Лексикография и прикладная лингвистика

УДК 81’322.2 + 003.26 + 004.415.24
ББК Ш100.3 + Ш11

ЛИНГВИСТИЧЕСКАЯ СТЕГАНОГРАФИЯ: СОВРЕМЕННЫЕ ПОДХОДЫ. ЧАСТЬ 1

О.И. Бабина
Южно-Уральский государственный университет, г. Челябинск

В статье определено место лингвистической стеганографии в системе методов сокрытия информации как способа внедрения тайного сообщения в текст-контейнер. Понятие собственно лингвистической стеганографии, подразумевающее при формировании стеготекста имитацию свойств текста на естественном языке, противопоставлено понятию текстовой стеганографии, где для кодирования сообщения применяются характеристики формата. Выделены уровни моделирования стеготекста, которые учитываются при порождении текстов с характеристиками, типичными для немодифицированного текста на естественном языке. Рассмотрены методы внедрения тайного сообщения в стеготекст на основе моделирования статистических и синтаксических характеристик текста. Выявлены достоинства и недостатки использования описанных методов.

Ключевые слова: лингвистическая стеганография, стеготекст, сокрытие информации, защита информации, автоматическая обработка текстов.

Введение

Стеганография часто определяется как наука и искусство сокрытия информации (сообщения) в так называемом контейнере, передача которого от одного лица другому не вызывает подозрений. В отличие от криптографии, в задачи которой входит защита собственно сообщения, основной целью стеганографии является защита самого факта наличия скрытого сообщения. На Первом семинаре по сокрытию информации в 1996 году стеганография была выделена как один из способов защиты информации от обнаружения, наряду со скрытыми каналами, а также средствами защиты от удаления информации – анонимностью и цифровыми отпечатками [17]. Несмотря на некоторую непоследовательность представленной классификации (смотри аргументацию, представленную, например, в [13]), стеганография выделилась в современную научную область (хотя практика стеганографии и сам термин насчитывает несколько столетий – термин был взят из названия книги Иоганна Тритемия и с греческого переводится как «тайнопись»), характеризующуюся современными компьютерными методами и стремительно развивающуюся в наши дни.

Стеганография охватывает широкую область исследований, так как в качестве контейнера могут выступать самые разнообразные объекты: начиная от листа бумаги, на котором сообщение записывается с помощью симпатических чернил или «встраивается» в виде микроточек, разработанных в Германии в первой половине XX века и практически используемых в военных целях (классическая, или техническая, стеганография), включая на современном этапе различные цифровые объекты – аудио, видео, текстуры 3D-объектов и другие типы файлов (цифровая стеганография), и заканчивая текстами на естественном языке (лингвистическая, или текстовая, стеганография).

При использовании лингвистической стеганографии передача скрытого сообщения производится посредством внедрения этого сообщения в некоторый текст (псевдотекст), который сам по себе не содержит «полезной» для получателя информации и для посторонних глаз выглядит «безобидным».

Следует остановиться на интерпретации «безобидности» текста: исходным предположением является то, что текст подвергается анализу на предмет содержания в нем какого-либо скрытого сообщения – стегоанализу. Текст вызывает подозрения в том случае, если он выглядит «неестественным» в определенном смысле. Например, если в тексте на русском языке содержится буква «ф» (одна из наиболее редких букв в русском языке согласно данным, полученным на материале Национального корпуса русского языка [2]) примерно так же часто, как самая частотная буква «о» [2], то частотный анализ такого текста позволит заключить, что такое распределение неестественно для текста на русском языке; значит, возникнет подозрение, что текст подвергся модификациям, в частности, с целью внедрения в него тайного сообщения.

В качестве анализатора текста на предмет содержания в нем скрытой информации может выступать человек или компьютерная программа, и, в общем случае, оценка этих типов анализаторов для

одного и того же текста может быть различной. В современных условиях, в эпоху «больших данных», первичный анализ выполняется автоматизированными средствами. В связи с этим, стеготекст (контейнер с содержащимся в нем тайным сообщением) должен, прежде всего, выдерживать стегоатаку со стороны машины, то есть алгоритмическую проверку «естественности» текста.

