© Tarasenko Ya., Piven O. Features of calculating the information entropy of the text in case of attacking the linguistic stegosystem by semantic compression // Ukrainian Scientific Journal of Information Security, 2018, vol. 24, issue 2, p.124-129.

СТЕГАНОГРАФІЯ ТА СТЕГОАНАЛІЗ / STEGANOGRAPHY & STEGANALYSIS

DOI: 10.18372/2225-5036.24.12954

ОСОБЛИВОСТІ ОБЧИСЛЕННЯ ІНФОРМАЦІЙНОЇ ЕНТРОПІЇ ТЕКСТУ В УМОВАХ ПРОВЕДЕННЯ АТАКИ СЕМАНТИЧНИМ СТИСНЕННЯМ НА ЛІНГВІСТИЧНУ СТЕГОСИСТЕМУ

Ярослав Тарасенко, Олег Півень

Черкаський державний технологічний університет, Україна

ТАРАСЕНКО Ярослав Володимирович

Рік та місце народження: 1993 рік, м. Черкаси, Україна.
Освіта: Черкаський державний технологічний університет, 2015 рік;
Черкаський державний технологічний університет, 2016 рік.
Посада: аспірант кафедри інформаційної безпеки та комп'ютерної інженерії.
Наукові інтереси: комп'ютерна лінгвістична стеганографія, математична та прикладна лінгвістика, інформаційна безпека, комп'ютерні системи.
Публікації: 14 наукових публікацій, в тому числі наукові статті, матеріали та тези доповідей на конференціях.
E-mail: yaroslav.tarasenko93@gmail.com

ПІВЕНЬ Олег Борисович, к.ф.-м.н.

Рік та місце народження: 1964 рік, м. Черкаси, Україна.
Освіта: Черкаський державний педагогічний інститут, 1986 рік.
Посада: професор кафедри інформаційної безпеки та комп'ютерної інженерії.
Наукові інтереси: взаємодія лазерного випромінювання зі світлочутливими кристалами AgHal, збільшення швидкодії арифметичних пристроїв.
Публікації: автор і співавтор 35 наукових публікацій.
E-mail: pivolegbor@gmail.com

Анотація. *У статті, на основі відомих методів обчислення ентропії тексту проводиться їх удосконалення та описуються особливості обчислення інформаційної ентропії тексту в умовах проведення атаки семантичним стисненням на лінгвістичну стегосистему, реалізовану в однойменному програмному комплексі, формалізується задача визначення ентропії тексту природньої мови в контексті подальшого дискурсного аналізу та видалення семантичної надлишковості. Вводяться додаткові параметри, що сприяють визначенню семантичної ентропії осмисленого та штучно згенерованого тексту для проведення атаки семантичним стиснення на лінгвістичну стегосистему, контейнером для якої виступає текстова інформація природньої (англійської) мова. Обґрунтовуються розбіжності ентропії для різних стилів мови та пояснюється її збільшення зі зміною стилю завдяки потребі додавання до використаного словнику загальної термінології спеціалізованих словників. Крім особливостей розрахунку умовної та безумовної ентропії у випадку використання програмного комплексу проведення атаки на лінгвістичну стегосистему, наведено розрахунок потужності використаного у ньому словнику та прописаних правил граматики, що і є додатковими параметрами, які зумовлюють обчислення ентропії в конкретному випадку, наводиться розрахунок максимальної ентропії (для неосмисленого тексту) та кількості інформації, що несе одне слово чи граматична форма у випадку максимальної та реальної ентропії. Крім того, наводиться обчислення межі семантичного стиснення та формалізовано задачу визначення надлишкової смислової інформації. Таким чином, стає можливим визначення якості проведення атаки стисненням, що проводиться на основі використання відповідного програмного комплексу. Отримані результати можуть бути використані в подальших дослідженнях для удосконалення засобів проведення атаки, що дозволить підвищити її ефективність за рахунок максимального наближення до межі семантичного стиснення.*

