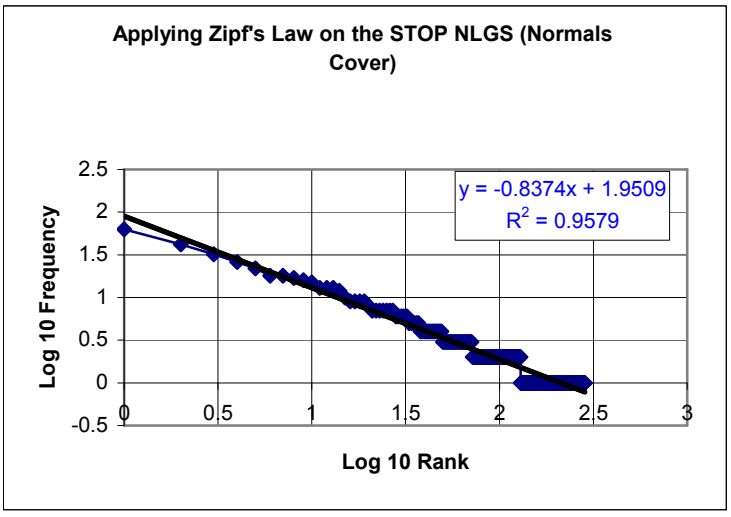


FIGURE 6. Illustrates Zipfian for the presented NORMALS Cover.

TABLE 3. The Zipfian distribution (logarithmic scale) for text without hidden message of two different domains: Helping to Quit Smoking and Consumer Prices Index (CPI). The equation is a linear curve fitting of the results. R2 is the squared error.

Text without hidden message of two different domains						
Helping to Quit Smoking				Consumer Prices Index (CPI)		
Text #	**Equation**	**R^2**	**Slope(–a)**	**Equation**	**R^2**	**Slope(–a)**
1	–0.7094x + 1.6976	0.9276	–0.7094	–0.8245x + 1.4915	0.9329	–0.8245
2	–0.6596x + 1.4729	0.9237	–0.6596	–0.8741x + 1.698	0.9467	–0.8741
3	–0.618x + 1.3766	0.9113	–0.618	–0.7412x + 1.266	0.9251	–0.7412
4	–0.7339x + 1.8687	0.9264	–0.7339	–0.8542x + 1.6855	0.9512	–0.8542
5	–0.6922x + 1.6727	0.9304	–0.6922	–0.9557x + 1.8569	0.9559	–0.9557
6	–0.6377x + 1.377	0.8922	–0.6377	–0.737x + 1.4103	0.9201	–0.737
7	–0.674x + 1.4475	0.9218	–0.674	–0.737x + 1.4103	0.9201	–0.737
8	–0.5745x + 1.3416	0.9012	–0.5745	–0.758x + 1.2825	0.9091	–0.758
9	–0.7227x + 1.6441	0.9244	–0.7227	–0.7493x + 1.428	0.9109	–0.7493
10	–0.6558x + 1.388	0.9146	–0.6558	–0.6697x + 1.4098	0.9173	–0.6697
11	–0.6141x + 1.4108	0.9145	–0.6141	–0.705x + 1.4186	0.9257	–0.705
12	–0.7221x + 1.6445	0.943	–0.7221	–0.6559x + 1.2942	0.8882	–0.6559
13	–0.8603x + 2.0621	0.9451	–0.8603	–0.7171x + 1.1889	0.9159	–0.7171
14	–0.8993x + 2.4766	0.9592	–0.8993	–0.6052x + 0.9868	0.8342	–0.6052
15	–0.899x + 2.4759	0.9591	–0.899	–0.9121x + 1.5605	0.9461	–0.9121
16	–0.6942x + 1.5498	0.9202	–0.6942	–0.8504x + 1.3719	0.9015	–0.8504
17	–0.6432x + 1.4241	0.887	–0.6432	–0.7116x + 1.3634	0.8902	–0.7116
18	–0.767x + 1.9058	0.9409	–0.767	–0.7093x + 1.363	0.9035	–0.7093
19	–0.7944x + 1.7776	0.9282	–0.7944	–0.7352x + 1.329	0.9185	–0.7352
20	–0.7018x + 1.6793	0.9279	–0.7018	–0.7085x + 1.3469	0.9021	–0.7085
21	–0.7441x + 1.9242	0.9434	–0.7441	–0.6697x + 1.4098	0.9173	–0.6697
22	–0.62x + 1.445	0.8853	–0.62	–0.6603x + 1.2676	0.8973	–0.6603
23	–0.8118x + 2.0752	0.9449	–0.8118	–0.671x + 1.3073	0.9037	–0.671
Average			–0.71518			–0.74835

