МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ АВІАЦІЙНИЙ УНІВЕРСИТЕТ

САМОЙЛИК ЄВГЕН ОЛЕКСАНДРОВИЧ

УДК 004.056.5

ЛЕКСИКОГРАФІЧНІ МЕТОДИ ЗАХИСТУ МОВНОЇ ІНФОРМАЦІЇ

Спеціальність 05.13.21 – Системи захисту інформації

Автореферат

дисертації на здобуття наукового ступеня

кандидата технічних наук

Київ – 2020

Дисертацією є рукопис.

Робота виконана на кафедрі телекомунікаційних та радіоелекторонних систем Національного авіаційного університету.

Науковий керівник: доктор технічних наук, доцент
Одарченко Роман Сергійович,
завідувач кафедри телекомунікаційних та радіоелектронних систем
Національного авіаційного університету

Офіційні опоненти: доктор технічних наук, професор
Васіліу Євген Вікторович,
директор навчально-наукового інституту кібербезпеки, комп'ютерних і радіо технологій
Одеської національної академії зв'язку
імені О.С. Попова

доктор технічних наук, професор
Фауре Еміль Віталійович,
проректор з науково-дослідної роботи
та міжнародних зв'язків
Черкаського державного
технологічного університету

Захист дисертації відбудеться «27» листопада 2020 р. о 12^{00} годині на засіданні спеціалізованої вченої ради Д.26.062.17 у Національному авіаційному університеті (м.Київ, пр. Любомира Гузара, 1, корп.11, ауд. 111).

З дисертацією можна ознайомитися у науково-технічній бібліотеці Національного авіаційного університету за адресою: 03058, м. Київ, пр. Любомира Гузара, 1.

Автореферат розісланий «27» жовтня 2020 року.

Учений секретар
спеціалізованої ради Д 26.062.17,
доктор технічних наук, доцент Гнатюк С.О.

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність роботи. Наразі розроблено різноманітні практично стійкі криптографічні системи, що знайшли застосування для вирішення широкого спектру прикладних задач, де необхідно забезпечити надійний захист від порушень конфіденційності інформації, що передається відкритими каналами зв'язку. Проте ці криптосистеми не гарантують формальну, теоретично доведену неможливість їхнього злому. Отже, існує проблема недовіри до надійності цих систем в задачах передавання інформації, що характеризуються високими рівнями секретності. Тому тема цієї дисертаційної роботи спрямована на створення засобів, що дозволяють відносно мало затратним шляхом забезпечити ефективний захист такої інформації.

З урахуванням вище зазначеного у даній роботі розглянуто стійкі криптосистеми захисту текстової інформації з теоретично доведеною ідеальною теоретико-інформаційною стійкістю, що гарантують неможливість однозначного відновлення відкритих текстових повідомлень навіть за умов, коли у розпорядженні криптоаналітика знаходяться зразки зашифрованих повідомлень скільки завгодно великої сумарної довжини, а криптоаналітик має необмежений час та необмежені обчислювальні ресурси для дешифрування перехоплених криптограм.

Проте, сучасним досконало стійким криптосистемам притаманні недоліки, що суттєво звужують області їхнього використання. Необхідно слідкувати, щоб поточна сумарна довжина зашифрованих текстових повідомлень у процесі шифрування не перевищувала довжину ключа шифру. Інакше виникає теоретична можливість розкриття ключа шифру.

Численні дослідження багатьох авторів вказують на відносно невеликі значення відстані єдиності при шифруванні повідомлень, складених із символів алфавіту будь-якої із природних мов. Це призводить до необхідності частої зміни ключової інформації, що є проблемою для багатьох застосувань. Окрім того, необхідно забезпечувати випадковість й однакову ймовірність вибору варіантів реалізації ключа. Так що створення нових методів побудови стійких криптосистем, що забезпечують більш великі значення відстані єдиності або, взагалі, забезпечують можливість шифрування скільки завгодно великих обсягів текстових, зокрема голосових, повідомлень незалежно від значень відстані єдиності являє актуальне завдання.

Аналіз потокових шифрів з рівно ймовірними ключами показує, що існує можливість збільшення відстані єдиності за рахунок синтезу штучної мови відображення прикладної області з алфавітом великої розмірності. У цьому випадку відкривається можливість передавання у режимі секретності більшого обсягу зашифрованої інформації без необхідності зміни ключа шифру. До того ж такий підхід не потребує визначення семантичних співвідношень між лінгвістичними конструкціями синтезованої мови і, отже, створення семантичного тезаурусу прикладної області. Тим не менш, неперевищення відстані єдиності є обов'язковою умовою дотримання режиму досконалої секретності у створених таким чином криптосистемах, що не усуває жорстких вимог до системи розповсюдження ключової інформації.

З іншого боку, можливо використати крипто-семантичний підхід до створення стійкої криптографічної системи. У цьому випадку відпадає необхідність дотримання умови неперевищення відстані єдиності і відкривається можливість вибору довжини ключів шифру незалежно від обсягів інформації, що потребують шифрування. Проте на цьому шляху виникає необхідність дослідження семантичних зв'язків між лінгвістичними конструкціями синтезованої мови відображення прикладної області і створення відповідного семантичного тезаурусу.

Вищезгадана задача, яка вирішувалась в даній дисертаційній роботі, обумовлює її **актуальність**.

Зв'язок роботи з науковими програмами, планами, темами. Тематика дисертаційної роботи та одержані результати безпосередньо пов'язані з «Основними науковими напрямами та найважливішими проблемами фундаментальних досліджень у галузі природничих,

технічних і гуманітарних наук НАН України на 2014-2018 роки», Стратегією національної безпеки України від 26 травня 2015 року № 287/2015, Стратегією кібербезпеки України від 15 березня 2016 року №96/2016 і Рамковою програмою ЄС з досліджень та інновацій «Горизонт 2020».

Висвітлені в дисертації наукові результати отримано, здебільшого, в рамках науково-дослідних робіт, які були виконані в Національному авіаційного університеті (НАУ), а також на ДП «Антонов»: шифр 874-ДБ13, тема «Створення та дослідження нових систем захищеного авіаційного радіозв'язку в рамках Концепції CNS/ATM ICAO» (НДР 0110U000225). Отримані результати використовуються у навчальному процесі НАУ. Використання результатів дисертаційної роботи підтверджено відповідними актами впровадження.

Мета роботи – підвищення ефективності шифрування мовної інформації за рахунок використання особливостей семантичних характеристик мовної інформації.

Задачі дослідження:

1. Проаналізувати ефективність сучасних методів захисту мовної інформації та можливість її підвищення за рахунок використання семантичних характеристик текстової інформації.

2. Розробити модель крипто-семантичного словника за рахунок уведення показників семантичних зв'язків між смисловими конструкціями мови відображення прикладної області і на цій основі здійснити синтез структури тезаурусу смислових образів.

3. Розробити метод побудови лексикографічної криптосистеми, заснованої на використанні механізму укрупнення алфавіту джерела текстових повідомлень.

4. Розробити лексикографічний метод захисту текстової інформації.

5. Розробити схему реалізації, програмне забезпечення крипто-семантичного захисту мовної інформації.

Об'єктом дослідження є процеси захисту мовної інформації.

Предметом дослідження є методи підвищення ефективності шифрування текстової інформації, що спрямовані на зменшення необхідності у періодичній зміні використання ключа шифру.

Методи дослідження. Розробка методів, що забезпечують підвищення ефективності шифрування мовної інформації при обміні інформацією через відкриті канали зв'язку, виконана на основі результатів теорії випадкових процесів, теорії телетрафіка, теорії криптографічних систем та теорії лексикографічних систем. В основу роботи покладена теорія секретного зв'язку К. Шеннона. Широко використано результати теорії лінгвістичних корпусів, що відображені, зокрема, у праці Широкова «Корпусна лінгвістика». Під час натурного моделювання запропонованих криптосистем використано сучасні методи комп'ютерного моделювання.

Наукова новизна одержаних результатів. Автором одержані наступні нові наукові результати:

1. В*перше розроблено модель крипто-семантичного словника*, яка за рахунок уведення в прикладну лексикографічну систему показників семантичних зв'язків між смисловими конструкціями мови відображення області застосування дозволяє визначити тезаурус бази захисту інформації у прикладній системі та семантичну структуру словників прикладної області.

2. *Вперше розроблено метод побудови лексикографічної криптосистеми*, який за рахунок розширення базового алфавіту лінгвістичної формальної системи джерела мовних повідомлень забезпечує збільшення відстані єдиності шифру за ключем та дозволяє збільшити довжину шифрованих повідомлень відносно довжини ключової інформації та відповідно зменшити частоту зміни ключів шифру.

3. В*перше запропоновано лексикографічний метод захисту текстової інформації*, що за рахунок випадкової заміни первинного смислового образу повідомлення на інший правдоподібний елемент, узятий із семантичного тезаурусу бази захисту прикладної області,

дозволяє забезпечити підвищену стійкість захисту при відсутності будь-яких обмежень на обсяг мовної інформації, що підлягає шифруванню, і, тим самим, усуває необхідність у періодичній зміні ключової інформації.

Практичне значення одержаних результатів. Автором були отримані нові результати для впровадження на практиці:

1. Розроблено методику автоматизації розробки тезаурусу бази захисту інформації у прикладній області під час побудови лексикографічної криптосистеми.

2. Запропоновано схему технічної реалізації методу побудови криптосистеми із збільшеною відстанню єдиності за ключем шифру для стійкої системи передавання текстових даних, представлених у вигляді табличних форм.

3. Розроблено програмне забезпечення крипто-семантичного захисту текстових даних, що засноване на використанні прикладного тезаурусу смислових образів, який здатний забезпечити режим підвищеної стійкості у рамках конкретно визначених прикладних систем.

Особистий внесок здобувача. Основні положення й результати дисертаційної роботи отримані автором самостійно. Роботи, виконані разом із співавторами, наведені в переліку публікацій. З робіт, що опубліковані у співавторстві, використовуються результати, отримані особисто здобувачем. У роботах, опублікованих у співавторстві, автору дисертації належить: розробка методу створення крипто-семантичної системи захисту інформації для використання в хмарних технологіях [1]; розробка криптосемантичної системи захисту текстової інформації [3]; розробка методу побудови семантичного словника у складі стійкої криптосистеми захисту текстової інформації [4]; визначення основних характеристик лексикографічних систем, придатних для створення стійких систем захисту текстової інформації [5]; удосконалення алгоритму шифрування даних в стільникових мережах із використанням лексикографічних підходів [6]; визначення характеристик лексикографічних систем для шифрування даних [7]; розробка методу створення семантичного тезаурусу для лексикографічних систем [8]; розробка структури крипто-семантичної системи захисту мовної інформації [9]; визначення характеристик лексикографічних систем з позицій технічного захисту текстової інформації [12].