Общепринятым стало выделение в текстовой стеганографии трех групп, предложенных в [5]: 1) стеганография, основанная на форматировании текста; 2) случайная и статистическая генерация; 3) лингвистические методы стеганографии. Стеганография, основанная на формате, подразумевает изменение определенных свойств форматирования текста с целью скрыть в нем информацию. К таким особенностям формата относят визуальные и текстовые семаграммы, намеренное использование опечаток, капитализация, применение определенных шрифтов, решетчатые и нулевые шифры и т. д. [10]. Одним из наиболее известных способов применения текстовых семаграмм является использование пробелов между словами, предложениями или в конце строк: в зависимости от количества пробелов, кодируется 0 или 1. В качестве вариации, пробелы, кодирующие двоичную информацию, могут использоваться в тексте CSS-файлов (использование пробела до или после символа знака «:» при описании атрибутов и их значений) или HTML-разметки (наличие или отсутствие пробела в метках для передачи двоичной информации). Обзор некоторых разновидностей форматной стеганографии можно найти, например, в [3, 10, 18]. Зачастую, термин «текстовая стеганография» используется узко для того, чтобы обозначить только этот тип стеганографии. Вместе с тем, данный тип, фактически, не использует лингвистические особенности текста. Другие два типа стеганографии в большей степени ориентированы на эти особенности. В дальнейшем, под термином «лингвистическая стеганография» мы будем подразумевать лишь последние два типа.

С точки зрения лингвистической стеганографии, в моделировании текста можно выделить следующие уровни:

– статистический,
– синтаксический,
– лексико-семантический,
– онтологический.

Тогда проблему лингвистической стеганографии можно переформулировать так: как создать стеготекст, характеризующийся на каждом из этих уровней свойствами, типичными для обычного, «естественного» текста (то есть такого, который не содержит тайного сообщения)?

Учитывая особенности естественного языка и достижения в области его компьютерного моделирования, данные уровни перечислены в порядке возрастания устойчивости к проверке на наличие в тексте скрытой информации. При этом текст, который не вызовет подозрений на предыдущем уровне, не обязательно выдержит проверку при оценке на последующем уровне.

Методы современной лингвистической стеганографии нацелены на автоматизацию создания стеготекста. В общем случае, в процедуре создания стеготекста в качестве входных данных используется собственно тайное сообщение M (традиционно представленное в двоичном коде) и, опционально, текст-контейнер C (любой текст, созданный человеком, зачастую имеющийся в открытом доступе). Применяя знания из лингвистической базы знаний, сообщение M внедряется в контейнер C. Возможен также случай, когда текст-контейнер не задан. Тогда процедура создания стеготекста состоит в том, чтобы преобразовать тайное сообщение M на основе созданной лингвистической базы знаний и разработанных процедур в автоматически сгенерированный связный псевдотекст. Такой псевдотекст, как правило, может удовлетворять требованиям к «естественному» тексту на статистическом и синтаксическом уровне, но семантика, будучи наиболее сложным объектом компьютерного моделирования естественного языка и трудно поддающимся автоматизации, в таких текстах страдает. Если такой текст будет предъявлен человеку, он легко идентифицирует его «нерегулярность», установив наличие нетипичных для обычного текста, созданного человеком, свойств. Поэтому основная задача современной стеганографии заключается в решении вопроса, какова должна быть процедура автоматизированной генерации стеготекста, который в определенном смысле (при оценке на каждом из перечисленных уровней) был бы похож на «естественный» текст, созданный человеком. Именно такой стеготекст, предположительно, не вызовет подозрений у стегоанализатора (человека или машины), и цель стеганографии (скрыть факт наличия тайного сообщения) будет достигнута.