***Ключові слова**: лінгвістична стеганографія, протидія методам стеганографії, інформаційна ентропія, семантичне стиснення, межа семантичного стиснення, семантична надлишковість, стегоаналіз, текстова стеганографія, видалення стегоповідомлення.*

Вступ

На сьогоднішній день, потреба автоматизації обробки текстових даних в комп'ютерних системах стегоаналізу для дослідження інформаційних потоків з метою виявлення і перекриття прихованих каналів зв'язку або в рамках редагування матеріалів залишених користувачами на веб-сайтах з вільним доступом та запобігання таким чином несанкціонованого зберігання інформації на вільно доступних ресурсах чи визначення тематики та мети написання тексту, яка пов'язана з приховуванням стегоповідомлення зумовлює необхідність створення ефективних методик та їх програмних реалізацій, що могли б ефективно протидіяти методам стеганографії для запобіганню незаконної передачі чи зберігання даних. Для цього доречно використовувати одну з описаних у [1] атак на стегосистему, а саме атаку проти вбудованого повідомлення, направлену на видалення чи спотворення стегоповідомлення в контексті її застосування у відношенні лінгвістичної стегосистеми, яка основана на використанні тексту природньої мови в якості контейнеру для приховування повідомлення. Саме в рамках науково-практичної задачі проведення автоматизованої атаки на лінгвістичну стегосистему шляхом семантичного стиснення, яка базується на результатах стегоаналізу, що проводиться за допомогою дискурсного аналізу та елементів інтенсіональної логіки, яка сформована в роботі [2], було реалізовано програмний комплекс проведення атаки на лінгвістичну стегосистему. Він реалізує атаку проти стегоповідомлення та описаний у статті [2]. Застосування цієї атаки саме для лінгвістичної стегосистеми основане на семантичному стисненні тексту. Підтвердження працездатності та ефективності програмного комплексу в контексті лінгвістичного стегоаналізу надано в роботі [3]. Одним з основних модулів програмного комплексу є система обчислення ентропії, що входить до складу модуля стиснення тексту, що головним чином відповідає за видалення стегоповідомлення. Проте, на відміну від звичного уявлення про ентропію тексту, описану Шенноном, що зазвичай використовується для визначення надлишковості текстової інформації шляхом дослідження статистики використання символів чи букв відповідного алфавіту, в контексті проведення атаки семантичним стисненням слід враховувати саме семантичну ентропію тексту природньої мови. Таким чином, аналізу підлягає імовірність появи тієї чи іншої лексичної або синтаксичної конструкції, притаманної тексту відповідного стилю.

Аналіз досліджень та постановка завдання

Обчислення ентропії для подібного типу атаки, що наносить шкоду тексту шляхом видалення надлишковості було розроблено в роботі [4], однак в такому разі семантика початкового тексту втрачається, оскільки враховується лише символьна надлишковість, тому особливості обчислення семантичної ентропії тексту є важливою задачею.

Що стосується стегоаналізу тексту, то у роботі [5] запропоновано метод, який базується на використанні інформаційної ентропії в ролі статичної змінної слів у досліджуваному тексті разом з його дисперсією як двокласифікаційні особливості.

В роботах [6-8] досліджуються особливості семантичної ентропії та підходи до її визначення, проте, не можливо ігнорувати твердження, описане в роботі [9], де говориться, що особливості використаного словника та описаних правил граматики впливають на ентропію тексту і в такому випадку ентропія для кожного тексу буде обраховуватись по унікальним правилам. Звідси витікає необхідність формалізації обчислення ентропії тексту в умовах проведення атаки на лінгвістичну стегосистему за допомогою програмного комплексу [2]. Цим зумовлюється актуальність статті. Якщо мова йде про стиснення тексту на смисловому рівні, де ентропія одна з характеристик, що зумовлюють визначення семантичної надлишковості текстової інформації, видалення якої зумовлює видалення і стегоповідомлення, що змусить зловмисника виконати відповідні дії, які можуть скомпрометувати використаний алгоритм чи метод приховування повідомлення та підтвердити наявність прихованого каналу зв'язку, то для стегоаналізатору незвична ентропія буде ознакою наявності слідів модифікації тексту одним з можливих методів стеганографії.