5.5.2. *Letter Frequency.* Generally, in any language some letters appear at higher frequencies than others. For example, in the English language the letters "E", "T", and "A" are the most-frequently-occurring letters and "J", "Q", and "Z" appear the least. However, in some domain-specific subjects this general observation does not hold. For example, the words "judgment", "jurisdiction", "injured", "injuries", "judicial", "jury", and "subject" are used frequently in court related documents which gives the letter "J" an uncommonly high frequency. Similarly, the letter "Q" in the domain-specific subject of Queuing System (in a telecommunications field) which boosts the frequency of the letter "Q", and the letter Z in some domain-specific subjects such as Zoology have uncommonly high frequencies [34][35][36].

However, it was observed that the overall impact on the frequencies of the various letters is not that dominant since the words that increase the use of a certain letter also boost the appearance of others. Figure 7 confirms this observation by comparing the plot of the Letter Frequency Distribution (LFD) in documents from four different domain-specific subjects. The "multiple domain-specific subjects" set is based on the

2005-2006 Graduate Catalog of the University of Florida Gainesville [37], which contains over 1.4 million characters. The other sets are based on text from Queuing System (in a telecommunications field) [38], Zoology [39], and court documents [40]. The LFD of these four different sets of data are not identical but roughly obey the characteristics of the letter frequency-distribution-plot of each other, as shown in Figure 7. In other words, the peaks and valleys of each plot of the LFD closely match each other. These four different sets of data are authenticated data and not used for concealing a message.

Similarly, comparing the plot of both the LFD of the NORMALS Cover (contains hidden message) and the relative LFD of the letters in the English language (without hidden message), as shown in Figure 8, shows that both plots roughly match. In other words, the peaks and valleys of each plot of the LFD closely match each other. Obviously, Normals Cover (contains hidden message) and the generated text that has no hidden message from the same NLGS will always hold identical LFD plot.

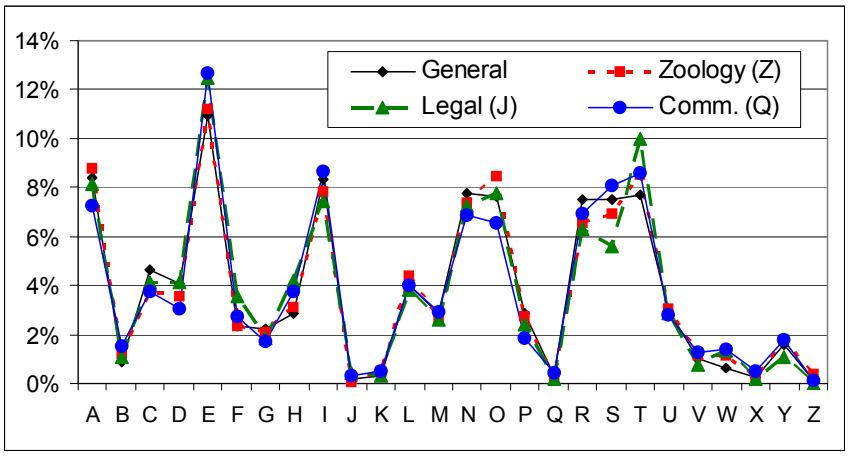


FIGURE 7. Distribution of letter usage in general and domain-specific literature.