Апробація результатів дисертації. Основні теоретичні та практичні результати дисертаційної роботи доповідались і обговорювались на таких конференціях та семінарах: Міжнародна науково-практична конференція молодих учених та студентів «Політ. Сучасні проблеми науки» (Київ, НАУ, 2017-2019 рр.); Міжнародний круглий стіл «Про національну і інформаційну безпеку РК» (Казахстан, Алмати, 2016 р.); IEEE International Scientific-Practical Conference «Problems of Infocommunications Science and Technology (PIC S&T)» (Харків, ХНУРЕ, 2018 р.); VIII Міжнародна науково-технічна конференція «Комп'ютерні системи і мережні технології» (Київ, НАУ, 2015 р.); Міжнародна науково-технічна конференція «ITSEC» (Київ, НАУ, 2017 – 2018 рр.); Міжнародна науково-технічна конференція «ABIA-2015» (Київ, НАУ, 2015 р.); Міжнародна науково-практична конференція «Інформаційна безпека та комп'ютерні технології» (Кропивницький, КНТУ, 2016 р.); Науково-технічна конференція «Проблеми розвитку глобальної системи зв'язку, навігації, спостереження та організації повітряного руху CNS/ATM» (Київ, НАУ, 2018 р.); XXI Міжнародна науково-технічна конференція «Сучасні засоби зв'язку» (Мінськ, 2018 р.); Науково-технічна конференція «Проблеми експлуатації та захисту інформаційно-комунікаційних систем» (Київ, НАУ, 2017 – 2019 рр.).

Публікації. За матеріалами дисертаційної роботи опубліковано 13 наукових праць, у тому числі 1 розділ колективної монографії, 4 статті у фахових виданнях, які входять в перелік наукових видань, затверджений МОН України, 1 працю, яка включена до науково-метричної бази Scopus, матеріали доповідей на науково-технічних конференціях – 6.

Структура та обсяг дисертації. Дисертаційна робота складається зі вступу, чотирьох розділів, висновків по кожному розділу та загальних висновків по роботі в цілому, списку використаних джерел (122 найменування). Повний обсяг дисертації - 157 сторінок, у тому числі 126 сторінок основного тексту, 17 рисунків, 2 таблиці.

ОСНОВНИЙ ЗМІСТ ДИСЕРТАЦІЇ

У **вступі** визначено проблему, що підлягає вирішенню, та обґрунтовано актуальність теми дисертації, сформульовано мету дослідження, визначено коло задач, що вирішуються, вказано на наукову новизну (розкрито чи сформульовано наукову новизну) та практичне значення отриманих результатів.

У першому розділі було проаналізовано ефективність сучасних методів захисту текстової інформації, зокрема, за критеріями стійкості, довжини ключа, складності реалізації, необхідності передачі ключа шифрування. Результати даного аналізу зібрані в табл. 1.

Таблиця 1

Результати аналізу методів захисту текстової інформації

Криптографічний алгоритм	**Порівняльні параметри ефективності шифрування**			
	Криптографічна стійкість	**Довжина ключа, біт**	**Складність реалізації**	**Передача ключа шифру**
RSA	Стійкий	2048	Складний	-
AES	Стійкий	128,192, 256	Складний	-
Шифр Вернама (одноразових блокнотів)	Абсолютно стійкий	Дорівнює довжині даного шифру	Відносно простий	Потребує передачі ключів шифру

Проведений аналіз показав, що підвищену стійкість має шифр Вернама, проте йому також притаманні деякі недоліки, зокрема він потребує періодичної передачі ключів шифрування, тому його дуже складно застосовувати на практиці. Тому необхідно проводити пошук нових шляхів забезпечення підвищеної стійкості із одночасним зменшенням частоти зміни ключів шифрування.

Показано, що досягнення поставленої мети слід шукати на шляху розробки нових методів побудови досконало стійких криптосистем, які здатні збільшити відстань єдиності за ключем або, в ідеалі, забезпечити досконало стійкий захист при відсутності будь-яких обмежень на обсяг текстової інформації, що підлягає шифруванню. Якщо збільшення відстані єдиності може бути досягнуто шляхом укрупнення алфавіту джерела відкритих текстових повідомлень, то уникнення виконання шеннонівської умови досконалої стійкості захисту є можливим лише за рахунок використання особливостей семантичних характеристик текстової інформації на основі поєднання методів симетричної криптографії і семантичної обробки змісту текстових повідомлень, зокрема за допомогою певним чином побудованих лексикографічних систем, реалізованих у вигляді проблемно-орієнтованих лінгвістичних корпусів. У цьому випадку має здійснюватися шифрування не форми відображення структури співвідношень між характеристиками об’єктів, що відображають текстові повідомлення, а безпосередньо сутність (смисл) цієї структури. Для адекватного відображення методів застосування лексикографічних систем в задачах ТЗІ надано під відповідним кутом зору визначення загальновідомих понять, таких як середовище, суб’єкт, інтелект, семантичний тезаурус, інформація, форма представлення та смисловий зміст інформації, а також надана модель розуміння суб’єктом смислового змісту інформації.

Окремо визначений смисловий потік розглядається як дискретна невипадкова послідовність смислових образів *SO*, що відображають хід думок людини. Вважається, що *SO* є дискретним елементом смислового потоку, що сприймається суб’єктом як віртуальна логічно несуперечлива смислова конструкція, що має ознаки смислової завершеності. Оскільки окремо визначений смисловий потік (*SP*) являє детерміновану послідовність *SO*, то цьому *SP* також притаманна ознака осмисленості.

В якості основної формальної характеристики як *SO*, так і *SP* прийнято **рівень абстрагування** їхнього представлення, а потік смислових образів (*SO*) представляється як

$$SP^{(i)} = SO_1^{(i-1)}, SO_2^{(i-1)}, \ldots, SO_k^{(i-1)}, \ldots, SO_N^{(i-1)}, \quad (1)$$

де $SP^{(i)}$ – смисловий потік i-го рівня абстрагування; i – показник рівня абстрагування; $SO_k^{(i-1)}$ – $к$-й елемент $SP^{(i)}$, що має більш деталізований рівень абстрагування відображення смислової одиниці; $к$ – порядковий номер SO у послідовності смислових образів, що відображають SP; N – довжина послідовності SO, що складають SP.

Визначено **тезаурус** TZ прикладної області як структурований у вигляді прошаркової (рос. – слоистой) коренево-подібної ієрархії набір смислових образів SO, що упорядковані за рівнями абстрагування цих смислових образів у напрямі усе більш конкретного (більш деталізованого) їхнього сприйняття суб'єктами. Тезаурус – це семантичний словник, структура якого однозначно відображає структуру семантичних зв'язків між смисловими конструкціями мови відображення області застосування. У досконало стійких криптосистемах структура семантичних зв'язків між елементами тезауруса має бути відображена на формальному рівні.

Абстрактна модель розуміння мови показана на рис. 1, де прийнято наступні позначення:

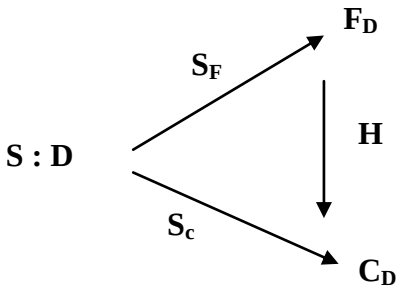


Рис. 1. Абстрактна модель розуміння текстової інформації

S – суб'єкт розумової діяльності (людина, комп'ютерна програма);

D – зразок текстового повідомлення, що осмислюється суб'єктом;

F_D – результат неосмисленого сприйняття відображення форми досліджуваного зразка текстового повідомлення;

C_D – результат сприйняття смислу досліджуваного зразка текстового повідомлення;

S_F – оператор відображення форми досліджуваного зразка текстового повідомлення на основі його безпосереднього сприйняття суб'єктом;

S_C – оператор безпосереднього відображення смислу досліджуваного зразка текстового повідомлення без аналізу його форми;

H – оператор відображення смислу досліджуваного зразка текстового повідомлення на основі аналізу його форми.

У процесі вирішення прикладних завдань суб'єкт формує упорядковані щодо смислу послідовності мовних одиниць шляхом осмисленого вибору цих одиниць із доступного йому тезаурусу. Так що, думка будь-якого суб'єкту являє собою дискретний часовий ряд смислових образів, що вибираються ним із доступного тезаурусу.

У другому розділі розроблено **модель семантичних тезаурусів**, необхідних для використання у складі криптосистеми захисту текстової інформації.

Тезаурус будь-якої мови взагалі TZ_M або будь-якого суб'єкту окремо TZ_S має ієрархічну прошаркову структуру і за ступенем абстрагування відображення смислових образів розподіляються на i рівнів, де $i = 1, 2, \ldots, I$ – кількість рівнів абстрагування відображення смислових образів, якими оперує інтелект у рамках заданої прикладної області.