Крім того, в статті [10] надано результат експериментального обчислення інформаційної ентропії природньої мови на прикладі російської та казахської. І, хоча, стаття присвячена дослідженню особливостей обчислення ентропії у випадку використання конкретної програми по відношенню до англійської мови, порівняння отриманих даних по стилям природньої мови може допомогти визначити перспективи застосування програми в майбутньому.

Актуальність роботи зумовлена також тим, що описаний в [2] програмний комплекс також стосується питань дискурсного аналізу, а як відомо, значення повідомлення підпорядковується своїм правилам імовірності, що відповідає ентропійним процесам [11], а звідси слідує, що ентропія також впливає на визначення дискурсу і навпаки, особливості дискурсу тексту визначають його ентропію.

На основі використання словників та прописаних правил граматики, які розраховані на подальший дискурсний аналіз, можна стверджувати, що ентропія тексту, визначена запропонованою програмою буде відрізнятися від типової ентропії тексту цього ж стилю, вирахуваної іншими програмними засобами, а отже також впливатиме на ефективність проведення атаки. Звідси слідує, що опис особливостей вирахування ентропії саме в контексті роботи програмного комплексу може підтвердити його вищу ефективність в порівнянні з аналогічними системами, а також визначити подальші дослідження в цьому напрямку і розвиток програмного продукту в напрямку вдосконалення стегоаналізу та проведення атаки. Таким чином формалізація задачі обчислення

ентропії тексту англійської мови в даному контексті є необхідною.

Метою роботи є наведення особливості обчислення інформаційної ентропії тексту в умовах проведення атаки семантичним стисненням на лінгвістичну стегосистему, реалізовану в однойменному програмному комплексі, формалізація задачі визначення ентропії тексту природньої мови в контексті подальшого дискурсного аналізу та видалення семантичної надлишковості.

Основна частина дослідження

Враховуючи описану в [12] особливість визначення семантичної надлишковості, що можна спостерігати у випадку, коли кілька елементів поверхневої структури представляють один елемент глибинної, можна використати ентропію смислової інформації тексту, вирахувану за допомогою формул, описаних в [9], за відмінністю, що вони адаптовані для задач скорочення тексту. Саме ці обчислення найкраще підходять в якості основи для задачі проведення атаки на лінгвістичну стегосистему, оскільки описаний підхід розрахований під літературні тексти, зокрема вірші, і говориться, що залишкова ентропія може бути використана для застосування відповідних літературних прийомів, а отже можна зробити висновок, що і для приховування інформації.

Таким чином у загальному вигляді інформація в реченні, що потребує видалення визначатиметься на основі формули, описаної в [9, с. 27-28] як різниця безумовної ентропії $H(x)$ і умовної ентропії $H(x/y)$ за формулою (1):

$$I(x/y) = H(x) - H(x/y), \qquad (1)$$

де об'єкт y – стиснене речення, об'єкт x – задане початкове речення. Тобто під умовною ентропією розуміється мінімальна кількість інформації необхідної для побудови стисненого речення у при заданому початковому реченні x. Інформація в початковому реченні відносно стисненого і буде тією надлишковістю, що повинна бути видалена, модифікована чи замінена в рамках проведення атаки на стегосистему.

Визначення умовної та безумовної ентропії з метою виявлення необхідної для видалення інформації залежить від використаного словнику та прописаних правил граматики. Таким чином, відомо, що безумовна ентропія [13] визначається за формулою (2):

$$H_{без.}(\alpha) = \sum_{i=1}^{k} P(\alpha_i) \cdot \log_2 \frac{1}{P(\alpha_i)}, \qquad (2)$$

