# Esteganografía Lingüística en Lengua Española: Aplicación en Acciones de Ciberseguridad y Cibercrimen dentro del Ejército Colombiano

Por

JUAN DANIEL GÓMEZ MOYA

Presentado ante el Departamento de Lingüística de la Facultad de Ciencias Humanas de la Universidad Nacional de Colombia, Bogotá, Colombia.

Con apoyo de:
**INTERPOL Information, communications and technology (ICT)**

**2023**

# MONOGRAFÍA

## LINGUISTICS GRADUATE DISSERTATIONS

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:

**LINGÜISTA**

**Metodología revisión: ©Methodologies in Computational Linguistics MIT & ©Writing Guides Strategic Resources & Digital Publications Yale University**


Director (a):

**John Freddy Chaparro Rojas**

**Magíster en Lingüística de la Universidad Nacional de Colombia**

Apoyo de:

**Capitán Laura Cristina Núñez Rivera**[1]

**Jefe del Observatorio del Delito de la Dirección de investigación Criminal e INTERPOL**


Línea de Investigación:

**Military Linguists**


Universidad Nacional de Colombia

Facultad de Ciencias Humanas

Departamento de lingüística

Bogotá, Colombia

**2023**


[1] Profesional en Relaciones Internacionales y Estudios Políticos Especialista en Gestión Pública. Cinco años de experiencia en análisis criminológico y administración de información criminal. Actualmente lidera la participación de la Policía Nacional en temas de Política Criminal, ante el Ministerio de Justicia y del Derecho.

*Dios, familia*

*Babel* בבל

“Solo podemos ver poco del futuro,
pero lo suficiente para darnos cuenta
de que hay mucho que hacer.”
**Alan Turing**
**CRIPTÓGRAFO Y LÓGICO INFORMÁTICO**

"טאָלפּ וואַ יימראַ וואַ טיִמ טקעלאַיד אַ זיא דאַרפש אַ"
“Una lengua es un dialecto con un ejército y una marina”
**Max Weinreich**
**LINGÜISTA**

# Sobre la guerra

Al seleccionar las palabras para esta disertación, me adentré con sumo cuidado en un territorio complejo. Reflexionar sobre la innata tendencia beligerante del ser humano se revela casi infructuosamente. En este terreno se entrecruzan el sufrimiento, la codicia, el poder, la muerte, la tortura, el tormento, la pena y la resistencia. Por otro lado, emerge la fuerza, la resiliencia, la paz y la esperanza, una dualidad que perpetuamente danza en el escenario mundial. Al comprometerme con este estudio monográfico, me vi confrontado tristemente con la lucha interna de mis valores éticos, espirituales y filosóficos en relación con esta innata naturaleza combativa. Más allá de mi orientación política, me impone la obligación de mantener una objetividad inflexible y un enfoque lógico, casi matemático, al abordar la criptografía y esteganografía en el contexto del poderío militar de las naciones.

Es innegable que pienso en las acciones del Estado de Israel, que inflige injusticia y sufrimiento a aquellos con afines al Islam, así como tambien reflexiono sobre el papel de Irán, Turquía, Afganistán, Palestina y Siria, por arrastrar, al estado sionista occidental, a una guerra religiosa. Del mismo modo, no puedo pasar por alto las intervenciones belicosas de la OTAN en conflictos transnacionales, como el generado en Ucrania, que vulneran el derecho de las fronteras y la elección de vivir y expresarse en una cultura rusa. Sin embargo, también considero necesario meditar sobre Rusia y su papel al abandonar la diplomacia y desencadenar acciones militares.No puedo evitar considerar el manejo del gobierno colombiano en el conflicto armado, sobre los presidentes de tendencia que, junto a grupos paramilitares, desplazaron, ultrajaron y asesinaron a personas en los Montes de María, Cauca, Putumayo y Arauca. De igual manera, no puedo eximir dentro de mi análisis el comportamiento de las guerrillas de izquierda, que han ocultado su presunta causa social bajo una capa de codicia ligada al narcotráfico, el secuestro, el terrorismo y el reclutamiento forzado.

No me atrevo a emitir un juicio sumario sobre la guerra en sí, pues hablar de un bando implica abordar también al otro, convirtiéndose en una lucha pírrica. Esta monografía, al igual que todo conocimiento científico, busca proporcionar una comprensión profunda de la ciencia esteganográfica. Mi esperanza es que pueda servir como una herramienta que contribuya a la seguridad nacional y, al mismo tiempo, a la protección de la vida humana, en lugar de convertirse en un instrumento de control y poder que busque arrebatarla.

# Resumen

**Esteganografía Lingüística en Lengua Española: Aplicación en Acciones de Ciberseguridad y Cibercrimen en el Ejército Colombiano**

Aunque la esteganografía no es una técnica novedosa para ocultar mensajes, los nuevos canales digitales sí lo son. Esto ha resultado en la aparición de innumerables formas de crear sistemas de comunicación innovadores. Las agencias estatales de seguridad, como la CIA o la FSB (ФСБ), han experimentado un aumento exponencial en recursos y personal para hacer frente a las nuevas estrategias que las organizaciones criminales están empleando para establecer canales de comunicación secretos. Este fenómeno, sumado al rápido desarrollo tecnológico, hace imperativo reconsiderar la seguridad nacional de los países. En la actualidad, es posible ocultar mensajes incluso en una cadena de ADN. Los países más susceptibles a que los grupos criminales desarrollen nuevas formas de comunicación oculta son aquellos que han estado en conflicto durante años, como: Israel, China. Rusia, Estados Unidos y Colombia. Según el informe de la ONU de 2021, Colombia se mantiene como el mayor productor de cocaína en el mundo, lo que constituye la principal fuente de financiación para los grupos guerrilleros y paramilitares. En vista de esta realidad, es esencial mejorar los estudios lingüísticos en el campo de la esteganografía digital en Colombia. La historia ha demostrado que la comunicación juega un papel fundamental en la inteligencia y contrainteligencia de una nación. Este trabajo tiene como objetivo describir el desarrollo y la importancia de esta ciencia en la aplicación de nuevas estrategias de seguridad nacional en Colombia. Para ello, se realizará una revisión bibliográfica que proporcionará las herramientas necesarias para proponer factores en los que la lingüística puede contribuir a la inteligencia y contrainteligencia en los nuevos entornos digitales.