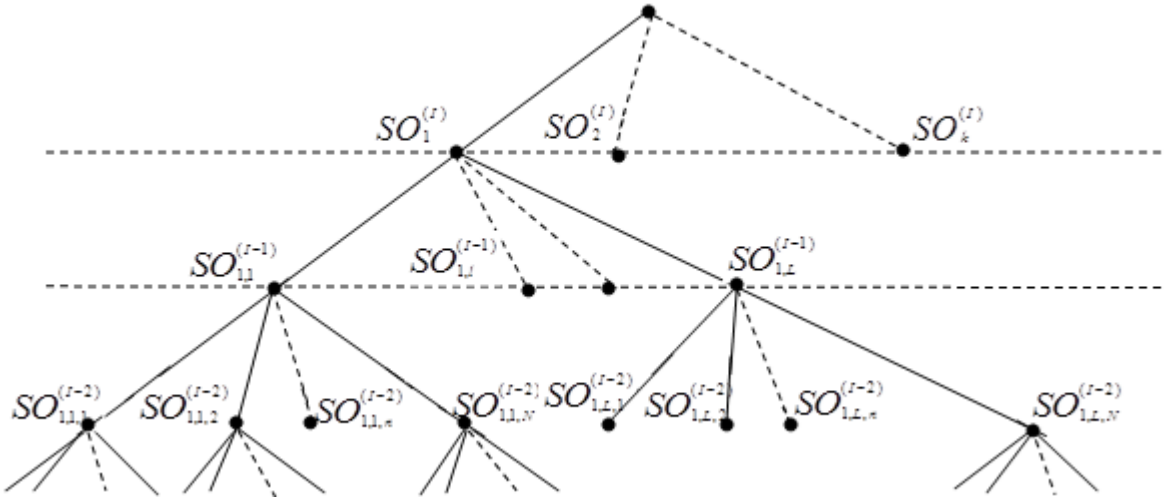


Рис. 2. Прошаркова коренево-подібна структура простору смислових образів, що складають тезаурус

За таким визначенням будь-який тезаурус *TZ* (зокрема, TZ_M або TZ_S) складається із сукупності підтезаурусів семантичних одиниць усіх доступних для розуміння рівнів абстрагування $TZ^{(i)}$, де $i=1,2, \dots ,I$. При цьому підтезауруси у складі *TZ* розташовані прошарками у вигляді гілко-подібної кореневої системи, що розростаються зверху униз. Структуру тезаурусу *TZ* можливо представити у вигляді рекурентної коренево-подібної схеми як

$$TZ \in \{TZ^{(I)}_{k}\},\ \text{де}\ TZ^{(I)}_{k} \in \{TZ^{(I-1)}_{k,l}\},\ \text{де}\ TZ^{(I-1)}_{k,l} \in \{TZ^{(I-2)}_{k,l,n}\}, \dots \quad (2)$$
$$K \qquad\qquad L \qquad\qquad N$$
$$\dots\ \dots\ \dots,\ \text{де}\ TZ^{(2)}_{k,l,n,\dots,p} \in \{SO^{(1)}_{k,l,n,\dots,j}\}.$$
$$J$$

У виразі (2) прийняті наступні позначення (див. рис.1): $TZ^{(I)}_k$ – тезаурус *I*-го рівня абстрагування представлення *SO*; $I \in \{1, 2, \dots\ i, \dots, \dots\ , I_{max}\}$ – кількість рівнів абстрагування відображення *SO*, що доступна інтелектові; I_{max} – теоретично будь-яке велике, але скінчене ціле; $k \in \{1, 2, \dots\ , K\}$ – порядковий номер елемента $TZ^{(I)}_k$ у множині тезаурусів, що у сукупності визначають простір смислових образів *I*-го рівня абстрагування, тобто $\{SO^{(I)}\}$; *K* – кількість тезаурусів *I*-го рівня абстрагування відображення *SO*, що входять до складу *TZ*; $TZ^{(I-1)}_{k,l}$ – тезаурус (*I-1*)-го рівня абстрагування представлення *SO*, що конкретизує смислові образи $TZ^{(I)}_k$, де $l \in \{1,2, \dots , L\}$ – порядковий номер тезаурусу (*I-1*)-го рівня абстрагування у множині тезаурусів, які у сукупності визначають простір смислових образів (*I-1*)-го рівня абстрагування у рамках тезаурусу $TZ^{(I)}_k$; *L* – кількість тезаурусів (*I-1*)-го рівня абстрагування відображення *SO*, що входять до складу $TZ^{(I)}_k$; $TZ^{(I-2)}_{k,l,n}$ – тезаурус (*I*-2)-го рівня абстрагування представлення *SO*, що конкретизує смислові образи $TZ^{(I-1)}_{k,l}$, де $n \in \{1,2, \dots , N\}$ – порядковий номер тезаурусу (*I-2*)-го рівня абстрагування у тезаурусі $TZ^{(I-1)}_{k,l}$; *N* – кількість тезаурусів (*I-2*)-го рівня абстрагування відображення *SO*, що входять до складу $TZ^{(I-1)}_{k,l}$; $TZ^{(2)}_{k,l,n,\dots,p}$ – тезаурус другого рівня абстрагування представлення *SO*, що конкретизує смислові образи $TZ^{(3)}_{k,l,n,\dots,s}$, де $p \in \{1,2, \dots , P\}$ – порядковий номер тезаурусу другого рівня абстрагування у складі тезаурусу третього рівня $TZ^{(3)}_{k,l,n,\dots,s}$, де $s \in \{1,2, \dots , S\}$ – порядковий номер тезаурусу третього рівня абстрагування; *P* – кількість тезаурусів другого рівня абстрагування, що входять до складу $TZ^{(3)}_{k,l,n,\dots,s}$; *S* – кількість тезаурусів третього рівня абстрагування, що входять до складу відповідного тезаурусу четвертого рівня абстрактності і т.д. уздовж ланцюга тезаурусів із зростанням значення індексу *i*.

Якщо розглядати структуру тезаурусів мови відображення будь-якої прикладної області відповідно до розтину за рівнями абстрагування відображення смислових образів, то у загальному випадку доцільно задати наступну ієрархію її семантичних одиниць:

прикладна область (напрямок знань)/ тема/ сценарій/ ситуація/

фраза/ слово/ символ алфавіту/ код символу алфавіту. (3)

Зокрема, у багатьох сферах прикладних застосувань, які у подальшому назвемо областями активності, доцільно призначити наступну ієрархію смислових одиниць, що відображають логічно завершені думки певного рівня абстрагування, де убування ступеня абстрагування прийнято зліва направо:

сценарій/ ситуація / фраза / слово. (4)

У цьому випадку, у виразі (3) $I = 4$, тобто будемо мати чотирьох ступеневу ієрархічну структуру у вигляді розгалуженого кореня тезаурусів з різним рівнем абстрагування представлення смислових образів.

$SO^{(1)}_{k,\,l,\,n,\,\ldots\,,\,j}$ – семантичний словник, що відображає тезаурус $TZ^{(2)}_{k\,,l,\,n,\,\ldots\,,\,p}$; J – кількість слів у тезаурусі $TZ^{(2)}_{k\,,l,\,n,\,\ldots\,,\,j}$.

Отже, структура тезаурусу TZ представляється у вигляді розгалуженого кореня підтезаурусів $TZ^{(i)}$, де $i \in \{1, 2, \ldots, \ldots, I_{max}\}$.

Необхідно надати формальні визначення таким смисловим відношенням між семантичними одиницями як **смислова ідентичність**, **смислова відмінність**, сумнівна смислова ідентичність (або **правдоподібність**), сумнівна смислова відмінність (або **фальшиво-подібність**). Якщо порівнювальні смислові образи сприймаються суб'єктом як ідентичні щодо смислу, але у нього існують небезпідставні сумніви щодо істинності такого сприймання, то у цьому випадку доцільно визначити смислове відношення як правдоподібне. Тобто, якщо в якості позначки відношення правдоподібності обрати символ $\wedge$ та врахувати, що $SO^{(i)}_k \equiv TZ^{(i)}_k$ для будь-яких значень i та k, то будь-які дві семантичні одиниці будь-якого, але однакового рівня абстрагування $TZ_a{}^{(i)}_k$ та $TZ_b{}^{(i)}_k$ із множини $\{TZ^{(I)}_k\} \rightarrow TZ$ знаходяться у відношенні смислової правдоподібності

$$TZ_a{}^{(i)}_k \wedge TZ_b{}^{(i)}_k, \tag{5}$$

якщо ймовірність того, що $TZ_a{}^{(i)}_k \equiv TZ_b{}^{(i)}_k$ є меншою, ніж 1.

Підкреслимо, що відношення правдоподібності визначаються між семантичними одиницями будь-якого, але однакового рівня абстрагування.

Відносно семантичних одиниць I-го (найвищого у даній мові даної області активності) рівня абстрагування буде справедливим наступне твердження:

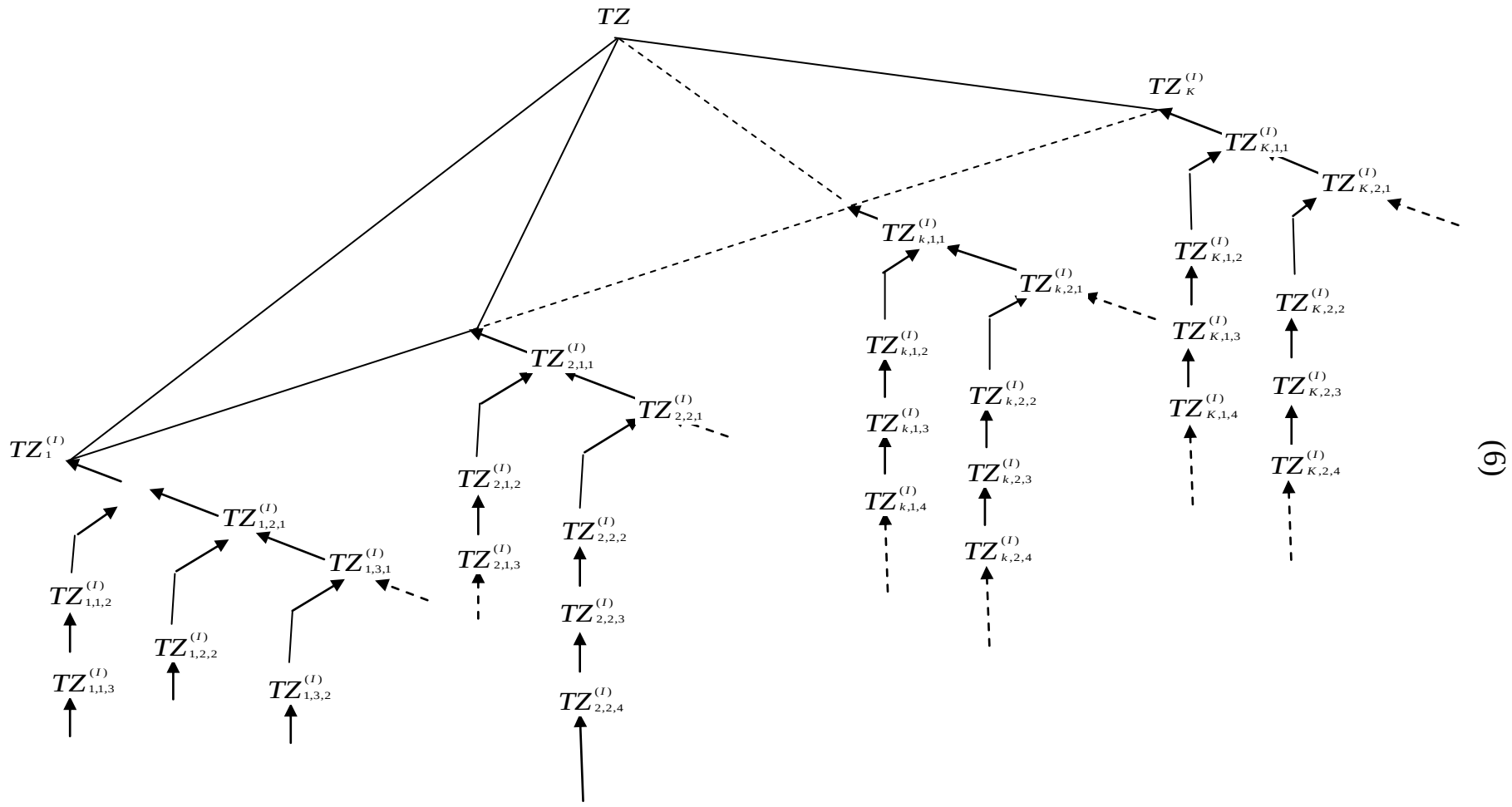
$$SO^{(I)}_1 \wedge SO^{(I)}_2 \wedge \ \ldots \wedge SO^{(I)}_k \wedge \ldots \wedge SO^{(I)}_K, \text{ якщо } \{TZ^{(I)}_k\} \rightarrow TZ\ , \tag{6}$$