Общим свойством современных подходов к лингвистической стеганографии является использование вариативности языковых средств для внедрения тайного сообщения в текст-контейнер: информация о значении очередного бита тайного сообщения M несет выбор в пользу одного из множества альтернативных способов репрезентации единиц на различных языковых уровнях.

В данной статье рассмотрим подробнее статистические и синтаксические подходы, в основе которых лежит опора на формальные характеристики текста.

Статистический подход

Методы стеганографии, связанные со статистическим уровнем, не используют собственно лингвистические свойства текстов, и выделены в [5] в отдельный тип – случайной и статистической генерации псевдотекстов с сохранением значений статистического распределения лингвистических

единиц в результирующем тексте. Начало для этого направления (и современной лингвистической стеганографии в целом) было положено с введением понятия мимических функций [23], которые представляют собой алгоритмы для автоматической генерации текста, соответствующего определенному статистическому профилю. То есть результирующий текст должен подражать некоторому реальному тексту, написанному на естественном языке, с точки зрения статистического распределения лингвистических единиц в нем. В качестве лингвистических единиц могут выступать, например, символы или слова. Например, такой алгоритм для генерации стеготекста *T*, моделирующего статистическое распределение лингвистических единиц (символов) в некотором реальном тексте-источнике *S*, может иметь следующий вид [22, с. 89]:

1. Построить частотный список *n*-грам в тексте *S* (*n*-грам интерпретируется как последовательность из *n* лингвистических единиц, в частности, символов).

2. Выбрать любой *n*-грам и поставить его в начало генерируемого текста *T*. Этот *n*-грам не содержит кодируемой информации и используется для начала процесса порождения.

3. Повторить в цикле следующие шаги:

а) взять последние ($n - 1$) символ в тексте *T*;

б) найти в построенном на шаге 1 списке все *n*-грамы, начинающиеся с этих ($n - 1$) символов;

в) сформировать множество возможных альтернатив для выбора последующего символа для текста *T* из *n*-х символов *n*-грамов, отобранных на предыдущем шаге (3б);

г) присвоить веса элементам множества в зависимости от частот;

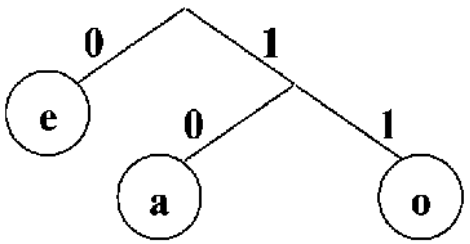
д) добавить символ из множества, учитывая его вес.

Наиболее существенная часть алгоритма заключается в выборе весов таким образом, чтобы самый частотный символ кодировал меньшее количество бит информации. Чем меньше частота возможной альтернативы, тем большее количество бит информации этот символ должен кодировать. За счет этого достигается более редкое использование в тексте *T* символов, которые реже использовались в исходном тексте *S*, и наоборот. Этот принцип реализуется в различных алгоритмах сжатия. Для его реализации при решении задачи автоматической генерации текста, имитирующего статистическое распределение лингвистических единиц некоторого реального текста, оказался весьма продуктивным алгоритм Хафмана [9]. Так, если для некоторого *n*-грама с определенным началом (первые $n - 1$ символы заданы) множество возможных альтернатив *n*-го символа включает символы «*е*» (предположим, с вероятностью 0,8), «*о*» или «*а*» (с вероятностями по 0,1), то на основании этой информации можно построить дерево Хафмана, представленное на рисунке.

Тогда если следующий бит тайного сообщения *M* содержит 0, в генерируемый текст *T* следует внести символ «*е*». Если следующий бит тайного сообщения *M* содержит 1, то необходимо взять еще один бит из сообщения *M* – тогда следующий символ текста *T* будет кодировать сразу 2 бита информации: если тайное сообщение *M* содержит последовательность «10», то следует выбрать символ «*а*», иначе – для кодирования последовательности «11» – следует использовать символ «*о*».