де $P(\alpha_i)$ – імовірність появи символу джерела алфавіту, k – потужність алфавіту. Безумовна ентропія максимальна і характеризується рівно очікуваними синтаксичними чи лексичними одиницями. Саме такою ентропією буде володіти штучно згенерований неосмислений текст. При обчисленні ентропії осмисленого тексту слід враховувати також потужність словника та прописані у даному середовищі правила граматики, що зумовлюють використання відповідних словоформ в англійській мові та розгалуження семантики, адже як таких словоформ в англійській мові не існує, а граматичні ознаки лексична одиниця отримує з граматичної структури, у якій вона вжита. Тому слід ввести додаткові параметри, що впливають на особливості визначення ентропії у програмному комплексі проведення атаки на лінгвістичну стегосистему. В такому разі, безумовна ентропія $H(x)$ визначатиметься за формулою (3):

$$H_{без.}(x) = \sum_{i=1}^{l} \sum_{j=1}^{d} P(x_{ij}) \cdot \log_2 \frac{1}{P(x_{ij})}, \qquad (3)$$

де l – потужність прописаного граматичного апарату, d – потужність використаного словника, $P(x_{ij})$ – імовірність появи слова джерела словника у відповідності до граматичного апарату. Це зумовлено тим фактом, що від використаного правила залежить значення (словоформа) одного і того ж слова. В програмному комплексі [2] використовується словник Мюллера (24 видання) об'ємом 66 000 слів, отже d = 66 000. Граматичний апарат складається з 20 основних правил, кожне з яких складається з 13-17 підпорядкованих правил, тому l = 290. Це значення приблизне і зумовлено тим, що хоча загальна кількість правил більша, проте саме ці правила є найбільш уживаними, та можуть значно впливати на ентропію тексту. Таким чином, одне слово буде нести D інформації: $D = \log_2 66000 \approx 16$ біт/слово. Словоформа чи лексична одиниця у відповідності до правил граматики буде нести L інформації: $L = \log_2 290 \approx 8$ біт на граматичну одиницю.

За статистичними даними, найбільш вживані слова зі словника складають близько 8 000 слів, якщо брати до уваги розмовний стиль. Це знижує кількість інформації одного слова, а відповідно і ентропію (імовірність передбачення зустрічі того чи іншого слова). Таким чином кількість інформації, що несе одне слово $D_1 = \log_2 8000 \approx 13$ біт/слово. Якщо також враховувати і найбільш вживані правила граматики (наприклад, 10 часових форм замість існуючих 24, використання допоміжного дієслова «will» замість застарілого «shall» та багато інших особливостей), то кількість інформації, що несе граматична одиниця $L_1 = \log_2 110 \approx 7$ біт на граматичну одиницю. Таким чином, врахування частоти вживання слова та його зв'язків з іншими словами і конструкціями, зменшує ентропію. Очевидно, що для різних текстів, які навіть належать до одного стилю ентропія буде різною, тому врахування особливостей кожного тексту, на основі правил граматики та використаного словника є необхідною умовою його стиснення, а отже і видалення потенційного стегоповідомлення, прихованого у ньому.

Що стосується специфічної літератури, написаної, наприклад, в науковому стилі, то використана лексика має свої особливості і не є широко розповсюдженою, а відповідно і виникає необхідність до словника Мюллера додавати словники фахової термінології, використаної у тексті. Це збільшує кількість інформації, що несе одне слово, оскільки до 66 000 описаних слів у словнику додатково приєднуються специфічні терміни. Саме цим і пояснюється описане в [12] різноманіття ентропії у різних стилях і менша ентропія тексту розмовного стилю в порівнянні, наприклад, з науковим. А саме, зазначена адекватність та принциповість ентропії для офіцій-

но-ділового, наукового, публіцистичного та художнього стилів, на основі чого можна зробити висновок, що ентропія офіційно-ділового та наукового стилю досить низька в порівнянні з публіцистичним, ентропія якого в свою чергу нижча, ніж у художнього.

Програмний комплекс враховує особливості кожного стилю, тому виникає необхідність використання частотного словнику англійської мови і, таким чином, визначення вживаності слова у досліджуваному тексті. Крім того, на ентропію впливає імовірність появи однієї лексичної конструкції після іншої, що визначається правилами граматики, тому ці поняття (прописані правила граматики та використане слово) розглядаються лише разом.