K

за умови $SO^{(I)}_k \equiv TZ^{(I)}_k$, для будь-яких значень I та k. (7)

У виразі (6) позначка $\rightarrow$ означає приналежність будь-якого тезаурусу $TZ^{(I)}_k$ із множини тезаурусів I-го рівня абстрагування до тезаурусу обраної мови TZ, а у виразі (7) позначка $\equiv$ означає відношення смислової ідентичності. Аналогічні вирази отримано й відносно інших рівнів абстрагування.

Отже, будь-яка семантична одиниця $TZ^{(I)}_k$ із множини семантичних одиниць I-го рівня абстрагування, що входять до складу тезаурусу мови обраної області активності TZ, під час криптоаналізу сприймається суб'єктом як можливий кандидат на смислову ідентичність із вихідною семантичною одиницею відкритого тексту $TZ^{(I)}_l$. Проте суб'єкт, якщо він має інформацію щодо прийнятої структури тезаурусу, яким він керується, розуміє, що $TZ^{(I)}_k \triangle TZ^{(I)}_l$ для будь-яких значень $k \neq l$, де $\triangle$ – позначка смислової відмінності. За цих умов суб'єкт усвідомлює, що вищезазначені семантичні одиниці знаходяться між собою у відношенні смислової правдоподібності.



(9)

Рис. 3. Тезаурус бази захисту текстової інформації TZ_{BZ}

Для того, щоб криптоаналітик ні за яких умов не мав можливостей скласти будь-яке уявлення щодо істинності смислу перехоплених мовних повідомлень, необхідно і достатньо замінити смисл *SO*, що входять до складу вихідних мовних повідомлень, на правдоподібні їхні відображення, що беруться із тезаурусу бази захисту у прикладній системі, що відображається тезаурусом обраної прикладної області.

Для здійснення такої заміни в автоматичному режимі необхідно мати формальні позначення місць розташування відображень *SO* у структурі TZ_{BZ}, тобто необхідно локалізувати мовні одиниці у структурі тезауруса TZ_{BZ}.

Необхідно виконати аналіз структури TZ_{BZ} та формалізувати цю структуру таким чином, щоб параметри локалізації були представлені у явному вигляді, наприклад у наступному вигляді.

Параметри локалізації цієї структури представлені у явному вигляді як

$$SO^{(i)}{}_k \in \left[\begin{array}{l} SO^{(i-1)}{}_{k,1,1};\ SO^{(i-1)}{}_{k,1,2};\ .....;\ SO^{(i-1)}{}_{k,1,n};\ .....;\ SO^{(i-1)}{}_{k,1,N(l=1)} \\ SO^{(i-1)}{}_{k,2,1};\ SO^{(i-1)}{}_{k,2,2};\ .....;\ SO^{(i-1)}{}_{k,2,n};\ .....;\ SO^{(i-1)}{}_{k,2,N(l=2)} \\ \text{. ……} \\ \text{……...} \\ \text{. ……. ……} \\ SO^{(i-1)}{}_{k,l,1};\ SO^{(i-1)}{}_{k,l,2};\ .....;\ SO^{(i-1)}{}_{k,l,n};\ .....;\ SO^{(i-1)}{}_{k,l,N(l)} \\ \text{……..} \\ \text{……...} \\ SO^{(i-1)}{}_{k,L,1};\ SO^{(i-1)}{}_{k,L,2};\ .....;\ SO^{(i-1)}{}_{k,L,n};\ .....;\ SO^{(i-1)}{}_{k,L,N(L)} \end{array}\right. \qquad (8)$$

Поточні значення нижніх індексів у позначеннях семантичних одиниць у виразі (8) однозначно визначають місця розташування цих семантичних одиниць у структурі бази захисту. Розроблений метод побудови тезаурусу смислових образів дозволяє визначити тезаурус бази захисту інформації у прикладній інформаційній системі, автоматизувати процес його створення під час побудови досконало стійкої криптосистеми.

У третьому розділі поставлено та вирішено завдання з розробки методу побудови лексикографічної криптосистеми, заснованої на використанні механізму укрупнення алфавіту джерела відкритих текстових повідомлень, що забезпечує суттєве збільшення відстані єдиності шифрів за ключем для розширення області застосування такої криптосистеми, для чого необхідно виконати наступні етапи.

Етап 1. Задання табличної форми подання інформації. Припустимо, що критично важлива текстова інформація поміщена у табличну форму розміром *s*x*L2*, де *s* – кількість стовпців таблиці, а *L2* – кількість рядків цієї ж таблиці. Вміст кожної клітини таблиці будемо розглядати як окреме повідомлення (слово або словосполучення) *M* із множини ***M***. Отже, дані таблиці – це текстова інформація, яка підлягає захисту, і при передаванні через канал зв'язку розглядається як потік текстових повідомлень $M_{j,v}$, де *j= 1, 2, ... , L* – індекс, що означає порядковий номер повідомлення у потоці, ν = *1, 2, ... ,N* – індекс, що означає смисловий варіант повідомлення із наперед визначеного словника повідомлень ***S***.

Таблиця 2

Задана таблична форма

	Найменування стовпця №*1*	Найменування стовпця №*2*	---	---	Найменування стовпця № *S*
Найменування рядка №*1*	***Повідомлення*$_{11}$**	***Повідомлення*$_{12}$**			***Повідомлення*$_{1S}$**
Найменування рядка №*2*	***Повідомлення*$_{21}$**	***Повідомлення*$_{22}$**			***Повідомлення*$_{2S}$**

Найменування рядка №*L*	***Повідомлення*$_{L1}$**	***Повідомлення*$_{L2}$**			***Повідомлення*$_{LS}$**

Етап 2. Створення семантичного словника. Уведемо поняття *семантичний словник таблиці* як множину ***M*** усіх можливих семантичних варіантів повідомлень $M_{j,i}$, що можуть знайти своє відображення у таблиці заданої форми. Розмір словника ***S*** це є розмір множини ***M***, що дорівнює *N*. Семантичний словник даної табличної форми складається із *s* семантичних підсловників відповідно до кількості стовпців у таблиці.

Етап 3. Синтез мови відображення текстової інформації, що поміщається у задану табличну форму. Якщо кожен рядок таблиці є літерою певним чином визначеного алфавіту даної табличної форми, то кількість літер у такому алфавіті *B* визначиться за формулою:

$$B = \prod_{i=1}^{s} S_i \text{ , де } s \text{ – кількість підсловників табличної форми,} \quad (9)$$

S_i – кількість елементарних повідомлень *M* в і-ому підсловнику.

Етап 4. Оцінка кількості можливих смислових варіантів повідомлень. Якщо припустити, що кожен рядок у таблиці буде зустрічатися тільки один раз, то кількість можливих варіантів повідомлень *N* довжиною *n* розраховується за формулою:

$$N = (S_1 S_2 \ldots S_s)\,(S_1 S_2 \ldots S_s - 1)\,\ldots\,(S_1 S_2 \ldots S_s - (s-1)), \quad (10)$$

де S_i – розмір і-го підсловника, i = 1, 2, 3, …, *s*.

Етап 5. Визначення відстані єдиності. Визначимо відстань єдиності для розроблюваної ключової системи за формулою:

$U = H(K)/D$, де $H(K)$ – ентропія ключової системи, *D* – надлишковість мови. (11)

Етап 6. Визначення надлишковості *D* штучної мови відображення інформації. За вищезазначених умов надлишковість *D* штучної мови відображення інформації, що поміщена у задану табличну форму, є надзвичайно малою, що, у свою чергу, згідно (9) визначає суттєво великі значення відстані єдиності *U*. Зокрема, у наведеному в роботі прикладі $D = 0{,}000042267419412$ [біт/літера], $U = 1480460{,}75583801$ [літер], а довжина ключового слова $k \approx 12$ [літер]. Як бачимо, обсяги текстової інформації, що потребують досконало стійкого захисту, у даному випадку можуть суттєво перевищувати довжину ключового слова.

В результаті проведених досліджень отримаємо вираз, що визначає залежність значень відстані єдиності *U* від довжини повідомлення *n* (див. рис. 4):

$$U(n) = \frac{\log_2 \prod_{s=1}^{n}\left[B-(s-1)\right]}{\log_2(B) - \frac{\log_2 \prod_{s=1}^{n}\left[B-(s-1)\right]}{n}}, \text{ де } B \text{ – кількість літер у алфавіті таблиці.} \quad (12)$$

Також отримаємо вираз, що визначає залежність ентропії ключа шифру *H(K)* від довжини повідомлення *n* (див. рис. 5):

$$H(K) = \log_2 \prod_{s=1}^{n}\left[B-(s-1)\right]. \quad (13)$$

З графіку на рис. 4 видно, що навіть при максимальній за цих умов довжині повідомлення $n=162$, відстань єдиності дорівнює $U \approx 680$ символів. Отже, шляхом укрупнення алфавіту мови відображення інформації у таблиці заданої форми можемо після відповідних перетворень мати шифротексти, які значно коротші за відстань єдиності.