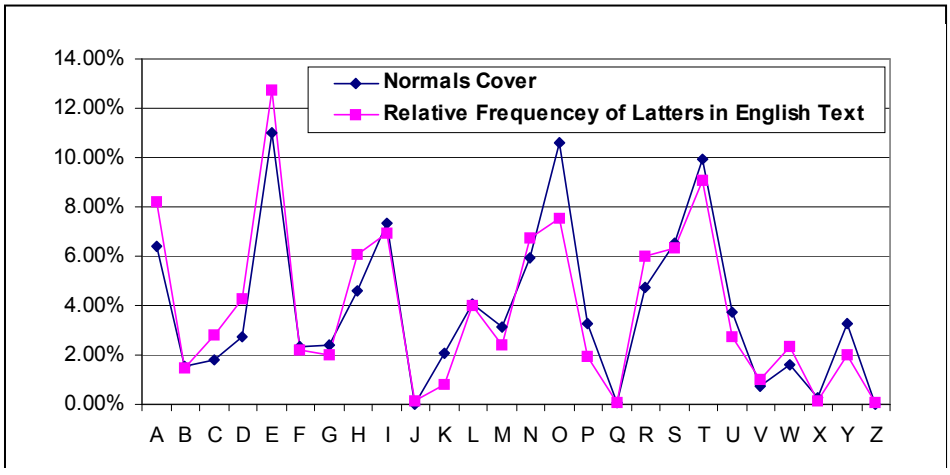


FIGURE 8. Distribution of Relative Frequency of Letters in English Text that contains no hidden message versus NORMALS Cover that contains a hidden message.

6. **Conclusions.** The generated text by NLG techniques is meaningful, syntactically correct, lexically valid, rhetorically sound, semantically coherent, and legitimate. Therefore, NORMALS takes advantage of the NLG techniques to generate NORMALS Cover

rendering it linguistically flawless (noiseless) and legitimate by manipulating the inputs' parameters of NLG system in order to camouflage data in the generated text. As a result, NORMALS is capable of fooling both human and machine examinations. NORMALS is a truly public methodology that does not rely on the secrecy of its approach. NLGS has plenty of room to conceal a message as demonstrated in this paper. To date, the NORMALS scheme presented achieves bitrate of 0.20% by encoding only the external inputs where the bitrate may differ from one NLG system to another. Obviously, by encoding both internal and external inputs of the NORMALS NLGS, the NORMALS bitrate will definitely be increased. In Matlist, the use of NLG techniques is applied to a domain-specific subject that is based on a random series (e.g. random series of binary, decimal, hexadecimal, octal, alphabetic, alphanumeric, or any other form). Inversely, NORMALS is capable of handling a non-random series domain. Regarding the translation-based approach, the continual improvement of machine translation will eliminate and ban the use of the translation-based approach. Conversely, improvement in natural language generation is promising and will make NORMALS more stable in future use. Unlike, translation-based approach, NORMALS can be applied to all known languages without any exceptions while the generated text-cover will remain linguistically legitimate. Improving the bitrate of NORMALS is feasible and worth investigating in the future.

REFERENCES

[1] A. Desoky, Matlist: Mature Linguistic Steganography Methodology, *Journal of Security and Communication Networks*.

[2] K. Bennett, Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text, *Technical Report CERIAS Tech Report*, Purdue University, 2004.

[3] N. F. Johnson and S. Katzenbeisser, A Survey of steganographic techniques, *S. Katzenbeisser and F. Petitcolas (eds.): Information Hiding*, pp. 43-78, 2000.

[4] G. C. Kessler, An Overview of Steganography for the Computer Forensics Examiner, *An edited version, issue of Forensic Science Communications. Technical Report*, vol. 6, no. 3, July 2004.

[5] F. A. P. Petitcolas, R. J. Anderson and M. G. Kuhn, Information Hiding V A survey, *Proceedings of IEEE*, vol. 87, pp. 1062-1078, July 1999.

[6] P. Wayner, Mimic Functions, *Cryptologia*, vol. 16, no. 3, pp. 193-214, July 1992.