Существуют многочисленные примеры использования алгоритма Хафмана для моделирования статистического распределения лингвистических единиц в стеготексте (в общем случае для тех же целей используются и другие алгоритмы сжатия). Вместе с тем, строго говоря, алгоритм не позволяет точно смоделировать распределение лингвистических единиц. Фактически, при использовании построенного дерева (см. рисунок) для символов «*е*»:«*а*»:«*о*» в результирующем тексте *T* моделируется распределение вероятностей 0,5:0,25:0,25 соответственно, а не 0,8:0,1:0,1, как это следует (исходя из нашего предположения) из частот данных символов, собранных на материале текста *S*.

Кроме того, текст, сгенерированный посредством посимвольной генерации, очевидно, не является собственно текстом – это лишь последовательность, в некоторой мере сохраняющая пропорции использования символов в определенном языке. Хотя экспериментальная проверка показала, что использование модели 5-го уровня (5-грам) позволяет получить текст, в основном содержащий реальные слова языка [22], все же проверка такого текста по словарю позволит обнаружить достаточное количество фиктивных слов; применение синтаксического анализатора в ходе стегоанализа немедленно выявит отсутствие релевантной для естественной языка синтаксической структуры предложений и т. д. Поэтому в современной стеганографии принцип сохранения статистического распределения лингвистических единиц используется в дополнение к собственно лингвистическим методам построения стеготекста.



Пример построения дерева Хаффмана (из [22])

Синтаксический подход

При генерации стеготекста для преодоления недостатков посимвольной генерации, в частности, отсутствия приемлемой для естественного

языка синтаксической структуры, а также наличия фиктивных слов, в [22] предлагается подход на основе применения контекстно-свободной грамматики – понятия, которое было введено Н. Хомским для объяснения функционирования языка [7]. Контекстно-свободная грамматика Хомского G представляется как четверка:

$$G = \{T, V, P, S\},$$

где T – множество терминальных символов; V – множество нетерминальных символов; P – конечный набор правил вывода; S – начальный символ. В основе такой грамматики лежит предложенный в [1] метод разложения на непосредственные составляющие: правила вывода моделируют процесс последовательного развертывания начального символа грамматики (обозначающего предложение у Н. Хомского) в дерево непосредственных составляющих, листы которого представлены терминальными символами, а внутренние узлы – нетерминальными символами.

Для целей лингвистической стеганографии может быть построена контекстно-свободная грамматика в нормальной форме Грейбах [8] (это требование предъявляется для предотвращения зацикливания при синтаксическом разборе), каждое правило которой строится как дизъюнкция способов разложения нетерминального символа на составляющие. Количество дизъюнктов должно быть 2^n $(n \in \mathbb{N})$, при этом n определяет, сколько бит информации может быть закодировано на данном шаге. Например, грамматика может иметь вид:

S := **subject predicate**
subject := **proper** || Эта **noun**
predicate := **adjective** || не **adjective**
adjective := привлекательная || черствая
noun := куница || головка сыра
proper := Анна || Светлана

В этом примере в каждом правиле используется 2^1 дизъюнктов; первый дизъюнкт в каждом правиле кодирует 0, второй – 1. Дерево конструируется в порядке сверху-вниз слева-направо. Последовательно выбирая альтернативу, соответствующую 0 или 1 (в зависимости от скрытого сообщения), на каждом этапе выделения непосредственных составляющих кодируется один бит информации. Например, при кодировании последовательности 1110 с использованием представленной контекстно-свободной грамматики, будет сгенерирован текст «*Эта головка сыра не привлекательная*».