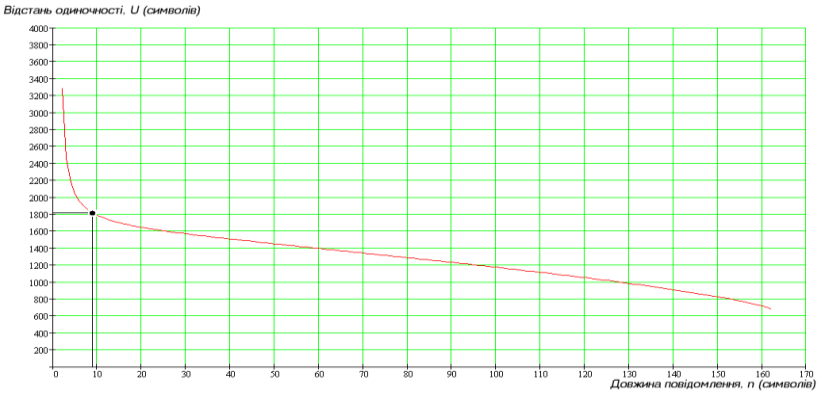


Рис. 4. Графік залежності відстані єдиності *U* від довжини повідомлення *n*

Визначимо виграш *Z* у довжині ключа (за інших рівних умов) у разі застосування у криптосистемі укрупненого алфавіту у порівнянні із методом одноразових блокнотів:

$$Z = \frac{H_1(K)}{H(K)}, \qquad Z = \frac{\log_2(B_1^{n_1})}{\log_2 \prod_{s=1}^{n}[B-(s-1)]} \tag{14}$$

У наведеному в роботі прикладі значення показника виграшу у довжині ключа *Z = 46,2*. Тобто, при застосуванні досконало стійкої криптосистеми з укрупненим алфавітом довжина ключа шифру може бути у *46,2* рази менша за довжину ключа шифру при застосуванні методу одноразових блокнотів.

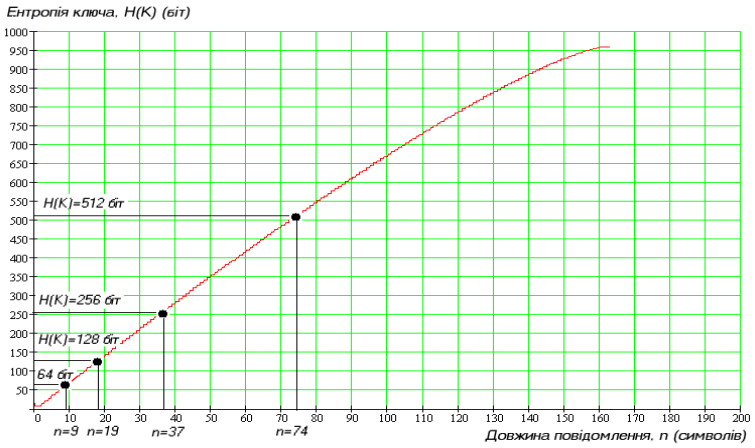


Рис. 5. Графік залежності ентропії ключа шифру *H(K)* від довжини повідомлення *n*

Розглянуто варіант побудови криптосистеми із збільшеною відстанню єдиності (КЗВЄ), що показаний на рис. 6. Принцип роботи КЗВЄ базується на синхронізації ГПВП, що розташовані на передавальній та приймальній сторонах каналу секретного обміну інформацією, за допомогою відомого ключа шифру. Схема містить усі основні елементи симетричної криптосистеми, яка за певних умов здатна функціонувати у режимі досконалої стійкості. Однак, окрім цього, використано додатковий елемент – словник предметної області, у рамках якої планується застосовувати КЗВЄ. Шифратор реалізує механізм укрупнення алфавіту мови відображення інформації. Зокрема, кожен окремий рядок таблиці представляється як окрема літера укрупненого алфавіту. За допомогою ключової інформації ГПВП ставиться у певний початковий стан. Генерується одне псевдовипадкове число на кожен рядок таблиці. Шифратор здійснює заміну істинного укрупненого елементу таблиці, у якості якого розглядається поточний рядок таблиці, на маскуючий рядок лінгвістичних елементів, узятих із словника.

При розмаху ГПВП, що дорівнює розмірності найбільшого підсловника, **рівняння шифрування**, яке являє собою формулу для визначення порядкового номеру розташування маскуючого слова у *к*-ому підсловнику тезаурусу, має наступний вигляд:

$$N_{\text{заш}} = \text{mod}_{p_k}(N_{\text{вих}} + N_{\text{вип}}), \quad (15)$$

де $N_{вих}$ – порядковий номер розташування у *к*-ому підсловнику *і*-го істинного мовного елемента, узятого із вихідної, що підлягає шифруванню, відкритої послідовності лінгвістичних одиниць, де *і* – порядковий номер розташування цього елементу у відкритій послідовності; $N_{зав.}$ – порядковий номер розташування у *к*-ому підсловнику маскуючого елементу, що поміщається в *і*-ту позицію (замість істинного елементу) вивідної зашифрованої послідовності лінгвістичних одиниць; $N_{вип}$ – псевдовипадкове ціле число, що згенероване ГПВП на *і*-ому кроці генерування для шифрування *і*-го істинного елементу із порядковим номером $N_{вих}$; $p_к$ – розмірність *к*-го підсловника; *к* – порядковий номер підсловника у словнику мови прикладної області.

Рівняння розшифрування інформації у даному випадку являє собою формулу для визначення порядкового номеру розташування в обраному підсловнику *і*-ої істинної лінгвістичної одиниці, що поміщається у вивідну розшифровану послідовність замість маскуючої лінгвістичної одиниці, узятої із вихідної зашифрованої послідовності:

$$N_{вих} = \text{mod}_{p_k}(N_{\text{заш}} - N_{\text{вип}}), \quad (16)$$

де $N_{вих}$ – порядковий номер лінгвістичної одиниці (згідно нумерації у *к*-ому підсловнику), що є ідентичним істинній одиниці, відображеній у вихідній відкритій послідовності лінгвістичних одиниць; $N_{заш}$ – порядковий номер зашифрованої лінгвістичної одиниці (згідно нумерації у *к*-ому підсловнику), узятої із зашифрованої послідовності, що надана на обробку, на поточному кроці процедури розшифрування; $N_{вип}$ – псевдовипадкове ціле число, згенероване ГПВП, для розшифрування зашифрованої одиниці із порядковим номером $N_{заш}$.

Вираз (14) є справедливим, якщо розмах ГПВП дорівнює розмірності підсловника, котрий має найбільшу кількість мовних елементів серед множини усіх підсловників використаного словника предметної області.

Слід зазначити, що для забезпечення випадкового характеру підмін кожна наступна таблиця шифрується послідовністю псевдовипадкових чисел без скидання ГПВП у початковий стан. Після кожного сеансу шифрування фіксується кількість псевдовипадкових чисел, що були використанні під час шифрування. І кожна наступна таблиця шифрується не із початкового стану ГПВП, а з тої точки послідовності, де вона була зафіксована на попередньому акті шифрування. Тим самим, нейтралізуються атаки типу імовірнісний аналіз частоти повторень символів вихідного тексту.

Визначено стійкість криптосистеми із збільшеною відстанню єдиності за різних ресурсних можливостей порушника. За умов, коли хоча б один зразок вихідного (відкритого) текстового повідомлення та зразок кореспондованої з ним шифрограми є відомими, стійкість синтезованої криптосистеми співпадає із стійкістю алгоритму шифроперетворень, що використовується. За умов відсутності кореспондованих зразків вихідних та зашифрованих текстових повідомлень та недотримання обмежень щодо відстані єдиності стійкість криптосистеми є статистичною величиною і у *v* разів вища за стійкість методу шифроперетворень, що використовується, де *v* – об'єм статистичної вибірки інформації секретного обміну. З ростом обсягу вибірки *v* стійкість криптосистеми зростає.

Отримано математичні вирази та надано графік залежності показника стійкості від розрядності ключа шифру. З ростом довжини ключа шифру стійкість системи захисту зростає. При збільшенні основи алфавіту мови ключової інформації значно збільшується показник стійкості криптосистеми при сталому значенні розрядності ключа шифру.

Визначено формальні умови коректного функціонування побудованої криптосистеми та розрядність ключа шифру у кореляції із довжиною повідомлень, що закриваються.

Критерій коректного функціонування побудованої КЗВЄ – дотримання відстані єдиності (умова того, що обсяг зашифрованого одним ключем шифру не перевищить відстань єдиності), що визначається як

$$U = \frac{\log_2(K)}{D}, \tag{17}$$

де U – відстань єдиності, K – максимально можлива кількість переборів ключа шифру, D – надлишковість прийнятої для відображення повідомлень мови.

Для виконання умови дотримання відстані єдиності необхідно коректно визначити розрядність (довжину) ключа k у кореляції з довжиною повідомлення n, маючи на увазі, що

$$k = \log_2 N, \tag{18}$$

де k – розрядність ключа, N – кількість можливих значень повідомлення довжиною n.