**Palabras clave: (Esteganografía lingüística, Lingüística Computacional, Lingüística Militar, Esteganotextos, Ciberseguridad)**

# Abstract


### Linguistic Steganography in Spanish Language: Application in Cybersecurity and Cybercrime Actions in the Colombian Army.[2]

Although steganography is not a novel technique for hiding messages, new digital channels are. This has resulted in the emergence of countless ways to create innovative communication systems. State security agencies, such as the CIA or the FSB (ФСБ), have experienced an exponential increase in resources and personnel to cope with the new strategies criminal organizations are employing to establish secret communication channels. This phenomenon, coupled with rapid technological development, makes it imperative to reconsider the national security of nations. Today, it is possible to hide messages even in a DNA chain. The countries most susceptible to criminal groups developing new forms of hidden communication are those that have been in conflict for years, such as Israel, China. Russia, the United States and Colombia. According to the 2021 UN report, Colombia remains the largest cocaine producer in the world, which is the main source of funding for guerrilla and paramilitary groups. In view of this reality, it is essential to improve linguistic studies in the field of digital steganography in Colombia. History has shown that communication plays a fundamental role in the intelligence and counterintelligence of a country.This paper aims to describe the development and importance of this science in the application of new national security strategies in Colombia. For this purpose, a literature review will be conducted to provide the necessary tools to propose factors in which linguistics can contribute to intelligence and counterintelligence in the new digital environments.


**Keywords: (Linguistic Steganography, Computational Linguistics, Military Linguistics, Steganotexts, Cybersecurity)**

[2] Revisado por **Lina Maria Robayo Acuña** Magíster en Lingüística Aplicada de la Universidad de York (York, Inglaterra) Licenciada en Filología en Inglés y Lingüística de la Universidad Nacional de Colombia.

# תקציר

**סטגנוגרפיה לשונית בשפה הספרדית: יישום בפעולות אבטחת סייבר ופשעי סייבר בצבא קולומביה**[3]

למרות שסטגנוגרפיה אינה טכניקה חדשה להסתרת מסרים, ערוצים דיגיטליים חדשים כן. זה הביא להופעתם של חוו ,(FSB (ФСБ-או ה CIA-אינספור דרכים ליצור מערכות תקשורת חדשניות. סוכנויות הביטחון הממלכתיות, כמו ה גידול אקספוננציאלי במשאבים ובכוח אדם כדי להתמודד עם האסטרטגיות החדשות שבהן משתמשים ארגוני פשע להקמת ערוצי תקשורת סודיים. תופעה זו, המתווספת להתפתחות טכנולוגית מהירה, מחייבת לשקול מחדש את הביטחון המדינות הרגישות ביותר לקבוצות פליליות .DNA הלאומי של מדינות. כיום ניתן להסתיר מסרים גם בשרשרת המפתחות צורות חדשות של תקשורת נסתרת הן אלו שנמצאות בעימות במשך שנים, כמו ישראל, סין. רוסיה, ארצות הברית וקולומביה. על פי דו"ח האו"ם לשנת 2021, קולומביה נותרה היצרנית הגדולה ביותר של קוקאין בעולם, המהווה את מקור המימון העיקרי של קבוצות גרילה ופארא-צבאיות. לאור מציאות זו, חיוני לשפר את לימודי הבלשנות בתחום הסטגנוגרפיה הדיגיטלית בקולומביה. ההיסטוריה הראתה שלתקשורת יש תפקיד מהותי במודיעין ובאינטליגנציה הנגדית של מדינה. עבודה זו נועדה לתאר את הפיתוח והחשיבות של מדע זה ביישום אסטרטגיות ביטחון לאומי חדשות בקולומביה. לשם כך תתבצע סקירה ביבליוגרפית שתספק את הכלים הדרושים להצעת גורמים שבהם הבלשנות יכולה לתרום למודיעין ולמודיעין נגדי בסביבות דיגיטליות חדשות.

**מילות מפתח: סטגנוגרפיה לשונית, בלשנות חישובית, בלשנות צבאית, סטגנוטקסטים, אבטחת סייבר**

[3] Revisado por **Christian Marcos Cantor** exdirector de la Lucha Contra la Proliferación de Armas del Ministerio de Relaciones Exteriores de Israel.

# 摘要

西班牙语语言隐写术:在哥伦比亚军队网络安全和网络犯罪行动中的应用[4]

尽管隐写术并不是一种新颖的信息隐藏技术，但新的数字渠道却是。因此，出现了无数种创建创新通信系统的方法。中央情报局或联邦安全局（ФСБ）等国家安全机构的资源和人员急剧增加，以应对犯罪组织为建立秘密通信渠道而采取的新策略。这一现象，再加上技术的飞速发展，使得我们必须重新考虑各国的国家安全问题。如今，即使在 DNA 链中也可以隐藏信息。最容易被犯罪集团开发新的隐蔽通信方式的国家是那些多年来一直处于冲突中的国家，如以色列、中国、俄罗斯、美国和哥伦比亚。俄罗斯、美国和哥伦比亚。根据联合国 2021 年的报告，哥伦比亚仍然是世界上最大的可卡因生产国，而可卡因是游击队和准军事组织的主要资金来源。鉴于这一现实，必须加强哥伦比亚数字隐写术领域的语言学研究。历史表明，通信在一个国家的情报和反情报工作中发挥着根本性的作用，本文旨在介绍这门科学在哥伦比亚实施新的国家安全战略中的发展和重要性。为此，本文将进行文献综述，以提供必要的工具，提出在新的数字环境中语言学可促进情报和反情报工作的因素。

关键词:（语言隐写术、计算语言学、军事语言学、隐写文本、网络安全）。

[4] Revisado por **Wuqiang Hou (**侯武强**)**: Profesor de chino con Maestría en Educación.

# Аннотация

**Лингвистическая стеганография на испанском языке: применение в кибербезопасности и борьбе с киберпреступностью в колумбийской армии[5]**

Хотя стеганография не является новой техникой сокрытия сообщений, новые цифровые каналы являются таковыми. Это привело к появлению бесчисленных способов создания инновационных систем связи. Государственные службы безопасности, такие как ЦРУ или ФСБ, испытывают экспоненциальный рост ресурсов и персонала, чтобы справиться с новыми стратегиями, которые используют преступные организации для создания секретных каналов связи. Это явление в сочетании с бурным развитием технологий заставляет пересмотреть вопрос о национальной безопасности государств. Сегодня можно скрыть сообщения даже в цепочке ДНК. Наиболее восприимчивыми к разработке преступными группировками новых форм скрытой связи являются страны, находящиеся в состоянии многолетнего конфликта, такие как Израиль, Китай, Россия, США и Колумбия. Россия, США и Колумбия. Согласно докладу ООН за 2021 год, Колумбия остается крупнейшим производителем кокаина в мире, который является основным источником финансирования партизанских и военизированных группировок. Учитывая эту реальность, необходимо совершенствовать лингвистические исследования в области цифровой стеганографии в Колумбии. История показывает, что коммуникация играет фундаментальную роль в разведке и контрразведке страны, и данная статья ставит своей целью описать развитие и значение этой науки в применении новых стратегий национальной безопасности в Колумбии. Для этого будет проведен обзор литературы, который даст необходимые инструменты для предложения факторов, с помощью которых лингвистика может способствовать разведке и контрразведке в новых цифровых средах.

**Ключевые слова: (лингвистическая стеганография, вычислительная лингвистика, военная лингвистика, стеганотекст, кибербезопасность)**

[5] Revisado por **Eduardo Rosero** Profesor de lingüística y dialectología hispanoamericana Российский Университет Дружбы Народов (РУДН).