[7] P. Wayner, *Disappearing Cryptography*, Morgan Kaufmann, pp. 81-128, 2nd Ed.2002.

[8] E. Reiter and D. Dale, *Building Natural Language Generation Systems Cambridge University Press*, 2000.

[9] D. Kahn, *The Codebreakers: The Story of Secret Writing*, revised ed. Scribner, December 1996.

[10] M. Chapman and G. Davida, Hiding the Hidden: A Software System for Concealing Ciphertext as Innocuous Text, *Proc. of International Conference on Information and Communications Security*, Lecture Notes in Computer Science, Springer, Beijing, P. R. China, vol. 1334, pp. 335-345, November 1997.

[11] M. Chapman, et al., A practical and effective approach to large-scale automated linguistic steganography, *Proc. of Information Security Conference (ISC 01)*, Malaga, Spain, pp. 156V165, 2001.

[12] C. Grothoff, et al., Translation-based steganography, *Technical Report CSD*, TR# 05-009, Purdue University, 2005.

[13] C. Grothoff, et al., Translation-based steganography. *Proc. of Information Hiding Workshop (IH 2005)*, Springer-Verlag, Barcelona, Spain, pp. 213V233, June 2005.

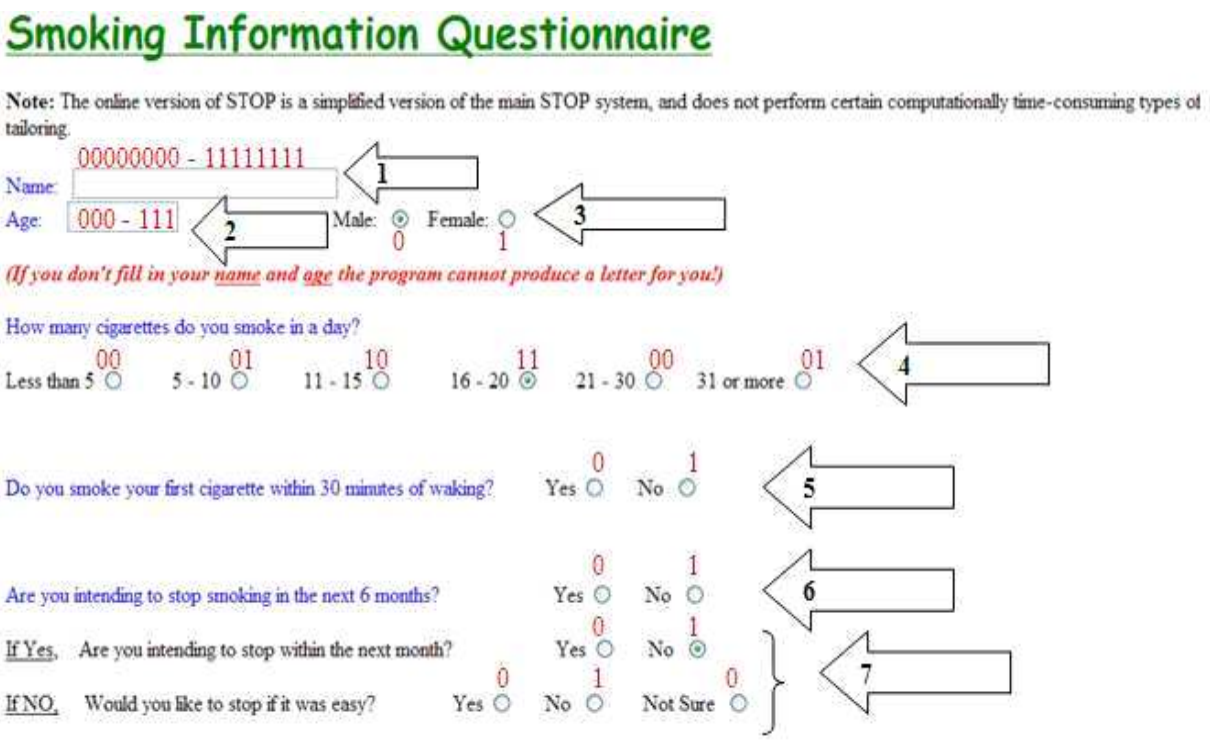
[14] R. Stutsman, et al., Lost in Just the Translation, *Proc. of the 21st Annual ACM Symposium on Applied Computing (SAC06)*, Dijon, France, April 2006.