Представленная ранее грамматика при генерации дает 4 точки ветвления в дереве непосредственных составляющих, при этом в каждом случае выбор осуществляется из двух альтернатив; значит, грамматика может породить всего $2^4 = 16$ различных предложений. Очевидно, что текст, составленный из различных комбинаций этих 16 предложений, не отличается разнообразием. Тогда если сообщение M достаточно длинное, подобные повторы укажут на неестественную природу такого текста, и цель стеганографии не будет достигнута. Для решения этой проблемы такая грамматика может быть расширена: вместо дизъюнкции двух альтернатив, в некоторых правилах может содержаться дизъюнкция 4, 8, 16 и т. д. альтернативных способов разделения на непосредственные составляющие; тогда каждая из таких альтернатив вместо 1 бита информации будет кодировать соответственно 2, 3, 4 и т.д. бит (например, для случая 4 альтернатив – первая альтернатива кодирует последовательность 00, вторая – 01, третья – 10, четвертая – 11). Вместе с тем наиболее существенным недостатком такого способа остается трудоемкость разработки самой грамматики: усилия, потраченные на составление достаточно устойчивой для стегоанализа (т. е. большой) и в то же время непротиворечивой грамматики, могут оказаться неоправданно высокими. При этом в результате использования грамматики будут составляться синтаксически верные последовательности, которые, однако, могут не иметь приемлемой семантической интерпретации (ср. известный пример Н. Хомского «*Colorless green ideas sleep furiously*»). Кроме того, если грамматика действительно покрывает многие явления естественного языка, она неизбежно допускает множественные синтаксические интерпретации для некоторых языковых выражений, так как синтаксическая омонимия в естественном языке – нередкое явление. В результате, для некоторых предложений при автоматическом парсинге могут возникать несколько вариантов разбора, что приведет к невозможности однозначной интерпретации скрытого сообщения.

Для минимизации усилий при использовании синтаксического подхода необходима автоматизация построения синтаксического компонента модулей генерации стеготекста. Известным примером реализации этого направления является система *NICETEXT* [6], в работе которой, наряду с возможностью генерации предложений с помощью контекстно-свободной грамматики, предлагается альтернатива применения корпусного подхода для извлечения поверхностно-синтаксических шаблонов из текста на естественном языке. Так, корпус текстов (текст-источник) S может быть автоматически таггирован. Например,

Мы/мест_лич *жили*/глал *в*/предл *болотистом*/прил *крае*/сущ *близ*/предл *большой*/прил *реки*/сущ**,** *в*/предл *двадцати*/числит *милях*/сущ *от*/предл *ее*/мест_притяж *впадения*/сущ *в*/предл *море*/сущ.

На основе размеченного таким образом корпуса S автоматически собирается база синтаксических шаблонов R, которые могут включать, кроме собственно меток частей речи, знаки пунктуации, скобки, специальные символы и т.д. Так, для приведенного в примере предложения, шаблон примет вид

[*мест_лич глал предл прил сущ предл прил сущ, предл числит сущ предл мест_притяж сущ предл сущ.*].

Также на основе корпуса S составляется словарь L, в котором все вхождения w_j соотнесены с меткой части речи и произвольно каждому вхождению приписано значение 1 или 0.

Тогда процесс генерации стеготекста T представляет собой процедуру, включающую циклическое применение следующих шагов до тех пор, пока скрытое сообщение M не будет полностью закодировано:

1) произвольно выбрать синтаксический шаблон r_i из базы синтаксических шаблонов R;

2) заполнить выбранный шаблон лексическими единицами из словаря L таким образом, что:

а) все позиции частей речи шаблона r_i заменяются на вхождение w_j из лексикона L, помеченное меткой, соответствующей данной позиции в шаблоне;

б) при заполнении каждой позиции из множества слов лексикона с одинаковой меткой выбирается такое, которое соответствуют текущему кодируемому биту информации в тайном сообщении M;

3) добавить заполненный шаблон к тексту T.

Метки частей речи в примере достаточно грубые (включают лишь собственно часть речи) – при генерации текста с их помощью предложения могут оказаться аграмматичными. Поэтому реальная система должна использовать более «тонкие» метки, включающие информацию о различных грамматических признаках слова в данной позиции, например, информацию о числе, роде и падеже для существительных и прилагательных (для генерации фраз, согласованных по этим грамматическим признакам). Кроме того, метка может включать информацию семантического характера (например, для существительных – абстрактное или конкретное, одушевленное или неодушевленное, человек или не-человек и т. д.)