Отже, залежність ентропії (тобто, довжини у бітах) ключа шифру від довжини повідомлення можна записати наступним чином:

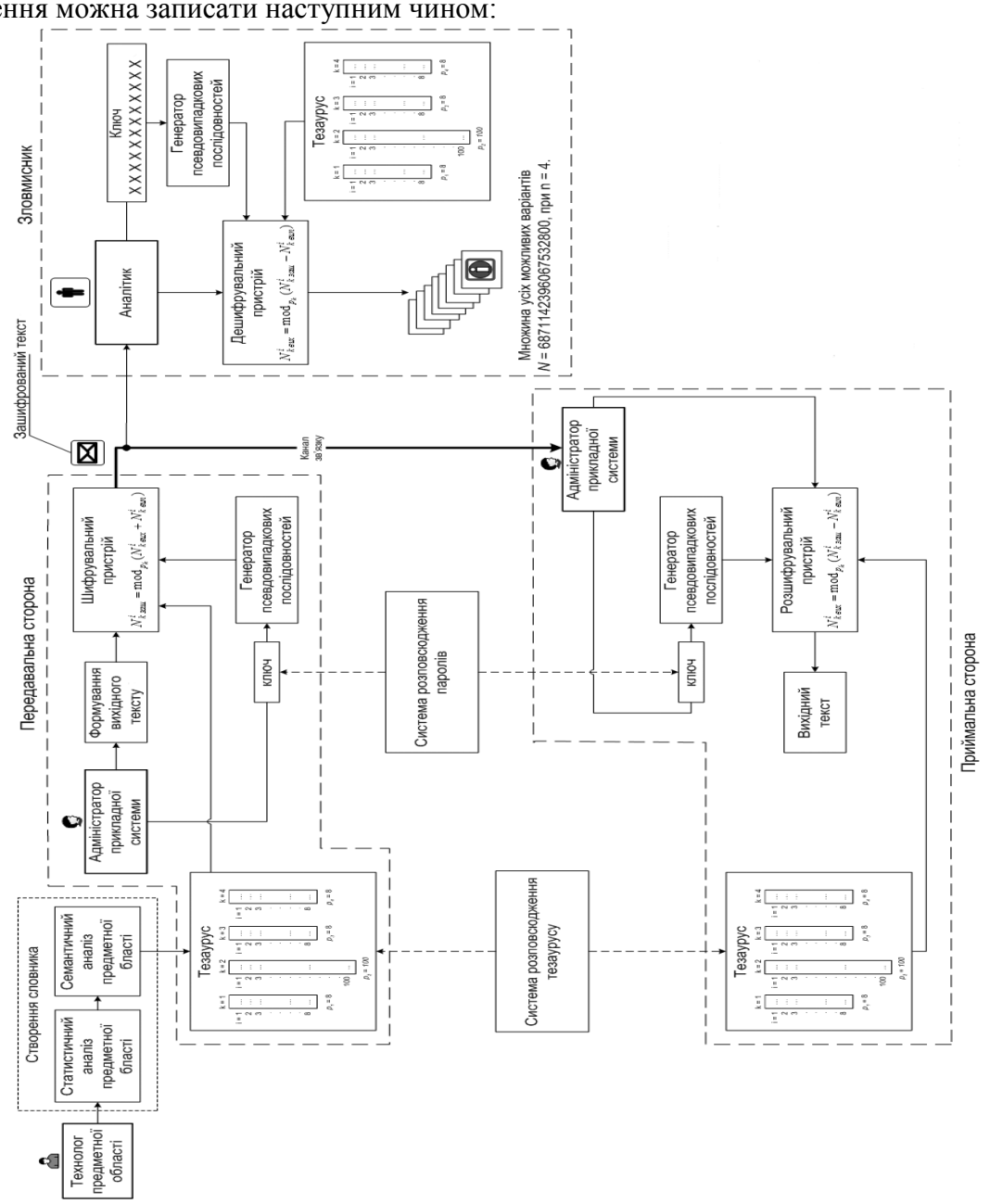


Рис. 6. Варіант побудови досконало стійкої криптосистеми з лексикографічною системою у своєму складі

$$H(K) = \log_2 \prod_{s=1}^{n} \left[B - (s-1) \right], \tag{19}$$

де B – кількість символів алфавіту мови, на якій представлене повідомлення.

Висновок щодо вищерозглянутої моделі атаки: КЗВЄ за умов дотримання відстані єдиності має властивості досконало секретної системи. Застосування методу є можливим лише для криптосистем, що використовуються у прикладних областях, для котрих створені відповідні штучні мови відображення цих прикладних областей, що характеризуються алфавітами великої розмірності. Метод не дозволяє позбавитись необхідності періодичної зміни ключів шифрів під час використання криптосистеми на практиці.

У четвертому розділі поставлено та вирішено завдання з розробки лексикографічного методу захисту текстової інформації, заснованої на використанні тезаурусу смислових образів прикладної області застосування, що забезпечує незалежність обсягу текстової інформації, що підлягає шифруванню, від відстані єдиності та довжини ключів шифру.

Спочатку надамо формальне обґрунтування можливості створення досконало стійких незалежних від довжини ключу шифру криптосистем.

Припустимо, що задано функціональний простір прикладної області застосування певної і-ої інформаційної системи $\mathbf{\Pi_i}$, цільове використання котрої пов'язано з передаванням, обробкою та зберіганням текстових повідомлень, у вигляді:

$$\mathbf{\Pi_i(\Phi_i,Z_i)} \in \{\mathbf{\Phi_i,Z_i}\} = \{\Phi_{i,1}, \Phi_{i,2}, \ldots \Phi_{i,N};\ Z_{i,1}, Z_{i,2}, \ldots Z_{i,M}\}, \tag{20}$$

де $\mathbf{\Phi_i}$ – функціональна структура і-ої інформаційної системи, що складається із елементів множини Φ_i припустимих для виконання функцій $\Phi_{i,к}$, де к=1, 2, …. , N; $\mathbf{Z_i}$ – функціональна структура і-ої інформаційної системи, що складається із елементів множини Z_i обмежувальних умов Z_i, де k = 1, 2,…,M, за яких виконання функцій і-ої інформаційної системи є припустимим; $\Phi_{i,k}$ – k-й функціональний елемент із заданої множини припустимих для виконання функцій і-ої інформаційної системи Φ_i, де k = 1, 2, …, N; $Z_{i,k}$ – k-та обмежувальна умова із заданої множини обмежувальних умов Z_i, де k = 1, 2, …, M; N, M – кількість елементів відповідно у множинах Φ_i та Z_i.

Вираз (20) задає функціональність простору $\mathbf{\Phi_i}$ прикладної області застосування і-ої інформаційної системи та простір умов $\mathbf{Z_i}$, за яких дана інформаційна система може бути використана за цільовим призначенням.

Необхідно забезпечити захист смислу інформації, що передається, оброблюється та зберігається засобами і-ої інформаційної системи, від порушень конфіденційності на рівні надання абсолютних гарантій неможливості злому системи захисту як з теоретичної (формальної), так і з практичної точок зору. При цьому обсяг інформації, що підлягає захисту, за будь-якої довжини ключа шифру, що використовується, має не обмежуватися довжиною цього ключа, а довжина ключа обиратися в залежності від припустимого значення ймовірності прийняття безпомилкових рішень в процесі дешифрування смислового потоку.

Формальний спосіб виконання завдання наступний. Простір заданої прикладної області $\mathbf{\Pi_i}$ відобразимо на простір смислових образів цієї системи $\mathbf{S_i}$ з урахуванням семантичних співвідношень між ними, що визначаються заданим простором обмежувальних умов $\mathbf{Z_i}$, тобто

$$\mathbf{\Pi_i(\Phi_i,Z_i)} \xrightarrow{\mathbf{P_{\Phi S(Zi)}}} \{\mathbf{S_i,Z_i}\}, \tag{21}$$

де $\mathbf{P_{\Phi S(Zi)}}$ – оператор відображення $\mathbf{\Pi_i}$ на $\mathbf{S_i}$ з урахуванням $\mathbf{Z_i}$.

Фактично $\{\mathbf{S_i,Z_i}\}$ задає структуру семантичного тезаурусу заданої прикладної області, формальний синтез котрого має здійснюватися згідно (4) – (8).

У структурі тезаурусу кожному смисловому образу із $\mathbf{S_i}$ поставимо у відповідність певним чином визначену множину інших смислових образів із цього ж $\mathbf{S_i}$, що знаходяться із

цим смисловим образом у відношенні смислової правдоподібності (див. вирази (5) та (6)). Така структура тезаурусу надає можливість за певних умов під час шифрування замінювати повідомлення, що підлягають шифруванню, на інші правдоподібні повідомлення, що не відображають істинний смисл вихідних повідомлень. А під час розшифрування, якщо пароль є відомим, здійснювати зворотні заміни правдоподібних зашифрованих повідомлень на істинні щодо смислу. Для зловмисника пароль є невідомим. Отже, можливостей щодо здійснення зворотних замін, як показав подальший аналіз цього методу захисту, він не має.

Процес шифрування текстових повідомлень представимо у вигляді:

$$\{\mathbf{S}\} \xrightarrow{\mathbf{P_{SF}}} \{\mathbf{F}\} \xrightarrow{\mathbf{P_{Ш}(x)}} \{\mathbf{F_{Ш}}\}, \tag{22}$$

де {**S**} – простір смислових образів, елементи котрого використовуються під час формування вихідних текстових повідомлень; {**F**} – простір відображень засобами обраної мови смислових образів, узятих із простору {**S**}; {$\mathbf{F_{Ш}}$} – простір зашифрованих відображень смислових образів із простору {**S**}; $\mathbf{P_{SF}}$ – оператор перетворення смислових образів у їхні відображення у рамках обраної мови, $\mathbf{P_{Ш}(x)}$ – оператор шифрування, що залежить від вибору ключа шифру **x.**

Процес розшифрування текстових повідомлень представимо у вигляді:

$$\{\mathbf{F_{Ш}}\} \xrightarrow{\mathbf{P_{рш}(x)}} \{\mathbf{F}\} \xrightarrow{\mathbf{P_{FS}}} \{\mathbf{S}\}, \tag{23}$$

де $\mathbf{P_{рш}(x)}$ – оператор розшифрування, що залежить від вибору ключа шифру **x**, а $\mathbf{P_{FS}}$ – оператор перетворення відображень смислових образів у вихідні смислові образи.

Для відтворення операторів $\mathbf{P_{SF}}$ та $\mathbf{P_{FS}}$ використаємо крипто-семантичний підхід до забезпечення досконало стійкого захисту інформації, а оператори $\mathbf{P_{Ш}(x)}$ та $\mathbf{P_{рш}(x)}$ реалізуємо будь-яким із відомих методів симетричної криптографії.

Підкреслимо, що у виразах (20) та (21) оператори $\mathbf{P_{Ш}(x)}$ та $\mathbf{P_{рш}(x)}$ не залежать від семантичних характеристик елементів простору {**S**}, тобто шифрування здійснюється без урахування смислу текстових повідомлень (саме тому методи класичної криптографії мають універсальний щодо смислу інформації характер).