[15] Mercan Topkara, Umut Topkara and Mikhail J. Atallah, Information hiding through errors: A confusing approach, *Proc. of SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents*, January 2007.

[16] M. Shirali-Shahreza, et al., Text Steganography in SMS, *International Conference on Convergence Information Technology*, pp. 2260 - 2265, Nov. 2007.

[17] A. Desoky, Nostega: A Novel Noiseless Steganography Paradigm, *Journal of Digital Forensic Practice*, Vol. 2, No. 3, pp. 132-139, March 2008.
[18] A. Desoky, Nostega: A Novel Noiseless Steganography Paradigm, Ph.D. Dissertation, University of Maryland, Baltimore County, May 2009.
[19] A. Desoky, et al., Auto-Summarization-Based Steganography, *Proc. of the 5th IEEE International Conference on Innovations in Information Technology*, December 2008.
[20] A. Desoky, Listega: List-Based Steganography Methodology, *International Journal of Information Security*, Springer-Verlag, April 2009.
[21] A. Desoky, Notestega: Notes-Based Steganography Methodology, *Information Security Journal: A Global Perspective*, vol. 18, No 4, pp. 178-193, January 2009.
[22] G. Kipper, *Investigator's Guide to Steganography*, CRS Press LLC, pp. 15-16, 2004.
[23] P. Davern, and M. Scott, Steganography its history and its application to computer based data files, *Technical Report Internal Report Working Paper: CA-0795*, School of Computing, Dublin City University 1995.
[24] K. Kukich, Design of a knowledge-based report generator, *Proc. of the 21st Annual Meeting of the ACL*, Massachusetts Institute of Technology, Cambridge, MA, pp. 145–150, June 1983.
[25] CoGenTex Inc., WeatherReporter
http://www.cogentex.com/
[26] Ana the Stock Reporter (StockReporter)
http://www.ics.mq.edu.au/ ltgdemo/StockReporter/about.html
[27] Chessmaster:
http://chessmaster.com
[28] Spam Mimic:
http://www.spammimic.com
[29] A. Desoky and M. Younis, PSM: Public Steganography Methodology, *Technical Report TR-CS-06-07*, Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, November 2006.
[30] A. Desoky and M. Younis, Graphstega: Graph Steganography Methodology, *Journal of Digital Forensic Practice*, vol. 2, No. 1, pp. 27-36, January 2008.
[31] A. Desoky and M. Younis, Chestega: Chess Steganography Methodology, *Journal of Security and Communication Networks*, March 2009.
[32] G. K. Zipf, (Introduction by Miller, G. A.) *The Psycho-Biology of Language: An Introduction to Dynamic Philology*, Cambridge, MA: MIT Press, 1968.
[33] W. Li, Random texts exhibit Zipf's-law-like word frequency distribution, *IEEE Trans. Information Theory*, vol. 38, no. 6, pp. 1842-1845, 1992.
[34] C. P. Pfleeger, *Security In Computing*, NJ: Prentice-Hall Inc., pp. 21-65, 2000.
[35] W. Stallings, *Cryptography and Network Security: Principles And Practices*, NJ: Prentice-Hall Inc., pp. 21- 50, 2003.
[36] R. E. Lewand, *Cryptological Mathematics*, DC: The MAA Inc., pp. 1-44, 2000.
[37] Graduate Catalog 2005-2006 University Of Florida, Gainesville:
http://gradschool.rgp.ufl.edu/currentfiles/current-catalog.pdf
[38] University of Minnesota: Modeling and Analysis of Flexible Queueing Systems
http://www.ie.umn.edu/faculty/faculty/pdf/nrl.pdf
[39] University of Otago, Postgraduate Catalog of Zoology:
http://www.otago.ac.nz/Zoology/pdf/postgraduate_handbook.pdf
[40] The Los Angeles Superior Court Civil, General Information:
http://www.lasuperiorcourt.org/civil/main.htm#3

Appendix A.
Figure A1 shows NORMALS scheme employing STOP NLGS:



If you were to try to stop smoking, how confident would you be about succeeding?

very confident	fairly confident	not confident
0	1	0

8

If you were to try to stop smoking, how ***supportive or helpful*** do you think the following people would be?