При тому, необхідно враховувати межу семантичного стиснення, перевищення якої приведе до втрати семантичної цілісності, в той же час, значне віддалення від цієї межі сприятиме зниженню ефективності подальшої атаки на стегосистему. Таким чином, застосовуючи описану в [13] особливість ефективного кодування до лінгвістичної стегосистеми та семантичного стиснення текстової інформації, можна стверджувати, що гранична межа $L_{гр.}$ буде тим обмеженням, що накладається на видалену семантичну інформацію з метою перешкоджання втрати семантичної структури початкового тексту та розраховуватиметься за формулою (4):

$$L_{гр.} = -\sum_{i=1}^{l}\sum_{j=1}^{d} f_{ij} \cdot \log_2 f_{ij}\,, \qquad (4)$$

де l – потужність прописаного граматичного апарату, d – потужність використаного словника, f_{ij} – частота вживання j-го слова з використаного словника у відповідності з і-м правилом граматики, що і вимагає використання частотного словника англійської мови в комплексі зі словниками фахової та загальновживаної термінології для підвищення точності дослідження та ефективності атаки.

Як відомо, умовна ентропія величини Y при спостереженні величини X [13] визначається за формулою (5):

$$H(Y/X) = \sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i y_j) \cdot \log_2 \frac{1}{P(y_j / x_i)}\,, \qquad (5)$$

де N – число можливих станів системи X, M – число можливих станів системи Y, P – імовірність входження системи X в стан x_i відносно стану y_j системи Y. Що стосується обчислення умовної ентропії з огляду проведення атаки на лінгвістичну стегосистему, то в такому разі слід враховувати імовірність появи лексичних одиниць, як в ансамблі X, так і в ансамблі Y. В такому разі ентропія заданого початкового речення x при спостереженні можливого стисненого речення y буде обчислюватись за формулою (6):

$$H_{ум.}(x/y) = \sum_{i=1}^{N}\sum_{j=1}^{M} P(y_i x_j) \cdot \log_2 \frac{1}{P(x_j / y_i)}\,, \qquad (6)$$

де N – число можливих станів початкового речення x, M – число можливих станів стисненого речення y, P – імовірність входження речення x в стан x_j відносно стану y_i речення y. Таким чином, згадана інформація $I(x/y)$, що потребує видалення повинна відповідати умові (7), що є різницею формул (3) та (6) та менша граничної межі (4):

$$L_{гр.} = -\sum_{i=1}^{l}\sum_{j=1}^{d} f_{ij} \cdot \log_2 f_{ij} > I(x/y) = \sum_{i=1}^{l}\sum_{j=1}^{d} P(x_{ij}) \cdot \log_2 \frac{1}{P(x_{ij})} - \sum_{i=1}^{N}\sum_{j=1}^{M} P(y_i x_j) \cdot \log_2 \frac{1}{P(x_j / y_i)}. \qquad (7)$$

Звідси слідує можливість ствердження, що в програмному комплексі проведення атаки на стегосистему, інформація, яка потребує видалення при семантичному стисненні тексту, а відповідно і знищенні стегоповідомлення в результаті проведення атаки на стегосистему, є різницею безумовної та умовної ентропії, що обраховуються у відповідності з особливостями, зумовленими лінгвістичним стегоаналізом та використаним словником і граматичним апаратом та за умови відповідності зазначеним межам семантичного стиснення.

На основі описаних принципів ентропії функціонує згаданий програмний комплекс проведення атаки на лінгвістичну стегосистему [2], роботу якого протестовано, в тому числі і в плані ентропії та її розрахунку та порівняно з існуючими аналогічними системами [3]. Такий підхід до обрахунку ентропії, введені параметри, та особливості використання словникових баз даних зумовлюють описане в статті [3]підвищення якості роботи багатьох модулів програмного комплексу.