Лексикографічна криптосистема щодо її побудови та принципу роботи не відрізняється від криптосистеми із збільшеною відстанню єдиності (КЗВЄ), що показана на рис. 6. (чи 7?) Відмінність – у структурі та вмісті семантичних тезаурусів. Зазвичай, у багатьох сферах прикладних застосувань (зокрема, у регламенті авіаційного радіообміну) призначають наступну ієрархію смислових одиниць, що відображають логічно завершені думки певного рівня абстрагування, де убування ступеня абстрагування прийнято зліва направо:

сценарій / ситуація / фраза / слово. (24)

У цьому випадку, будемо мати чотирьох ступеневу ієрархічну структуру у вигляді розгалуженого кореня тезаурусів з різним рівнем абстрагування представлення смислових образів. Для КЗВЄ достатньо забезпечити правдоподібність семантичних одиниць лише щодо одного обраного певним чином рівня абстрагування, але за умови дотримання відстані єдиності. У той час, як для лексикографічних криптосистем необхідно забезпечити правдоподібність семантичних одиниць на усіх чотирьох рівнях абстрагування, а також забезпечити семантичну пов’язаність усіх підсловників тезаурусу. Такими властивостями володіє так званий тезаурус бази захисту інформації, синтез котрого здійснено в роботі. Як результат, досконала стійкість лексикографічних криптосистем досягається без необхідності дотримання відстані єдиності.

В даному розділі також було розроблено алгоритм реалізації методу побудови лексикографічної криптосистеми.

Даний метод захисту смислу текстових повідомлень базується на забезпеченні неоднозначності зворотного переходу від відображення зашифрованої сутності (тобто, смислу або, іншим словом, суті змісту) вихідного текстового повідомлення, представленого у вигляді однієї із можливих форм цього відображення, до відображення істинної суті змісту вихідного текстового повідомлення. Будь-яка сутність може бути відображена різноманітними формами. Теорія крипто-семантики базується на гіпотезі, що, у загальному випадку, однозначного зв'язку між сутністю текстових повідомлень та формою їхнього відображення не існує. Отже, якщо перехоплені криптоаналітиком зразки зашифрованих повідомлень (у рамках формально визначених обмежень обраної області прикладних застосувань) являють собою семантично правдоподібні повідомлення, а будь-який результат застосування будь-якого із можливих способів криптоаналізу приводить до отримання нехай іншого щодо смислу, але теж семантично правдоподібного повідомлення, то істинність вихідного повідомлення не є можливим виявити, оскільки криптоаналітик не має можливостей визначити факт успішного завершення своєї роботи.

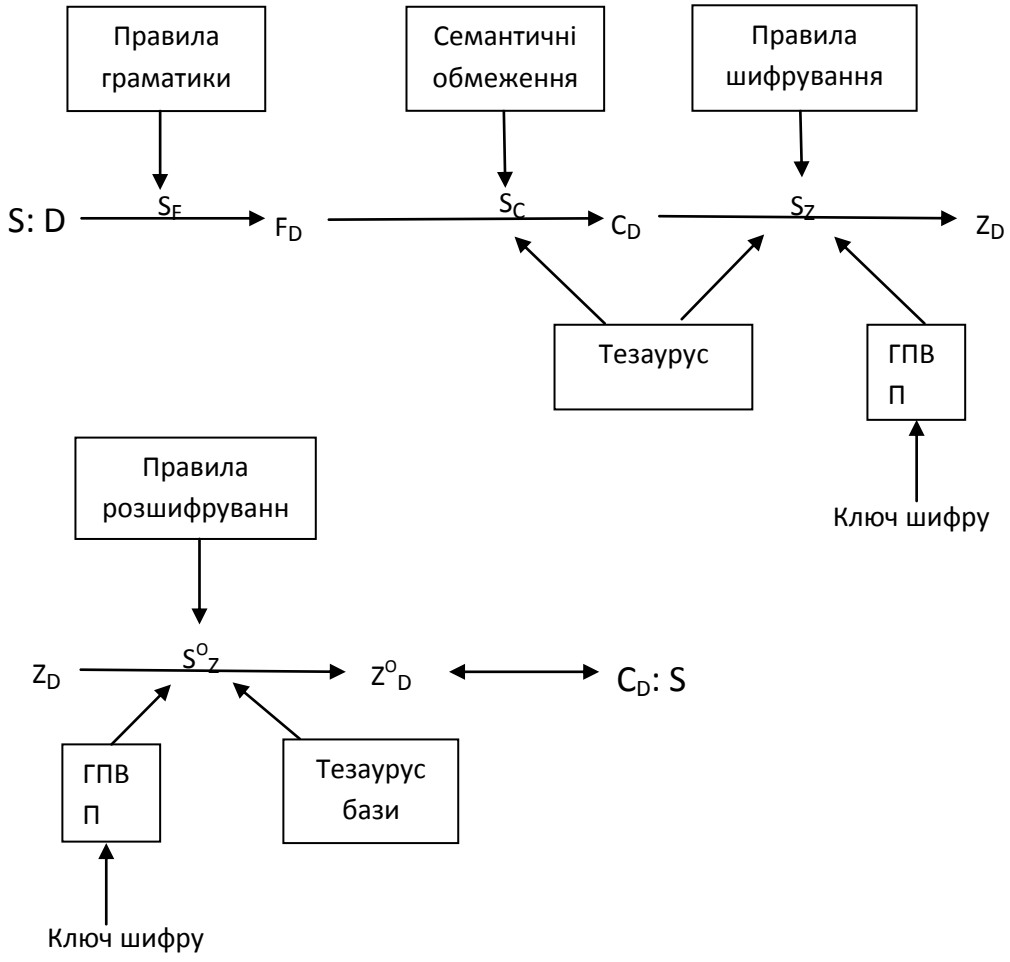


Рис. 7. Алгоритм реалізації лексикографічної криптосистеми

Умови для реалізації на симетричні крипто-семантичні системи:

а) усі семантичні одиниці вихідних текстових повідомлень є елементами тезаурусу смислових образів інформаційної системи, що захищається:

$$\{SO\} \in \{TZ_M\}, \quad (25)$$

де **{*SO*}** – кінцева множина смислових образів, що використовують різні суб'єкти в процесі виконання завдань у рамках визначеної області прикладних застосувань; **{TZ_M}** – тезаурус прикладної області, тобто упорядкований набір тлумачних (сценарних, фразеологічних, семантичних тощо) словників, що упорядковані за рівнями абстрагування смислових образів;

б) структура тезаурусу відображає структуру співвідношень між семантичними одиницями прикладної області, в рамках якої використовується інформаційна система, що захищається:

$$⟰\{SO\} \equiv ⟰\{TZ_M\}, \quad (26)$$

де ⟰ – знак профілю функціональності;

в) тезаурус побудовано таким чином, що кожній семантичній одиниці, що входить до складу тезаурусу **{TZ_M}**, відповідає кілька інших семантичних одиниць, що знаходяться між собою у відношенні семантичної правдоподібності:

$$SO_k^{(1)} \wedge \{SO_1^{(1)} \wedge SO_2^{(1)} \wedge \ \dots \ \dots \ \wedge SO_M^{(1)}\}, \quad (27)$$

де $SO_k^{(1)}$ – *к*-та семантична одиниця *і*-го рівня абстрагування; $\{ SO_1^{(1)} \wedge SO_2^{(1)} \wedge \ \dots \ \dots \ \wedge SO_M^{(1)} \}$ – множина із *M* правдоподібних щодо $SO_k^{(1)}$ семантичних одиниць, що входять до складу семантичного тезаурусу; $к = \{1, 2, \dots , M\}$; $i = \{1, 2, \dots , I\}$.

Також в рамках четвертого розділу даного дисертаційного дослідження було розроблено програмне забезпечення, що здійснює шифрування та відповідно розшифрування мовної інформації, що передається. Для реалізації даного прикладу було обрано використовувати прикладну область «Радіообмін диспетчер-пілот». Для реалізації поставленого завдання було вирішено використовувати мову програмування C# та вбудовані бібліотеки операційної системи Windows для розпізнавання мовної інформації. Скріншот цього працюючого додатку приведено на рис. 8.

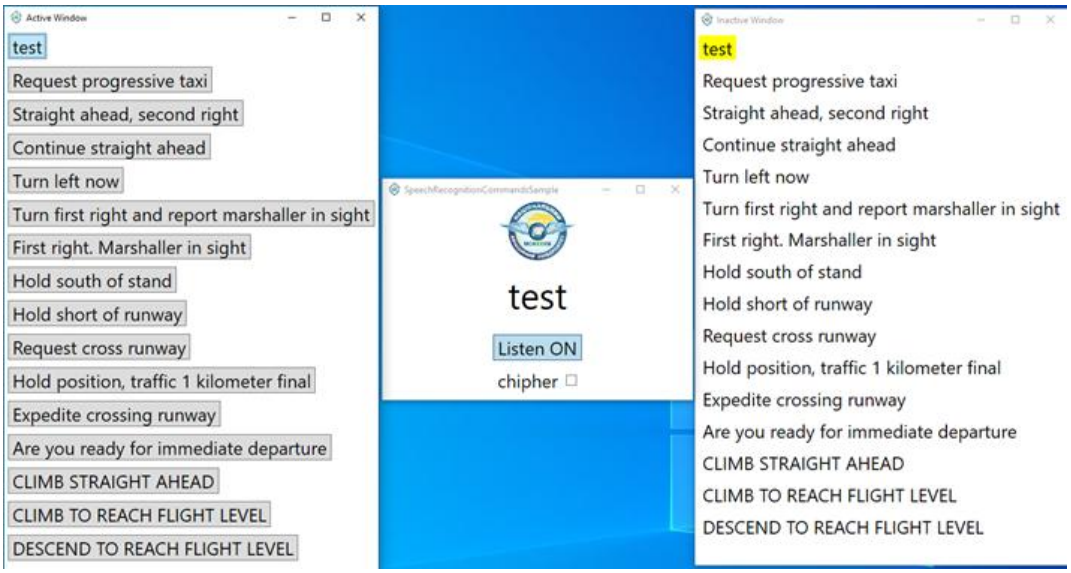


Рис. 8. Вигляд працюючого програмного додатку

ВИСНОВКИ

Сукупність наукових положень, сформульованих та обґрунтованих в дисертаційній роботі, складає вирішення науково-технічної задачі, яка полягала в необхідності підвищення ефективності шифрування текстових даних. У дисертаційній роботі отримані такі теоретичні та практичні результати.

1. В результаті проведеного аналізу було встановлено, що існуючі визначення і моделі, що використовують у задачах теорії інформації і зв'язку, не повною мірою придатні для опису систем семантичної криптографії. Для адекватного опису цих систем представлена

модель осмислення суб'єктом змісту інформації стосовно завдань семантичної криптографії.