		00 Very	01 Quite	10 Not at all	11 N/A
your husband/wife/partner	9	○	○	○	◉
other members of your family	10	○	○	○	◉
your friends	11	○	○	○	◉
your workmates	12	○	○	○	◉

9 - 12

What are the ***good*** things for you about smoking?

13 - 20

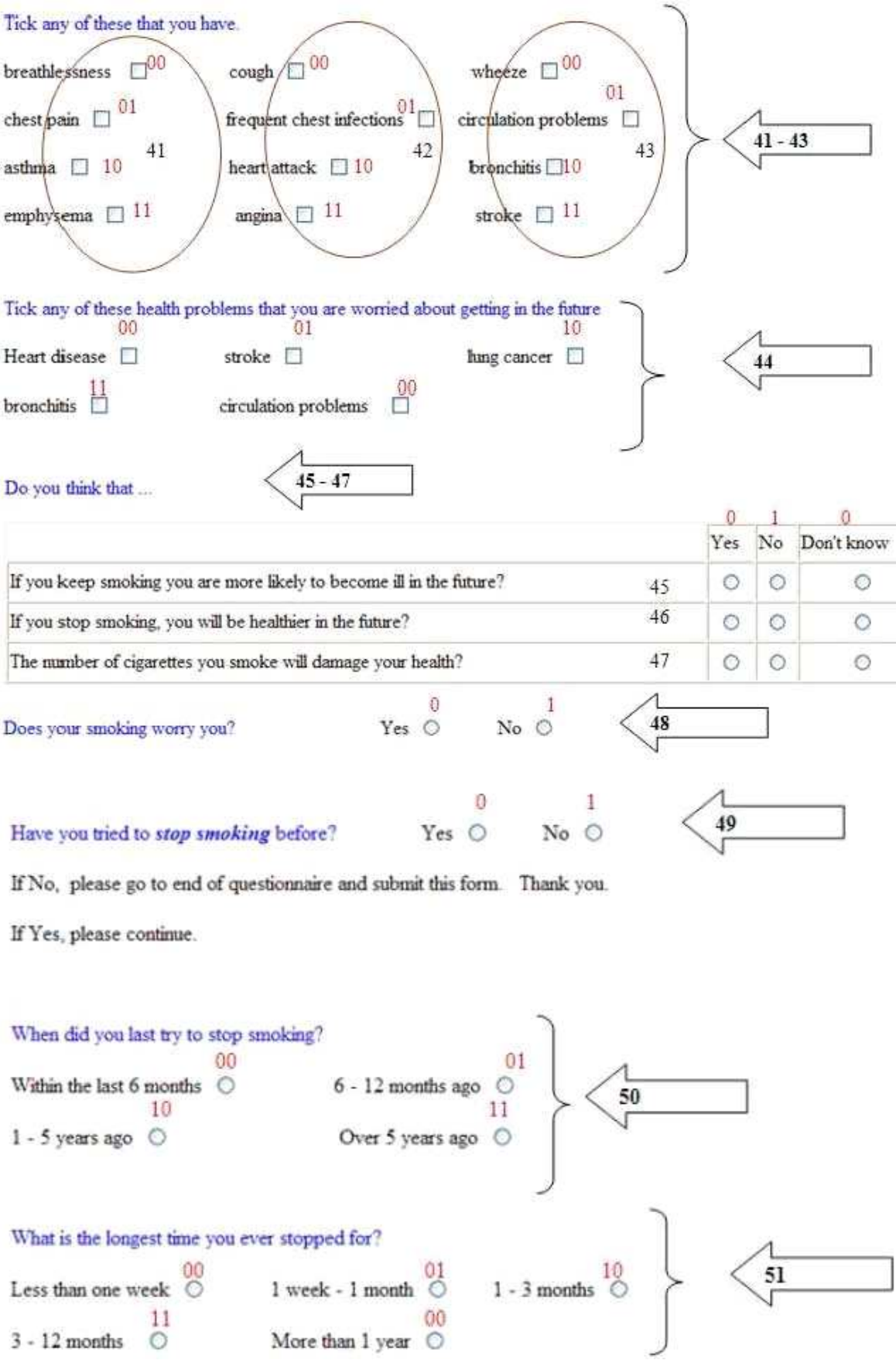
		0 Very important	1 Quite important	0 Not important
it helps me to relax	13	○	○	◉
it helps to break up my working time	14	○	○	◉
it is something to do when I am bored	15	○	○	◉
it helps me to cope with stress	16	○	○	◉
I enjoy it	17	○	○	◉
it is something I do with my friends or family	18	○	○	◉
it stops me putting on weight	19	○	○	◉
it stops me getting withdrawal symptoms	20	○	○	◉