Моделирование синтаксического уровня текста также выполняется при встраивании тайного сообщения M посредством синтаксических трансформаций текста-субстрата C, созданного человеком и использующегося как контейнер [4, 12, 14, 16, 19, 21]. Для реализации такого подхода текст C следует подать на вход синтаксического анализатора (для полного или частичного парсинга), чтобы получить на выходе синтаксическую структуру предложения (в частности, в виде дерева непосредственных составляющих). Далее эта структура может быть подвержена синтаксическим трансформациям таким образом, что смысл в значительной степени не изменится. Трансформации, как правило, заключаются в перестановке, вставке или замещении непосредственных составляющих. При этом в языках с относительно-свободным порядком слов практически любые перестановки составляющих дают грамматически верные конструкции (например, такие трансформации используются в качестве инструмента лингвистической стеганографии в [15] для турецкого и в [11] для греческого языка), что позволяет ограничиться частичным парсингом при автоматической обработке текста и довольно свободно оперировать этим видом трансформаций. В языках с фиксированным порядком слов синтаксические трансформации также возможны, но лишь ограниченный их набор приводит к построению грамматически верного предложения.

В [4] в качестве наиболее частотных синтаксических трансформаций в английском языке выделяются изменение позиции обстоятельства в предложении (*Mary reads books in the evening – In the evening, Mary reads books*), вставка наречия (*Mary often reads books in the evening*), пассивизация (*Books are read by Mary in the evening*), преобразование в расщепленное предложение (*It is Mary who reads books in the evening*). В [20] дополнительно выделяются трансформации топикализации (*Mary reads books – Books, Mary reads*), экстрапозиции (*To believe that is difficult – It is difficult to believe that*), преобразование в псевдорасщепленное предложение (*I like apples – Apples are what I like*), преобразование к конструкции с вводным *there* (*Apples are on the table – there are apples on the table*), замещение составляющих на дейктические слова (*Apples are on the table – Apples are there*). При таком подходе каждая из синтаксических структур (тип предложения, порядок слов в предложении) соотнесена с двоичным кодом, и в зависимости от того, какая следующая последовательность бит сообщения M должна быть закодирована, для предложения текста-контейнера C используется трансформация с соответствующим кодом. Трансформированное предложение включается в стеготекст T.

Ненадежность такого подхода с точки зрения стегоанализа заключается в том, что здесь не учитываются требования сохранения естественного для языка статистического распределения лингвистических единиц: некоторые синтаксические конструкции довольно редки в языке (например, предложения с топикальным выделением), и их частое появление в тексте может свидетельствовать о вмешательстве с целью передачи тайного сообщения. Кроме того, дискурсивный анализ может выявить необычное распределение синтаксических типов. Например, если в качестве текста-субстрата C выступает текст инструкции, предложения в нем представляют собой последовательность императивов. Внезапное появление, например, пассивной конструкции приведет к тому, что это предложение будет «выбиваться» из профиля типичной структуры дискурса в данном типе текстов.

Заключение

Рассмотренные в данной статье подходы представляют собой лишь часть современных методов, применяющихся в области лингвистической

стеганографии. В основе этих подходов лежит опора на «внешние» свойства лингвистического знака – синтаксическая структура и частотные характеристики. Их использование направлено на сокрытие факта передачи сообщения от машины. Однако такой автоматически сгенерированный стеготекст наверняка не пройдет проверку, выполненную человеком вручную, так как такой текст не учитывает семантическую составляющую, и, в результате, сгенерированный стеготекст не имеет смысловой структуры. Это приводит к мысли о том, что для большей надежности сокрытия информации необходимо прибегнуть к подходам, принимающим во внимание также смысл при генерации стегостекста. Эти подходы рассмотрим в последующих публикациях.