Висновок

Для формалізації задачі визначення ентропії тексту природньої мови в контексті подальшого дискурсного аналізу та видалення семантичної надлишковості було наведено особливості обчислення умовної та безумовної інформаційної ентропії тексту в умовах проведення атаки семантичним стисненням на лінгвістичну стегосистему, реалізовану в однойменному програмному комплексі. Формалізовано розрахунок межі семантичного стиснення, перевищення якої спричинить втрату важливої початкової семантичної інформації тексту. Введено додаткові параметри, що характеризують особливості використаного словника та зв'язок слів за допомогою прописаних у програмному комплексі правил англійської граматики. Максимальна потужність словника (за умови розгляду не фахового тексту без спеціалізованої термінології) складає 66 000. Максимальна потужність зазначених правил граматики, що впливають на ентропію 290. Дійсна потужність словника 8 000, правил граматики 110. Кількість інформації, що несе одне слово від 13 до 16 біт/слово. Кількість інформації, що несе граматична одиниця від 7 до 8 біт на граматичну одиницю.

Описаний підхід до обрахунку ентропії забезпечує якісне покращення стегоаналізу, а саме незалежність результатів атаки на стегосистему від ентропії тексту.

Результати дослідження можуть бути використані для визначення якості проведення атаки стисненням на лінгвістичну стегосистему. В подальших дослідженнях формалізоване визначення ентропії

дасть змогу удосконалити засоби проведення атаки. Крім того дозволить підвищити ефективність атаки за рахунок максимального наближення до межі семантичного стиснення. Є основою для проведення подальших дій та етапів методу семантичного стиснення текстової інформації для протидії комп'ютерній лінгвістичній стеганографії, оскільки ці дії виконуються з оглядом на обраховану максимально можливу інформацію, що може бути видалена на основі особливостей обчислення ентропії.

Література

[1] В. Грибунин, И. Оков, И. Туринцев. «Цифровая стеганография». М.: СОЛОН-ПРЕСС. 2009. 263 с.

[2] Я. Тарасенко. «Програмний комплекс проведення атаки на лінгвістичну стегосистему», *Безпека інформації*. №24(1). 2018. С. 56-61.

[3] Я. Тарасенко. «Експериментальне дослідження роботи програмного комплексу проведення атаки на лінгвістичну стегосистему». *Захист інформації*. Т.20. № 2. 2018. С. 79-88.

[4] В. Мищенко, Ю. Виланский. «Ущербные тексты и многоканальная криптографія». Минск. Энциклопедикс. 2007. 292 с.

[5] Z. Chen, L. Huang, Z. Yu, Xi. Zhao, Xu. Zhao. «Effective Linguistic Steganography Detection». 8th International Conference on Computer and Information Technology Workshops. Sidney, Australia. July 08-11. 2008. PP. 224-229.

[6] C. Bentz, D. Alikaniotis, M. Cysouw, R. Ferrer-i-Cancho. «The Entropy of Words – Learnability and Expressivity across More than 1000 Languages». *Entropy*. 2017. №19(6):275. URL: http://www.mdpi.com/1099-4300/19/6/275/htm.

[7] A. Herbelot, M. Ganesalingam. «Measuring semantic content in distributional vectors». Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria. August 04-09. Vol. 2. 2013. PP. 440-445.

[8] Z. Jiapeng, Y. Yang, L. Tingwen, S. Jinqiao. «Towards Personal Relation Extraction Based on Sentence Pattern Tree». China Conference on Knowledge Graph and Semantic Computing. Beijing, China. September 19-22. Vol. 650. 2016. PP. 92-103.

[9] В. Иванов. «Избранные труды по семиотике и истории культуры. Том 4: Знаковые системы культуры искусства и науки». М.:, Языки славянских культур. 2007. 792 с.

[10] R. Ospanova. «Calculating Information Entropy of Language Texts». *World Applied Sciences Journal*. №22(1). 2013. PP. 41-45.

[11] С. Гусаренко. «О семантических структурах дискурса и семантической энтропии». Известия Волгоградского государственного педагогического университета. № 5. 2007. С. 71-74.

[12] Н. Валгина. «Теория текста». М.: Логос. 2003. 280 с.