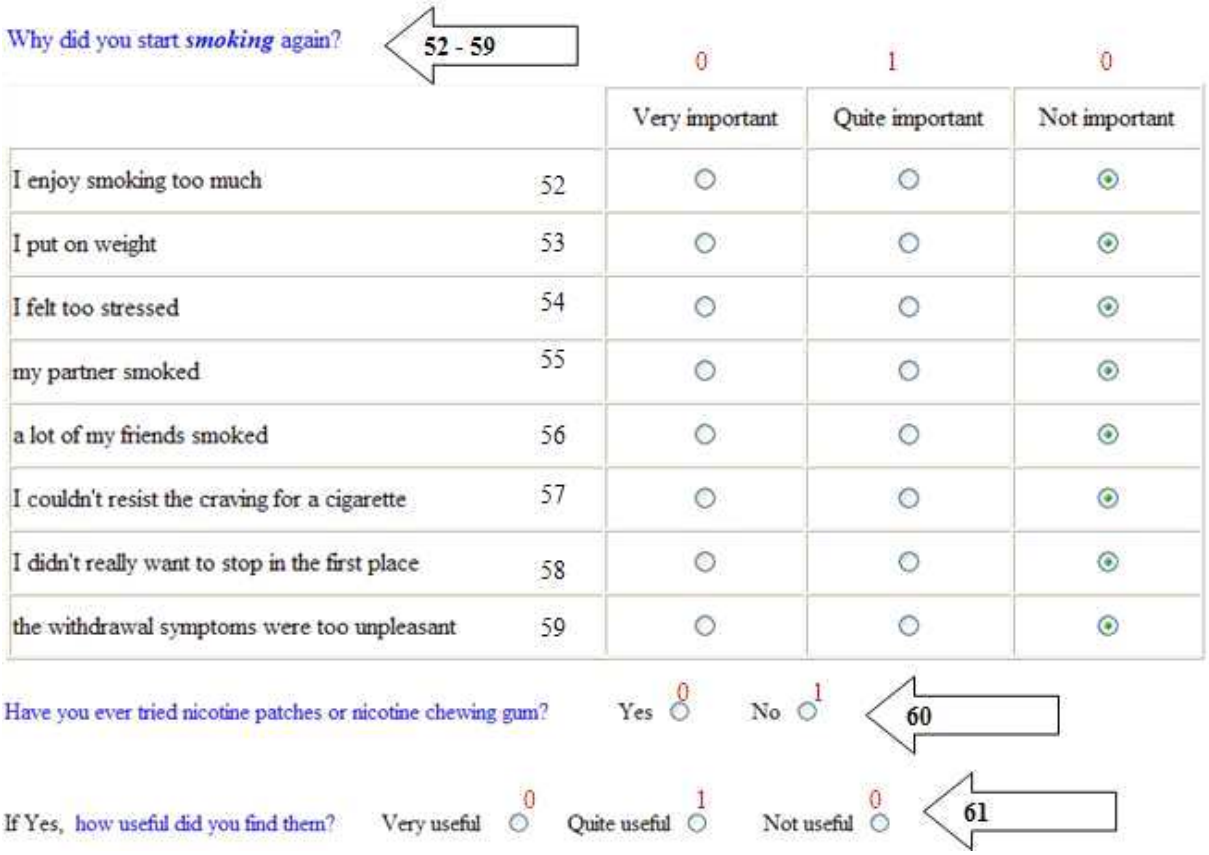
What are the things you ***don't like*** about your smoking? **21 - 29**