Литература/References

1. Блумфильд Л. Язык Изд. 2, стереотип. М.: Изд-во УРСС, 2002. 608 с. [Bloomfield L. *Yazyk* (The Language), 2nd edition, Moscow, Izdatel'stvo URSS Publ., 2002, 608 p.]

2. Ляшевская О.Н., Шаров С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. 1112 с. [Lyashevskaya O.N., Sharoff S.A. *Chastotnii slovar sovremennogo russkogo yazyka (na materialakh Natsionalnogo korpusa russkogo yazyka)* (Frequency List of the Modern Russian Language (based on the Russian National Corpus), Moscow, Azbukovnik Publ., 2009, 1112 p.]

3. Agarwal Monika. Text Steganographic Approaches: A Comparison, *International Journal of Network Security & Its Applications*, Vol. 5, No. 1, January 2013, pp. 91–106.

4. Atallah M. J., Raskin V., Crogan M., Hempelmann C., Kerschbaum F., Mohamed D., and Naik S. Natural Language Watermarking: Design, Analysis, and a Proof-of-Concept Implementation, I.S. Moskowitz (ed.), *Information Hiding: Fourth International Workshop*, Lecture Notes in Computer Science 2137, Springer, April 2001, pp. 185–199.

5. Bennett Krista. Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text, CERIAS TR 2004-13, Tech. Report, Center for Education and Research in Information Assurance and Security (CERIAS), Purdue University, May 2004, 30 p.

6. Chapman M.T., Davida G.I. Hiding the Hidden: A Software System for Concealing Ciphertext as Innocuous Text, O. S. Q. Yongfei Han Tatsuaki (ed.) *Information and Communications Security: First International Conference* (Beijing, China, November 11–14, 1997), *Lecture Notes in Computer Science 1334*, Springer, August 1997.

7. Chomsky Noam, Miller George A. Finite State Languages, *Information and Control*, Vol. 1, No. 2, May 1958, pp. 91–112.

8. Greibach Sheila. A New Normal-Form Theorem for Context-Free Phrase Structure Grammars, *Journal of the ACM*, Vol. 12, No. 1, January 1965, pp. 42–52.

9. Huffman D.A. Canonical Forms for Information Lossless Finite-State Logical Machines, *IRE Transactions on Circuit Theory,* Special supplement, Vol. CT-6, May 1959, pp. 41–59.

10. Kaleem M.K. An Overview of Various Forms of Linguistic Steganography and Their Applications in Protecting Data, *Journal of Global Research in Computer Science*, Vol. 3, No. 5, May 2012, pp. 33–38.

11. Kermanidis Katia Lida. Capacity-Rich Knowledge-Poor Linguistic Steganography, Journal of Information Hiding and Multimedia Signal Processing, Vol. 2, No. 3, July 2011, pp. 247–258.

12. Kim Mi-Young. Text Watermarking by Syntactic Analysis, Proceedings of the 12th WSEAS International Conference on COMPUTERS (Heraklion, Greece, July 23-25, 2008), 2008, pp. 904–909.

13. Lubacz Józef, Mazurczyk Wojciech, Szczypiorski Krzysztof. Principles and Overview of Network Steganography, *IEEE Communications Magazine*, Vol. 52, No. 5, May 2014, pp. 225–229.

14. Meral H. M., Sevinc E., Unkar E., Sankur B., Ozsoy A. S., Gungor T. Syntactic Tools for Text Watermarking, Edward J. Delp, Ping Wah Wong (eds.), *Proceedings of the SPIE Electronic Imaging Conference*: *Security, Steganography, and Watermarking of Multimedia Contents,* Vol. 6505, San Jose, CA, January 2007, pp. 65050X-1-12.

15. Meral Hasan Mesut, Sankur Bülent, Özsoy A. Sumru, Güngör Tunga, Sevinç Emre. Natural Language Watermarking via Morphosyntactic Alterations, Computer Speech and Language, Vol. 23, 2009, pp. 107–125.