[13] Е. Зверева, Е. Лебедько. «Сборник примеров и задач по основам теории информации и кодирования сообщений». СПб, НИУ ИТМО, 2014. 76 с.

УДК 003.26 (045)

Тарасенко Я. В., Пивень О.Б. Особенности вычисления информационной энтропии текста в условиях проведения атаки семантического сжатием на лингвистическую стегосистему

Аннотация. *В статье, на основе известных методов вычисления энтропии текста, производится их усовершенствование и описываются особенности исчисления информационной энтропии текста в условиях проведения атаки семантическим сжатием на лингвистическую стегосистему, реализованную в одноименном программном комплексе, формализуется задача определения энтропии текста естественного языка в контексте дальнейшего дискурсного анализа и удаления семантической избыточности. Вводятся дополнительные параметры, способствующие определению семантической энтропии осмысленного и искусственно сгенерированного текста для проведения атаки семантическим сжатия на лингвистическую стегосистему, контейнером для которой выступает текстовая информация естественного (английского) языка. Обосновывается различие энтропии для разных стилей языка и объясняется ее увеличение с изменением стиля благодаря необходимости добавления к использованному словарю общей терминологии специализированных словарей. Кроме особенностей расчета условной и безусловной энтропии, в случае использования программного комплекса проведения атаки на лингвистическую стегосистему, приведен расчет мощности использованного в нем словаря и прописанных правил грамматики, являющихся дополнительными параметрами, которые обусловливают вычисления энтропии в конкретном случае. Приводится расчет максимальной энтропии (для неосмысленного текста) и количества информации, которую несет одно слово или грамматическая форма в случае максимальной и реальной энтропии. Кроме того, приводится вычисление предела семантического сжатия и формализована задача определения избыточности смысловой информации. Таким образом, становится возможным определение качества атаки сжатием, проводимой на основе использования соответствующего программного комплекса. Полученные результаты могут быть использованы в дальнейших исследованиях для совершенствования средств проведения атаки, что позволит повысить ее эффективность за счет максимального приближения к границе семантического сжатия.*

Ключевые слова: *лингвистическая стеганография, противодействие методам стеганографии, информационная энтропия, семантическое сжатие, предел семантического сжатия, семантическая избыточность, стегоанализ, текстовая стеганография, удаление стегосообщения.*

Tarasenko Ya. Piven O. Features of calculating the information entropy of the text in case of attacking the linguistic stegosystem by semantic compression

Abstract. *The article deals with the improvement of well-known methods for calculating the entropy of the text, and the description of the information entropy of the text calculating peculiarities in case of the semantic compression attack on the linguistic stegosystem, implemented in the cognominal program complex. The problem of determining the natural language text entropy in the context of further discursive analysis and semantic redundancy removal is formalized. Additional parameters that contribute to determining the semantic entropy of meaningful and artificially generated text for a semantic compression attack on the linguistic stegosystem, the container of which is textual information of natural (English) language are entered. The entropy variety for different language*

styles is substantiated and its changing according to the style is explained due to the need of adding specialized terminology dictionaries to the general terminology dictionary. In addition to the calculation features of conditional and unconditional entropy in case of using the software complex for attack the linguistic stegosystem, the dictionary size used in it and the prescribed grammar rules size are given, which are the additional parameters determining the entropy calculation in a particular case. The maximum entropy calculation for meaningless texts and the amount of information of a single word or a grammatical form calculation in case of maximum and real entropy are shown. In addition, the calculation of the semantic compression limit is given and the task of determining the semantic information redundancy is formalized. Thus, it becomes possible to determine the quality of the compression attack, carried out on the basis of the software complex use. The obtained results can be used in further research to improve the means of conducting an attack, which will increase its efficiency by maximally approximating the semantic compression limit.
***Key words**: linguistic steganography, counteraction the steganography methods, information entropy, semantic compression, semantic compression limit, semantic redundancy, steganalysis, textual steganography, removal of the stegomessage.*

Отримано 8 *липня 2018 року*, затверджено редколегією 29 *липня 2018 року*