## TABLA DE CONTENIDO

1. Introducción ..... 14
1.1. Contextualización del Tema ..... 14
1.2. Declaración del Problema ..... 14
1.3. Objetivos de la Investigación ..... 17
1.4. Justificación y Relevancia ..... 18
1.4.1 Contribución al Conocimiento Científico ..... 18
1.4.2 Aplicación Práctica en la Ciberseguridad Militar……18
1.4.3 Impacto Ético y Legal ..... 18
1.5. Metodología de Revisión ..... 19

2. Fundamentos de Esteganografía ..... 20
2.1. Fundamentos Básicos ..... 20

3. Historia ..... 21
3.1 Los Orígenes de la Esteganografía en la Civilización Clásica …….. .. 22
3.2 El Renacimiento de la Esteganografía en la Edad Media ..... 22
3.3 La Esteganografía en Épocas de Guerra: Siglos XVIII y XIX …… ... 22
3.4 La Revolución Digital: La Esteganografía en el Siglo XX ..... . 23

4. Lingüística Computacional y Modelos Lingüísticos ..... 24
4.1. Principales Aspectos de la Lingüística Computacional ..... 25
4.1.1 Procesamiento del Lenguaje Natural (PLN) ..... 25
4.1.2 Modelos Lingüísticos ..... . 25
4.1.3 Algoritmos de Procesamiento de Lenguaje Natural ..... 25
4.1.4 Aprendizaje Automático y PLN ..... 25
4.1.5 Recursos Lingüísticos ..... 25
4.1.6 Aplicaciones Prácticas ..... 25

5. Procesamiento de Lenguajes Naturales (PLN) ..... 26
5.1. Pioneros del PLN ..... 26
5.2. Alan Turing, Enigma y criptografía ..... 27
5.3. Matemáticas en Enigma ..... 31
5.3.1 Posibles Combinaciones en el Panel de Enchufes ..... 32
5.4 BOMBE ..... 33
5.4.1 Explicación del Código de “Bombe” en Python Básico ..... 33
5.5. Claude Shannon ..... 34
5.5.1 Obras Destacadas ..... 34
5.6.2 Impacto en el Procesamiento del Lenguaje Natural ……..... 36
5.6.3 Complejidades Sintácticas de Chomsky ..... 36
5.6.4 Ejemplo de Gramática Sensible en Python ..... 37
5.6.5 Ejemplo de Gramática Sensible en Java ..... 38

6. Entropía y Esteganografía ..... 39

7. Criptografía Cuántica y Lingüística Computacional ..... 45

8. Sobre Algunas Consideraciones Filosóficas ..... 48
8.1 Aspectos Éticos y Legales de la Esteganografía ..... 50
8.1.1 Antecedentes de “Chuzadas” por parte del DAS ..... 50

9. Técnicas de Esteganografía ..... 52
9.1 Estudios Relacionados en Estados Unidos ..... 52
9.1.1 Manuscrito Voynich ..... 52
9.1.2 Lengua Natural o Mentira Elaborada ..... ..... 54
9.1.3 Antecedentes del Manuscrito ..... 55
9.1.4 Entropía ..... 58
9.1.5 Reflexiones sobre el Manuscrito Voynich y la Esteganografía …….. 60
9.2 Illegals Program ..... 60
9.3 Estudios Relacionados en Rusia ..... 64
9.3.1 Malwares en la Ciberguerra Kaspersky Monty Three ..... 65
9.3.2 Módulo HttpTransport: Tareas de Red ..... 76
9.3.3 ¿Quién Está Detrás de Este Malware? ..... 77
9.4 Estudios Relacionados en China ..... 78
9.4.1 Errores Tipográficos en el Chino ..... 78
9.4.2 Textos y Redes Sociales en China ..... 79
9.4.4 Ejemplificación y Experimentación ..... 81
9.5 Estudios Relacionados en Israel ..... 85
9.5.1 Universidad de Tel Aviv ..... 85
9.5.2 Israel y la Cooperación en Medio de la Crisis ..... 86
9.5.3 La Transferencia de Información entre Entidades ..... 88

10. Esteganografía Digital ..... 91
10.1 Imagen y Audio Digital ..... 93
10.2 Métodos de Portadores Digitales ..... 96
10.3 Ejemplos de Esteganografía ..... 97
10.4 Detección de Esteganografía ..... 101
10.5 Herramientas para la Detección de Esteganografía ..... 103

11. Esteganografía Digital Avanzada ..... 111
11.1 Técnica ADN ..... 111
11.2 Identificación de SNP y Puntos de Acceso de SNP ..... 112

12. Características Lingüísticas ..... 114
12.1 Lengua Inglesa ..... ..... 114
12.2 Lengua Rusa ..... 115
12.2.1 Características Lingüísticas y Esteganográficas del Ruso ……. 116
13.3 Lengua China ..... 118
12.4 Lengua Hebrea ..... 119

13. Comparación Sociopolítica ..... 125
13.1 Narcotráfico y Conflictos Armados: Colombia y Afganistán ..... 126
13.2 Desafíos Emergentes y Esteganografía Criminal ..... 127
13.3 La Divisa Naranja en Colombia ..... 127
13.3.1. Comunicaciones Militares en el ENC ..... 128

14. Estudios de Esteganografía en Lengua Española ..... 130
14.1 Herramientas y Recursos Tecnológicos ..... 136

15. Conclusiones ..... 137
15.1 Recomendaciones ..... 137
15.2 Resumen de los Hallazgos Clave ..... 138
15.3 Contribuciones al Campo ..... 138
15.4 Recomendaciones para Futuras Investigaciones ..... 138

16. Referencias ..... 139

# 1. Introducción

## 1.1. Contextualización del tema

La esteganografía lingüística en lengua española se sitúa en la intersección de la lingüística computacional y la ciberseguridad, representando un área de investigación vital en un mundo digitalmente interconectado. En la era actual, la protección de la información sensible se ha convertido en una prioridad, especialmente en contextos militares como el Ejército Colombiano, donde la ciberseguridad y la gestión de cibercrimen son elementos críticos para la defensa y la seguridad nacional. Por ello surge la necesidad de comprender y desarrollar técnicas avanzadas de esteganografía lingüística específicamente adaptadas al español. A medida que las amenazas cibernéticas evolucionan, es esencial explorar nuevas formas de ocultar información en el lenguaje, aprovechando las complejidades lingüísticas únicas que presenta el español. Esta contextualización se erige sobre la base del crecimiento exponencial de las comunicaciones digitales en el ámbito militar y la urgente demanda de salvaguardar la integridad y confidencialidad de la información en el ciberespacio.