		0 Very important	1 Quite important	0 Not important
it is expensive	21	○	○	◉
it is bad for my health	22	○	○	◉
I don't like feeling dependent on cigarettes	23	○	○	◉
it makes my clothes and breath smell	24	○	○	◉
it is a bad example for children	25	○	○	◉
it is unpleasant for people near me	26	○	○	◉
it makes me less fit	27	○	○	◉
people around me disapprove of my smoking	28	○	○	◉
it is bad for the health of people near me	29	○	○	◉

What might make it ***difficult*** for you to stop smoking? **30 - 40**

		0 Very important	1 Quite important	0 Not important
I enjoy smoking too much	30	○	○	◉
I don't think I have enough willpower	31	○	○	◉
I think I would put on weight	32	○	○	◉
I would be too stressed	33	○	○	◉
I think I am too addicted to cigarettes	34	○	○	◉
my partner smokes	35	○	○	◉
I would miss smoking with friends	36	○	○	◉
I can't resist the craving for a cigarette	37	○	○	◉
I don't really want to stop	38	○	○	◉
I would be bored	39	○	○	◉
I would miss smoking breaks at work	40	○	○	◉


Tick any of these that you have.
breathlessness 00
chest pain 01
asthma 10
emphysema 11
41
cough 00
frequent chest infections 01
heart attack 10
angina 11
42
wheeze 00
circulation problems 01
bronchitis 10
stroke 11
43
41 - 43
Tick any of these health problems that you are worried about getting in the future
Heart disease 00
stroke 01
lung cancer 10
bronchitis 11
circulation problems 00
44
Do you think that ...
45 - 47
0 1 0
Yes No Don't know
If you keep smoking you are more likely to become ill in the future? 45
If you stop smoking, you will be healthier in the future? 46
The number of cigarettes you smoke will damage your health? 47
Does your smoking worry you?
Yes 0 No 1
48
Have you tried to stop smoking before?
Yes 0 No 1
49
If No, please go to end of questionnaire and submit this form. Thank you.
If Yes, please continue.
When did you last try to stop smoking?
Within the last 6 months 00
6 - 12 months ago 01
1 - 5 years ago 10
Over 5 years ago 11
50
What is the longest time you ever stopped for?
Less than one week 00
1 week - 1 month 01
1 - 3 months 10
3 - 12 months 11
More than 1 year 00
51



Why did you start ***smoking*** again?

52 - 59

		0 Very important	1 Quite important	0 Not important
I enjoy smoking too much	52	○	○	◉
I put on weight	53	○	○	◉
I felt too stressed	54	○	○	◉
my partner smoked	55	○	○	◉
a lot of my friends smoked	56	○	○	◉
I couldn't resist the craving for a cigarette	57	○	○	◉
I didn't really want to stop in the first place	58	○	○	◉
the withdrawal symptoms were too unpleasant	59	○	○	◉

Have you ever tried nicotine patches or nicotine chewing gum? Yes 0 ○ No 1 ○

60

If Yes, how useful did you find them? Very useful 0 ○ Quite useful 1 ○ Not useful 0 ○

61

Thank you for filling this in, please SUBMIT the form now by pressing the 'Submit' button.

Submit

Return to STOP homepage