2. Розроблено модель семантичних тезаурусів, придатних для використання у складі стійкої криптосистеми захисту текстової інформації. А саме: уведено показники семантичних зв’язків між смисловими конструкціями мови відображення прикладної області, насамперед показник правдоподібності, і на цій основі здійснити синтез структури тезаурусів. Надано структуру тезауруса у розтині за координатою, що характеризує рівень абстрагування відображення смислових образів. Ця структура однозначно визначає місця розташування семантичних одиниць у тезаурусі, що дозволяє автоматизувати процес його створення.

3. Вперше розроблено метод побудови лексикографічної криптосистеми, який за рахунок розширення базового алфавіту лінгвістичної формальної системи джерела мовних повідомлень забезпечує збільшення відстані єдиності шифру за ключем та дозволяє збільшити довжину шифрованих повідомлень відносно довжини ключової інформації та відповідно зменшити частоту зміни ключів шифру у порівнянні із схемою Вернама. У наведеному прикладі маємо майже 50-ти кратний виграш за даним критерієм ефективності.

4. Вперше запропоновано лексикографічний метод захисту текстової інформації, що за рахунок випадкової заміни первинного смислового образу повідомлення на інший правдоподібний елемент, узятий із семантичного тезаурусу бази захисту прикладної області, дозволяє забезпечити підвищену стійкість захисту при відсутності будь-яких обмежень на обсяг мовної інформації, що підлягає шифруванню, і, тим самим, усуває необхідність у періодичній зміні ключової інформації.

5. На базі розроблених методів та моделей було розроблене відповідно програмне забезпечення, яке надало змогу шифрувати/розшифровувати смислові образи.

ПУБЛІКАЦІЇ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Klimchuk V., Samoylik Y., Gnatyuk V., Prysiazhnyy D., Buryachok V. Synthesis of quite proof cryptosystem with increased unicity distance for cloud computing CEUR Workshop Proceedings 2104, pp. 596-607.

2. Самойлик Є.О. Ефективність досконало стійкої криптосистеми із збільшеною відстаню єдиності. Захист інформації. - Том 19, № 2 (2017). – С. 184-192.

3. Одарченко Р.С., Самойлик Є.О., Сімахін В.М., Боровик В.О., Тимчишин Р.М. Криптосемантична система захисту текстової інформації. Control Systems and Computers, N1, 2020. С. 35-46.

4. Р.С. Одарченко, Є.О. Самойлик, А.О. Абакумова Метод побудови семантичного словника у складі досконало стійкої криптосистеми захисту текстової інформації. Наукоємні технології. - № 3 (39), 2018. – С. 355-361.

5. Самойлик Є.О., Одарченко Р.С. Визначення характеристик лексикографічних систем, придатних для створення досконало стійких систем захисту текстової інформації. Вісник інженерної академії наук. - № 2, 2018. – С. 96-102.

6. Одарченко Р.С., Гнатюк В.О., Самойлик Є.О. Удосконалена архітектура системи безпеки стільникових мереж нового покоління. Наукоємні технології в інфокомунікаціях: обробка, захист та передача інформації [Текст] : монографія / під заг. ред.: В. М. Безрука, В. В. Баранніка ; Харків. нац. ун-т радіоелектроніки. – Харків : Бровін О. В., 2018. – 327 с.

7. Самойлик Є.О., Одарченко Р.С. Визначення характеристик лексикографічних систем з позицій технічного захисту текстової інформації. Збірник тез всеукраїнської науково-практичної конференції молодих учених і студентів «Проблеми навігації та управління рухом розвитку глобальної системи зв’язку, навігації, спостереження та організації повітряного руху CNS/ATM»: 21-23 листопада 2018 р. - С. 28.

8. Самойлик Є.О., Одарченко Р.С., Жмурко Т.О., Лукашенко В.В. Структура семантичного тезаурусу для лексикографічних систем. Збірник тез ІІІ міжнародної науково-технічної конференції «Інформаційна безпека у сучасному суспільстві»: 29-30 листопада 2018 р., С. 34-35.

9. Hrystak A., Kinzeryavyy V., Prysiazhnyi D., Burmak Y., Samoylik Y. High-speed and hash function for blockchain security mechanisms. Scientific and practical cyber security journal (SPCSJ) 4(1): 65-70 ISSN 2587-4667 Scientific Cyber Security Association (SCSA).

10. Самойлик Е.А., Одарченко Р.С. Крипто-семантическая система защиты текстовой информации. Тезисы докладов I Международной научно-практической конференции «Современные технологии кибербезопасности», 14 червня 2019 р., м. Алмати, Казахстан. - С. 102-111.

11. Самойлик Є.О. Структура системи криптосемантичного захисту інформації // Збірник тез міжнародної науково-практичної конференції молодих учених і студентів «Політ-2019.Сучасні проблеми науки»: 3–4 квітня 2019 р.: тези доп. – К., 2019. – С. 114.

12. Самойлик Є.О. Практичні рекомендації по організації криптосемантичного каналу захисту текстової інформації // Збірник тез науково-практичної конференції «Проблеми експлуатації та захисту інформаційно-комунікаційних систем»: 7 – 9 червня 2019 р.: тези доп. – К., 2019. – С. 67-68.

13. Самойлик Є.О., Одарченко Р.С. Визначення характеристик лексикографічних систем з позицій технічного захисту текстової інформації. Збірник тез всеукраїнської науково-практичної конференції молодих учених і студентів «Проблеми навігації та управління рухом розвитку глобальної системи зв'язку, навігації, спостереження та організації повітряного руху CNS/ATM»: 21-23 листопада 2018 р. - С. 28.

АНОТАЦІЯ

Самойлик Євген Олександрович Лексикографічні методи захисту мовної інформації. – Рукопис.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.21 – Системи захисту інформації. – Національний авіаційний університет, Київ, 2020.

Дисертаційна робота присвячена створенню стійких симетричних криптосистем, що не пред'являють жорстких вимог до системи розповсюдження ключової інформації. Розроблено метод побудови семантичного словника, який задає семантичну структуру словника прикладної області. Розроблено метод побудови лексикографічної системи захисту мовної інформації, який за рахунок укрупнення алфавіту джерела текстових повідомлень забезпечив збільшення відстані єдиності шифру, що надало змогу суттєво збільшити довжину шифрованих повідомлень відносно довжини ключової інформації. Розроблено метод побудови стійкої криптосистеми, заснованої на використанні лексико-графічних систем захисту текстової інформації, що забезпечує незалежність обсягу текстової інформації, що підлягає шифруванню, від довжини ключів шифру, що використовуються, у рамках прикладних застосувань, для яких створено прикладні семантичні тезауруси.

Ключові слова: захист інформації, симетричні криптосистеми, лексикографічні системи, відстань єдиності шифру за ключем, семантичний тезаурус.

АННОТАЦИЯ

Самойлик Евгений Александрович Лексикографические методы защиты речевой информации. – Рукопись.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.21 – Системы защиты информации. – Национальный авиационный университет, Киев, 2020.

Диссертационная работа посвящена созданию устойчивых симметричных криптосистем, которые не предъявляют жестких требований к системе распространения ключевой информации. Разработан метод построения семантического словаря, который задает семантическую структуру словаря предметной области. Разработан метод построения лексикографической системы защиты речевой информации, который за счет укрупнения алфавита источника текстовых сообщений обеспечил увеличение расстояния единственности шифра, что дало возможность существенно увеличить длину шифрованных сообщений относительно длины ключевой информации. Разработан метод построения устойчивой криптосистемы, основанной на использовании лексико-графических систем

защиты текстовой информации, который обеспечивает независимость объема текстовой информации, подлежащей шифрованию, от длины ключей шифра, используемых в рамках прикладных применений, для которых созданы прикладные семантические тезаурусы.
Ключевые слова: защита информации, симметричные криптосистемы, лексикографические системы, расстояние единственности, семантический тезаурус.

ABSTRACT

Samoylik Yevgen Olexsandrovych Lexicographical methods for speech information security. - Manuscript.

Thesis for the degree of candidate of technical sciences, specialty 05.13.21 - Information security systems - National aviation university, Kyiv, 2020.

The dissertation is devoted to the creation of stable symmetric cryptosystems. A method for constructing a semantic dictionary suitable for use as part of a robust cryptosystem for protecting textual information has been developed. The structure of the thesaurus in the coordinate section is presented, which characterizes the level of abstraction of mappings of semantic images. This structure uniquely determines the location of semantic units in the thesaurus, which allows to automate the process of its creation. Formal definitions of indicators of semantic relations between semantic units, such as semantic identity, semantic difference, doubtful semantic identity (or plausibility), doubtful semantic difference (or false similarity) are introduced. A methodological approach to protecting the meaning of text messages based on the use of the relationship of semantic plausibility is proposed. In order for a cryptanalyst under no circumstances to form any idea of the truth of the meaning of intercepted text messages, it is necessary and sufficient to replace the meaning of semantic images contained in the original text messages with their plausible representations, which are taken from the corresponding thesauri of the chosen language. areas of activity.

A method for constructing a robust cryptosystem for protection against breaches of confidentiality of textual information is proposed, which is taken from the semantic dictionary of a predefined application area and placed in a given tabular form. The method is based on the application of the mechanism of enlargement of the alphabet of the language of reflection of textual information, which uses the dictionary of the application area and takes into account the structure of the tabular form. As a result of the enlargement of the alphabet, the so-called uniqueness distance increases, which is the main threshold indicator of the cryptosystem's belonging to the class of stable protection systems with theoretically proven ideal theoretical and information stability.

A synthesis of a crypto-semantic method for protecting the meaning of text messages has been performed. In accordance with this method, the space of a given application area of application of the i-th information system is mapped to the space of semantic images of this system, taking into account the semantic relationships between them, which are determined by a given space of restrictive conditions. That is, the structure of the semantic thesaurus of a given application area is synthesized, which allows each semantic image from the space of semantic images to match many other semantic images from the same space, which are with this semantic image in terms of semantic plausibility. This structure allows to replace the original messages with other plausible messages during encryption, which do not reflect the true meaning of the original messages. And during decryption at the known password information to carry out return replacements of plausible false messages on true in sense. The intercepted samples of messages encrypted by this method are presented in the form of semantically plausible messages. Any result of applying any of the known methods of cryptanalysis leads to the receipt of plausible messages. So it is impossible to reveal the truth of the original message, because the cryptanalyst is unable to determine the fact of successful completion of its work.

Keywords: technical priv, quite proof symmetric cryptosystem, lexical-graphic systems, distance of unicity after the key, semantic thesaurus.