En los últimos 12 meses, Colombia ha emergido como el tercer país más atacado por ciberdelincuentes en América Latina, con un preocupante registro de 117 ataques por minuto. A pesar de una ligera disminución del 3% en comparación con el mismo periodo del año anterior, la nación ha enfrentado una asombrosa cantidad de 1.196 ,256,196 ataques de malware. Estas cifras revelan la creciente sofisticación y persistencia de las amenazas cibernéticas en la región. En contraste, Brasil lidera la tabla con 1,515 ataques por minuto, seguido de cerca por México con 275. Fabio Assolini, director del equipo de investigación y análisis para América Latina de Kaspersky[6], destaca la urgencia de fortalecer las medidas de ciberseguridad en Colombia ante este panorama, donde se reportan 3.277,414 ataques diarios en América Latina, equivalente a 37.9 por segundo.

## 1.2. Declaración del problema

El problema central radica en la necesidad de desarrollar y adaptar métodos de esteganografía lingüística que sean efectivos y seguros en la lengua española. A medida que la información sensible se convierte en un activo estratégico, la capacidad de ocultar datos en el propio tejido lingüístico se vuelve esencial para preservar la confidencialidad y la integridad de la información. A pesar de la relevancia del tema, la literatura actual carece de un análisis exhaustivo y aplicado de la esteganografía lingüística específica para el español, especialmente en el ámbito militar colombiano. La falta de investigaciones especializadas limita la capacidad de los profesionales de la ciberseguridad y las fuerzas armadas para desarrollar y aplicar estrategias efectivas de ocultación de información en una lengua que presenta particularidades sintácticas, semánticas y gramaticales únicas.La resolución de este problema no solo contribuirá al conocimiento académico en ciberseguridad, sino que también proporcionará herramientas y enfoques prácticos para fortalecer las capacidades de protección de información en el Ejército Colombiano y otros contextos de seguridad.

[6] Es una compañía internacional dedicada a la seguridad informática con presencia en aproximadamente 195 países del mundo. Su sede central se encuentra en Moscú, Rusia, mientras que el holding está registrado en Reino Unido. Recuperado de: https://latam.kaspersky.com/

Jeimy Cano[7], en su análisis sobre la ciberseguridad en Colombia durante el 2021, destaca la complejidad de los 20 meses, en su estudio, marcados por la pandemia y sus consecuencias. En este periodo, tanto individuos como organizaciones tuvieron que enfrentar desafíos inesperados, llevándolos a cuestionar su situación actual y reconocer la falta de preparación para enfrentar una prolongada e inestable crisis. En el contexto digital, Cano señala que los adversarios han capitalizado la transformación digital de empresas y personas, aprovechando entornos híbridos para perpetrar distracciones y engaños tanto en el ámbito físico como digital.

Cano destaca la necesidad de prestar atención a las formas tradicionales e híbridas de fraude, especialmente en un escenario global marcado por tensiones sanitarias, polarizaciones políticas y el aumento acelerado de lo digital. El análisis se centra en eventos de seguridad y ciberseguridad en Colombia, destacando incidentes mediáticamente relevantes, como la fuga de información del Ejército Nacional, el ciberataque a la Universidad El Bosque, la Aeronáutica Civil, las Empresas Públicas de Cali (EMCALI), el Departamento Administrativo Nacional de Estadística (DANE), y la Pontificia Universidad Javeriana. Añadiendo a esto Cano extrae lecciones aprendidas de cada incidente, resaltando la importancia de protocolos eficientes para reportar fugas, la gestión de vulnerabilidades, la preparación en casos de crisis, y la conciencia situacional de las personas como primera línea de defensa.

La autora hace énfasis en el apuro de entender la ciberseguridad, no solo como un referente tecnológico sino, como un componente esencial de la misión y esencia de las organizaciones. Subraya la importancia de una acción coordinada y comunicada frente a eventos adversos, buscando mantener el orden y la disciplina interna para limitar los daños. A continuación una síntesis de la problemática:

| **Ataque** | **Organización Afectada** | **Breve Descripción** | **Lecciones Aprendidas** |
|---|---|---|---|
| **Fuga de Información en las Fuerzas Militares** | Ejército Nacional de Colombia | Agentes de inteligencia de un país cercano provocaron la fuga de información sensible. | 1. Protocolo eficiente para reportar fugas.<br><br>2. Identificación y control de la información clave.<br><br>3. Entender el contexto de la fuga para un mejor tratamiento de la información. |
| **Ciberataque a la Universidad El Bosque** | Universidad El Bosque | Atacantes tomaron control de sistemas internos, incluyendo cuentas de correo y redes sociales. | 1. Gestión de vulnerabilidades y parches.<br><br>2. Establecimiento de un plan de gestión de |

[7] Directora de Revista SISTEMAS at Asociación Colombiana de Ingenieros de Sistemas - ACIS // orcid.org/0000-0001-6883-3461

| | | | |
|---|---|---|---|
| | | | crisis definido y simulado.<br><br>3. Conciencia de que las redes sociales pueden ser utilizadas por los atacantes. |
| Ciberataque a la **Aeronáutica Civil** | Aeronáutica Civil | Malware tipo ransomware en el correo interno, llevando a la desactivación de servicios. | 1. Conciencia situacional y formación continua de personal.<br>2. Establecer umbrales de operación necesarios para condiciones adversas.<br><br>3. Desarrollar un "libro de jugadas" para eventos adversos. |
| **Ciberataque a EMCALI** | Empresas Públicas de Cali (EMCALI) | Ataque a infraestructura crítica con ransomware afectando servicios y clientes. | 1. Inventario de puntos de conexión y equipos.<br>2. Plan de ejercicios de simulación para infraestructuras críticas.<br>3. Participación de grupos de interés en prácticas de seguridad. |
| **Ciberataque al DANE** | Departamento Administrativo Nacional de Estadística (DANE) | Ataque a entidad gubernamental con malware, afectando datos estadísticos y sistemas. | 1. Definición de infraestructuras críticas y marcos de aseguramiento.<br>2. Sistema de alertas operacionales basadas en el funcionamiento normal.<br>3. Asegurar copias de seguridad de datos según reglas establecidas. |
| **Ciberataque a la Pontificia Universidad Javeriana** | Pontificia Universidad Javeriana | Pocos detalles disponibles, pero indicativos de ransomware y medidas preventivas. | 1. No depender de "seguridad por oscuridad", compartir detalles básicos de los ataques.<br>2. Contar con una red de apoyo en casos de crisis.<br>3. Trabajo conjunto en monitoreo convergente de tecnología y seguridad. |

Finalmente, Cano reflexiona sobre la opacidad generada por el temor de las organizaciones de comunicar detalles de los ataques, señalando que cerrar los canales de información puede proporcionar un espacio cómodo para los adversarios. Destaca que la ciberseguridad no es simplemente un problema técnico, sino un componente fundamental en un entorno digital en constante evolución, donde proteger la promesa de valor de una organización implica desarrollar la confianza digital con sus clientes.

## 1.3. Objetivos de la investigación

### Objetivo General

Explorar, desarrollar y evaluar técnicas avanzadas de esteganografía lingüística específicamente adaptadas a la lengua española, con un enfoque aplicado en acciones de ciberseguridad y cibercrimen en el Ejército Colombiano.