16. Murphy B., Vogel C. The Syntax of Concealment: Reliable Methods for Plain Text Information Hiding, *Proceedings of the SPIE Conference on Security, Steganography, and Watermarking of Multimedia Contents*, Vol. 6505, San Jose, CA, 2007, pp. 65050Y-1-12.

17. Pfitzman B. Information hiding terminology – results of an informal plenary meeting and additional proposals, R. J. Anderson (ed.), *Proceedings of Information Hiding: First International Workshop, Lecture Notes in Computer Science,* Vol. 1174, Isaac Newton Institute, Cambridge, England, May 1996, Springer-Verlag, Berlin, Germany, ISBN 3-540-61996-8, pp. 347–350.

18. Singh H., Singh P. K., Saroha K. A Survey on Text Based Steganography, *Proceedings of the 3rd National Conference INDIACom-2009* (New Delhi, February 26 – 27, 2009), 2009, pp. 3–9.

19. Topkara M., Topkara U., Atallah M.J. Words are not Enough: Sentence Level Natural Language Watermarking, *Proceedings of the ACM Workshop on Content Protection and Security*, Santa Barbara, CA, 2006, pp. 37–46.

20. Topkara Mercan, Taskiran Cuneyt M., Delp Edward J. Natural Language Watermarking, Edward J.

Delp, Ping Wah Wong (eds.), *Proceedings of SPIE and IS&T Electronic Imaging*: *Security, Steganography, and Watermarking of Multimedia Contents VII*, Vol. 5681, 2005, pp. 441–452.

21. Wai Ei Nyein Chan, Khine May Aye. Modified Linguistic Steganography Approach by Using Syntax Bank and Digital Signature, International Journal of Information and Education Technology, Vol. 1, No. 5, December 2011, pp. 410–415.

22. Wayner Peter. *Disappearing Cryptography: Information Hiding: Steganography & Watermarking,* Third edition, Burlington, MA, Morgan Kauffman Publishers, 2009, XV, 439 p. (Series in Software Engineering and Programming).

23. Wayner Peter. Mimic Functions, *Cryptologia*, Vol. 16, No. 3, July 1992, pp. 192–214.

Бабина Ольга Ивановна, кандидат филологических наук, доцент, доцент кафедры лингвистики и межкультурной коммуникации, Южно-Уральский государственный университет (г. Челябинск), babinaoi@susu.ac.ru

Поступила в редакцию 16 декабря 2014 г.

LINGUISTIC STEGANOGRAPHY: STATE-OF-THE-ART. PART ONE

O.I. Babina, *South Ural State University, Chelyabinsk, Russian Federation, babinaoi@susu.ac.ru*

In the article the notion of linguistic steganography is positioned amongst the system of information hiding methods as a means to generate a (pseudo-) text with the embedded covert message. The notion of linguistic steganography proper, implying mimicking of linguistic features of a "natural" human-made text when composing a stegotext, and text steganography, including format-based transformations of the cover text, are opposed. The levels of steganographic text modeling are distinguished. These levels are accounted for when generating a stegotext characterized by the feature distribution typical of a non-modified text in a natural language. The techniques used to model statistical and syntactic features of the text are considered. Advantages and disadvantages of the techniques presented are analyzed.

Keywords: linguistic steganography, stegotext, information hiding, information security, natural language processing.

Olga I. Babina, Candidate of Philology (PhD), Associate Professor, Associate Professor of the Department of Linguistics and Intercultural Communication, South Ural State University (Chelyabinsk), babinaoi@susu.ac.ru

Received 16 December 2014

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Бабина, О.И. Лингвистическая стеганография: современные подходы. Часть 1 / О.И. Бабина // Вестник ЮУрГУ. Серия «Лингвистика». – 2015. – Т. 12, № 3. – С. 27–33.

FOR CITATION

Babina O.I. Linguistic Steganography: State-of-the-Art. Part One. *Bulletin of the South Ural State University. Ser. Linguistics*. 2015, vol. 12, no. 3, pp. 27–33. (in Russ.)