### Objetivos Específicos

- Analizar críticamente las limitaciones actuales de las técnicas de esteganografía lingüística aplicadas al español, identificando vacíos y áreas de mejora en la literatura existente.

- Desarrollar y adaptar metodologías de esteganografía lingüística que aprovechen las particularidades sintácticas, semánticas y gramaticales del español, con el objetivo de ocultar información de manera eficaz.

- Validar experimentalmente las técnicas propuestas utilizando conjuntos de datos relevantes para el ámbito militar colombiano, evaluando la eficacia de la ocultación de información y la resistencia a los intentos de detección.

- Establecer pautas y mejores prácticas para la implementación de técnicas de esteganografía lingüística en entornos de ciberseguridad militar, considerando aspectos éticos y legales.

- Contribuir al cuerpo de conocimientos científicos y técnicos en el campo de la esteganografía lingüística, proporcionando un marco teórico sólido y aplicado que pueda ser referente para investigaciones futuras.

- Investigar las implicaciones de la esteganografía en Colombia, identificando los desafíos y consideraciones éticas asociados con su aplicación en operaciones de ciberespionaje y seguridad nacional comparándolo con 4 países potencias en esteganografía: Estados Unidos, Rusia, China e Israel.

## 1.4. Justificación y Relevancia

La esteganografía lingüística adquiere una importancia significativa en el panorama de la ciberseguridad y acciones contra el cibercrimen, especialmente en el contexto del Ejército Colombiano. La justificación de esta investigación se cimienta en diversas consideraciones científicas y prácticas. Siendo Colombia el país 18 con más ciberataques a nivel mundial, se deben generar proyectos independientes a proveedores extranjeros que busquen una autosuficiencia, científica, criptografía y esteganografía en el país.

**1.4.1 Contribución al Conocimiento Científico:** Esta investigación aspira a llenar este vacío mediante un análisis crítico, riguroso y novedoso de las técnicas existentes y la propuesta de metodologías adaptadas a las complejidades lingüísticas del español. Al contribuir al cuerpo de conocimientos científicos, se abre un espacio para el desarrollo y la expansión de la esteganografía lingüística como disciplina científica.

**1.4.2 Aplicación Práctica en la Ciberseguridad Militar:** La implementación exitosa de técnicas de esteganografía lingüística diseñadas específicamente para el español puede fortalecer las capacidades de ciberseguridad y contrainteligencia del Ejército Colombiano. Al ocultar información en el propio tejido lingüístico, se abre la posibilidad de comunicación segura y confidencialidad, contribuyendo así a la seguridad nacional.

La elección de contextualizar esta investigación en el español y en el contexto colombiano no solo responde a la necesidad de abordar problemas específicos, sino que también reconoce la importancia de considerar factores culturales y lingüísticos en la ciberseguridad. La riqueza del español y sus variaciones regionales exigen un enfoque especializado, y esta investigación busca llenar este vacío mediante la aplicación de técnicas ajustadas a las particularidades de la lengua.

**1.4.3 Impacto Ético y Legal:** El estudio de la esteganografía lingüística también plantea consideraciones éticas y legales cruciales. La justificación de esta investigación incluye la necesidad de establecer pautas éticas para el uso de estas técnicas, así como consideraciones legales para su implementación en contextos militares. La transparencia y la responsabilidad son aspectos fundamentales que se abordarán para garantizar un uso ético y legal de las técnicas propuestas.

**1.5. Metodología de Revisión**

La metodología de revisión adoptada para esta monografía se basa en las directrices proporcionadas por las metodologías en Lingüística Computacional del MIT de Michael Collins & Stuart Shiebery así mismo los recursos estratégicos de redacción de la Universidad de Yale.

- **Enfoque General:** Nos sumergimos en la Lingüística Computacional, que combina la ciencia de la computación con el estudio del lenguaje humano. La revisión se estructura en dos vertientes principales, siguiendo el modelo de desarrollo propuesto por Warren Weaver en 1949 y la aplicación de razonamiento computacional de Noam Chomsky en la década de 1950.

- **Límites del Estudio:**Nos centraremos específicamente en la Esteganografía Lingüística, explorando cómo el ocultamiento de información en el lenguaje puede afectar la ciberseguridad. Definiremos los límites de nuestra investigación para garantizar la profundidad y la relevancia, abordando de manera específica los aspectos relacionados con las técnicas esteganográficas y su aplicación en el ámbito de la seguridad cibernética.

- **Selección de Casos y Evaluación:** La revisión se llevará a cabo a través de estudios de caso seleccionados, presentados por expertos en el campo de la Lingüística Computacional. Se evaluarán estos casos para comprender la gama completa de metodologías de investigación en Esteganografía Lingüística y su conexión con la ciberseguridad. Esta metodología asegura una exploración exhaustiva y contextualizada de la Esteganografía Lingüística, destacando su importancia en el ámbito de la seguridad cibernética.

  *1. Establecimiento de la pregunta*
  *2. Cuantificación de los efectos*
  *3. Localización de los estudios*
  *4. Criterios de inclusión/exclusión*
  *5. Búsqueda de información y datos*
  *6. Evaluación de la calidad de los estudios incluidos*
  *7. Análisis de la heterogeneidad de los estudios*
  *8. Combinación de resultados*
  *9. Identificación del sesgo de publicación*
  *10. Análisis de sensibilidad*

# 2. FUNDAMENTOS DE ESTEGANOGRAFÍA

## 2.1. Fundamentos básicos

La esteganografía, derivada de las palabras griegas "steganos" (oculto) y "graphien" (escribir), es el arte y la ciencia de ocultar información dentro de datos aparentemente normales. A diferencia de la criptografía, que se centra en cifrar la información para proteger su contenido, la esteganografía se enfoca en esconder la existencia misma del mensaje. Esta disciplina milenaria tiene aplicaciones tanto en contextos históricos como en el mundo digital contemporáneo. En esencia, la esteganografía busca camuflar la información sensible dentro de archivos o medios de comunicación cotidianos, conocidos como "portadores". Estos portadores pueden ser imágenes, archivos de audio, videos, documentos de texto, entre otros. El objetivo es que, a simple vista, el portador no revele la presencia del mensaje oculto. La raíz de la esteganografía radica en ocultar el mensaje encubierto sin oscurecer el hecho de que dos partes están comunicándose entre sí. El proceso esteganográfico generalmente implica colocar un mensaje oculto en algún medio de transporte incómodo para los contextos cotidianos. Así mismo el recado secreto se incrusta en el portador para formar el medio esteganográfico por lo tanto puede emplearse una clave esteganografía para cifrar el mensaje oculto o para la aleatorización en el esquema esteganográfico. Su estructura básica se fundamenta en la figura 1.1.

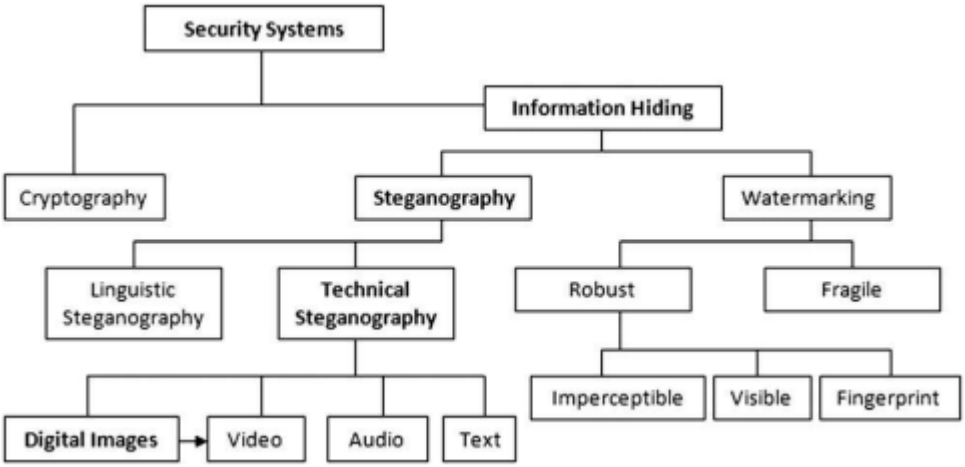



*Figura 1.1. Esteganografía lingüística en los sistemas de seguridad*

Las técnicas esteganográficas varían según el tipo de medio y la complejidad deseada. En el ámbito digital, se modifican en espacios creativos e incómodos. La esteganografía no solo ha evolucionado con la tecnología, sino que también ha desempeñado un papel crucial a lo largo de la historia. Desde antiguas técnicas como escribir mensajes en la piel tatuada de esclavos hasta métodos digitales sofisticados, la esteganografía ha sido utilizada en diversos contextos, como la inteligencia militar, la protección de secretos comerciales y la preservación de la privacidad. La detección de esta ciencia a menudo implica análisis estadísticos avanzados y la aplicación de contramedidas específicas. En el entorno digital, existen herramientas especializadas tanto para ocultar como para descubrir información esteganográfica. Aunque la esteganografía puede tener aplicaciones legítimas, su uso plantea desafíos éticos y legales. Así mismo puede ser utilizada con

fines maliciosos, como el intercambio encubierto de información en operaciones clandestinas, o con propósitos más benignos, como garantizar la seguridad de datos sensibles.

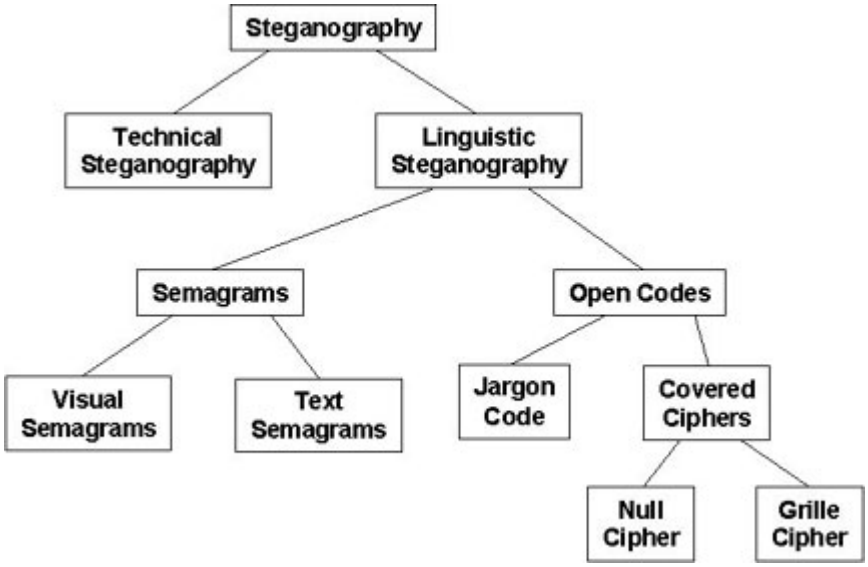



*Figura 1.2. Técnicas de esteganografía lingüística*

## 3. HISTORIA

### 3.1 Los Orígenes de la Esteganografía en la Civilización Clásica

Según Tábara (s. f.) en el texto "Breve historia de la Criptografía Clásica" la esteganografía tiene sus raíces en las civilizaciones clásicas de la antigüedad. Durante los siglos V a.C. al IV d.C., en el auge de la antigua Grecia y la Roma Imperial, surgieron los primeros registros de esta técnica, marcando el nacimiento de una disciplina que perduraría a través de los siglos.Uno de los primeros indicios de esteganografía en la civilización clásica se encuentra en los relatos del historiador griego Heródoto, quien vivió en el siglo V a.C. Heródoto, conocido como "el padre de la historia", relató la intrigante historia de Histiaeus, un antiguo gobernante de Mileto que había sido exiliado y buscaba comunicarse con su hijo Aristágoras, quien estaba bajo vigilancia persa. La solución ingeniosa de Histiaeus fue la de escribir su mensaje en la cabeza afeitada de un esclavo y permitir que el cabello volviera a crecer, ocultando así la información dentro del cuero cabelludo. Una vez que el cabello había crecido lo suficiente, el esclavo se envió a Mileto, donde se afeitó la cabeza y reveló el mensaje secreto.

Además de esta historia, se atribuye a Demeritus, un escriba griego de la antigüedad, la invención de una técnica esteganográfica ingeniosa. Demeritus desarrolló una práctica que implicaba escribir mensajes en tablillas de cera y luego recubrirlas con una fina capa de cera transparente. A simple vista, las tablillas parecían documentos comunes e inofensivos. Sin embargo, aquellos que conocían el método podían raspar la capa superior de cera y descubrir el mensaje oculto en la cera inferior. Esta técnica permitía la comunicación discreta en un momento en que la información podía ser un recurso preciado y peligroso. La esteganografía de la civilización clásica no se limitaba a la ocultación de mensajes en objetos físicos. También se extendía a la comunicación oral y gestual, donde se utilizaban señales secretas, códigos y metáforas para transmitir información de manera encubierta.

### 3.2. El Renacimiento de la Esteganografía en la Edad Media

La Edad Media fue testigo de un resurgimiento de la esteganografía, marcando una época en la que esta disciplina se desarrolló y se perfeccionaron técnicas más avanzadas. En medio de la oscuridad y el misterio de esta era histórica, la esteganografía se convirtió en un instrumento valioso para la comunicación secreta y la protección de información sensible.Durante este período, que abarca desde los siglos V al XV, la esteganografía se transformó y evolucionó, alejándose de las técnicas primitivas de la antigüedad para dar paso a métodos más elaborados. Uno de los momentos más destacados de esta era fue la creación de una obra que se convertiría en un referente en la historia de la esteganografía: "Steganographia" escrita por Johannes Trithemius en el siglo XV.

Johannes Trithemius, un monje y erudito alemán, publicó su monumental obra "Steganographia" en 1499. Aunque el título podría sugerir que se trata de un manual de esteganografía convencional, el contenido de la obra de Trithemius iba mucho más allá. En sus páginas, Trithemius no solo detallaba métodos para ocultar mensajes en textos, sino que también exploraba otros temas como la criptografía y la comunicación con seres sobrenaturales. La obra estaba diseñada para ser hermética y confusa, con el propósito de que solo aquellos con el conocimiento adecuado pudieran descifrar su verdadero significado.

Uno de los aspectos más intrigantes de "Steganographia" fue su enfoque en la magia y la comunicación con seres espirituales. Trithemius argumentaba que se podían utilizar técnicas de esteganografía para invocar a ángeles y demonios y comunicarse con ellos de manera segura. Si bien estas ideas pueden parecer extrañas a los estándares modernos, reflejan la mentalidad mística y esotérica de la época medieval. La obra de Trithemius ilustra cómo la esteganografía no solo se utilizaba con fines militares o de espionaje, sino que también estaba ligada a la espiritualidad y la búsqueda del conocimiento oculto. Esto demuestra la versatilidad de la esteganografía y su capacidad para adaptarse a las necesidades y creencias cambiantes a lo largo de la historia.

### 3.3. La Esteganografía en Épocas de Guerra: Siglos XVIII y XIX

En los turbulentos siglos XVIII y XIX, la esteganografía demostró ser una herramienta crucial para la comunicación secreta y la protección de información sensible, particularmente durante conflictos armados como la Guerra de Independencia de los Estados Unidos. En este período histórico, la esteganografía continuó evolucionando y adaptándose a las necesidades cambiantes de la guerra y la diplomacia. Durante la Guerra de Independencia de los Estados Unidos (1775-1783), un conflicto que vio a las colonias americanas luchando por su libertad contra el Imperio Británico. Según Jennifer Wilcox directora del museo de criptología de los Estados Unidos[8] La guerra de Independencia sirvió como un potente catalizador de técnicas esteganográficas.

Los combatientes, conscientes de la importancia de mantener sus comunicaciones fuera del alcance de los británicos, recurrieron a técnicas de esteganografía para garantizar la seguridad de la información. Una de las técnicas más notables utilizadas durante la Guerra de Independencia fue la ocultación de mensajes en periódicos y cartas. Los mensajes secretos se camuflaban en artículos periodísticos o en cartas aparentemente comunes. Estos mensajes eran cuidadosamente redactados y a menudo requerían de conocimientos compartidos entre el remitente y el destinatario para ser decodificados correctamente. Esta práctica permitía a los patriotas compartir información crucial sin alertar a las fuerzas británicas.Además de los periódicos y las cartas, se emplearon otras formas de esteganografía, como la escritura invisible y el uso de microtextos. La escritura invisible

[8] Cryptology in the American Revolution: Steganography 1 https://www.youtube.com/watch?v=07rL5c8wiVM

implicaba el uso de sustancias químicas o tintas especiales que solo se volvían visibles bajo ciertas condiciones, como la aplicación de calor o la exposición a la luz ultravioleta. Los microtextos eran mensajes escritos en letras extremadamente pequeñas, lo que los hacía prácticamente invisibles a simple vista.

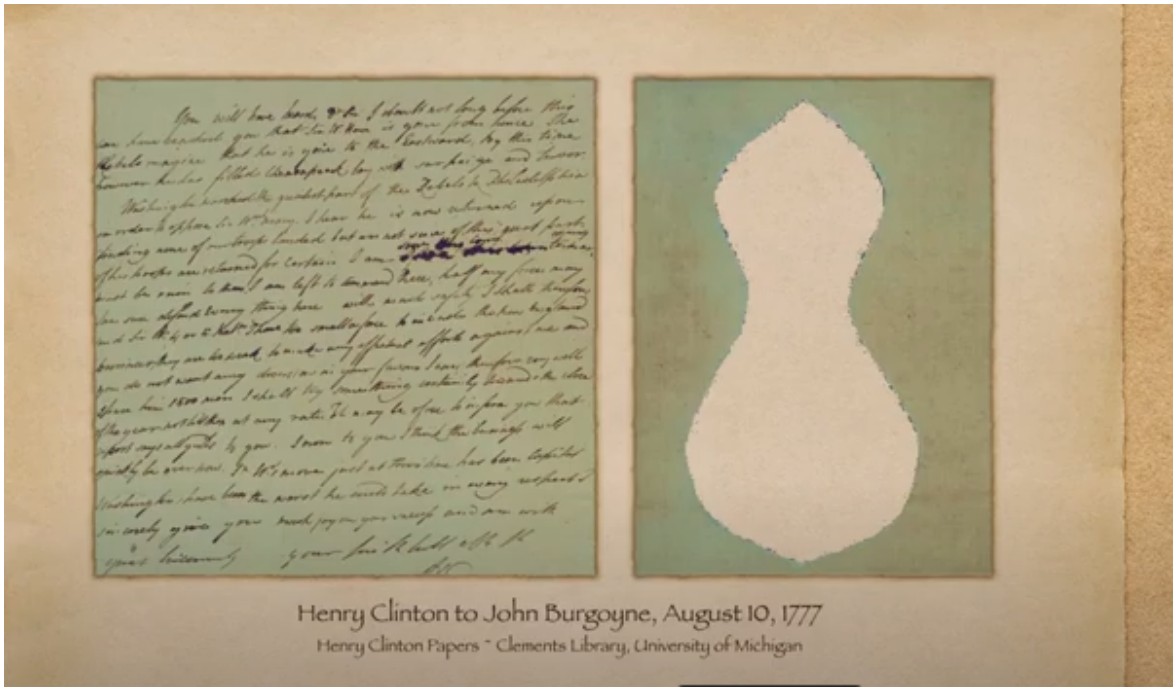



***Imagen 1***. *Mensaje de Henry Clinton con la respectiva máscara que, al sobreexponerse, mostraba el verdadero mensaje para Jonh Bungoyne. Recuperado de University of Michigan (2020)*

La Guerra de Independencia de los Estados Unidos no fue el único conflicto en el que se utilizó la esteganografía en esta época. Durante las guerras napoleónicas y otros conflictos europeos, se aplicaron técnicas similares para garantizar la seguridad de la comunicación militar y diplomática. Estos ejemplos históricos ilustran cómo la esteganografía se convirtió en una herramienta indispensable en tiempos de guerra y conflicto, permitiendo a los actores involucrados proteger sus comunicaciones y mantener la ventaja estratégica. Además, señalan la importancia de la evolución continua de las técnicas esteganográficas para adaptarse a las necesidades cambiantes de la seguridad de la información en contextos de conflicto.

### 3.4. La Revolución Digital: La Esteganografía en el Siglo XX

El siglo XX marcó un punto de inflexión en la evolución de la esteganografía, impulsado por el advenimiento de la informática y la era digital. En esta era de avances tecnológicos acelerados, la esteganografía dio un salto significativo al entrar en el mundo digital. Se desarrollaron técnicas innovadoras para ocultar datos dentro de archivos digitales, como imágenes, audio y video, lo que transformó por completo el panorama de la seguridad de la información. Con el aumento de la conectividad y el intercambio de datos a través de redes informáticas, surgió la necesidad de proteger información confidencial de manera más efectiva. La esteganografía digital se convirtió en una herramienta esencial para garantizar la confidencialidad y la integridad de los datos en un entorno cada vez más propenso a amenazas cibernéticas.

Una de las aplicaciones más notables de la esteganografía digital fue la capacidad de ocultar mensajes dentro de archivos multimedia, como imágenes JPEG, archivos de audio MP3 y videos.

Estos formatos digitales proporcionaron un lienzo perfecto para ocultar datos, ya que permitían introducir información adicional sin que fuera perceptible para el ojo humano o el oído humano. Por ejemplo, en imágenes digitales, los datos se podían insertar en los valores de los píxeles de manera que los cambios fueran imperceptibles a simple vista. En el caso del audio, se podían modificar ligeramente las frecuencias o los niveles de audio para ocultar información sin que fuera detectable para el oído humano. Esta técnica se utilizó tanto en aplicaciones de seguridad como en la distribución de contenido multimedia con marcas de agua digitales para identificar la propiedad intelectual.

La esteganografía digital también encontró aplicaciones importantes en el ámbito forense y legal. Se convirtió en una herramienta esencial para rastrear y descubrir pruebas digitales en casos de ciberdelitos y amenazas cibernéticas. Los investigadores forenses podían utilizar técnicas esteganográficas para descubrir información oculta en archivos digitales y determinar su autenticidad.Además de su uso en aplicaciones de seguridad y forenses, la esteganografía digital también encontró su lugar en la industria del entretenimiento y la gestión de derechos de autor, donde se aplicaron marcas de agua digitales para proteger contenido multimedia.

## 4. LINGÜÍSTICA COMPUTACIONAL Y MODELOS LINGÜÍSTICOS

La lingüística computacional es un campo interdisciplinario que combina la lingüística y la informática para abordar cuestiones relacionadas con el procesamiento del lenguaje natural (PLN). Se centra en el desarrollo de algoritmos y modelos computacionales que permiten a las máquinas entender, interpretar y generar lenguaje humano de manera efectiva. La lingüística computacional, abordada por Javier Gómez Guinovart de la Universidad de Vigo, se presenta como un campo interdisciplinario con aproximadamente cincuenta años de investigación y desarrollo. Su objetivo principal es integrar la capacidad de procesamiento del lenguaje humano en las computadoras.

Desde la perspectiva de la informática, se la considera una subdisciplina de la inteligencia artificial, que a su vez se enfoca en comprender la inteligencia y diseñar máquinas inteligentes con características humanas. Por otro lado, desde la perspectiva de la lingüística, se puede ver como una subdisciplina de la lingüística teórica, ya que busca desarrollar modelos formales del lenguaje humano implementables informáticamente. Además, como disciplina experimental con aplicaciones comerciales e investigativas, forma parte de las industrias de la lengua, proporcionando datos y programas informáticos aplicados al tratamiento del lenguaje. Este enfoque práctico vincula la lingüística computacional con la lingüística aplicada. El artículo destaca tres vertientes de trabajo: informática aplicada a la investigación lingüística, implementación de teorías lingüísticas y aplicaciones lingüísticas de la informática (Gómez Guinovart, 1998, p. 135).

La aplicación de la informática a la investigación lingüística, conocida como lingüística informática o informática aplicada a la lingüística, ha avanzado significativamente en las últimas décadas (Lezcano et al., 1997). Se ilustra este progreso mediante un ejemplo que destaca la utilidad de los ordenadores en el análisis morfológico del inglés temprano, específicamente del sufijo -able. Se compara un enfoque tradicional, basado en la observación manual de textos, con la eficiencia que proporciona la informática aplicada al análisis de un corpus digital de medio millón de palabras compilado entre 1500 y 1710 (Lezcano et al., 1997).

Este enfoque informático facilita la rápida obtención de datos sobre la frecuencia y ubicación de palabras con el sufijo -able. Además, destaca la importancia de los corpus etiquetados morfológicamente para la investigación lingüística y el desarrollo de aplicaciones de

procesamiento del lenguaje natural. Estos corpus permiten trabajar con palabras no ambiguas respecto a su categoría gramatical, lo que simplifica la detección de patrones morfosintácticos superficiales. La implementación de teorías lingüísticas, la segunda línea de trabajo, se centra en la elaboración de modelos lingüísticos formalmente implementables. Ejemplos incluyen la gramática léxica funcional (LFG), la gramática sintagmática generalizada (GPSG), y la gramática sintagmática dirigida por el núcleo (HPSG), todos agrupados bajo el término de gramáticas de unificación (Shieber, 1986).

Por último, la tercera línea se centra en las aplicaciones lingüísticas de la informática, como las tecnologías del habla, la traducción automática y los sistemas de extracción de información. Estos desarrollos tecnológicos tienen un impacto significativo en campos como la verificación lingüística automática, diccionarios electrónicos y sistemas de diálogo persona-máquina (Lezcano et al., 1997). Este resumen destaca la intersección entre la lingüística y la informática, resaltando el papel crucial de la tecnología en la investigación y aplicación de principios lingüísticos.

### 4.1. Principales aspectos de la lingüística computacional

**4.1.1 Procesamiento del Lenguaje Natural (PLN):** La lingüística computacional se ocupa del PLN, que es una rama de la inteligencia artificial que se centra en la interacción entre las computadoras y el lenguaje humano. El PLN implica tareas como análisis morfológico, análisis sintáctico, análisis semántico, y generación de lenguaje.

**4.1.2 Modelos Lingüísticos:** Los modelos lingüísticos son representaciones matemáticas y computacionales de la estructura y el funcionamiento del lenguaje. Estos modelos pueden incluir reglas gramaticales, estadísticas de frecuencia de palabras y patrones semánticos. Los modelos lingüísticos son esenciales para desarrollar sistemas de PLN efectivos.

**4.1.3 Algoritmos de Procesamiento de Lenguaje Natural:** Se desarrollan algoritmos específicos para llevar a cabo tareas de PLN. Esto puede incluir algoritmos de análisis sintáctico para comprender la estructura gramatical de las oraciones, algoritmos de reconocimiento de entidades para identificar nombres propios, y algoritmos de traducción automática para convertir texto de un idioma a otro.

**4.1.4. Aprendizaje Automático y PLN:** El aprendizaje automático, y en particular el aprendizaje profundo (deep learning), ha tenido un impacto significativo en la lingüística computacional. Los modelos basados en redes neuronales profundas han demostrado ser eficaces para tareas como el reconocimiento de voz, la traducción automática y la generación de texto.

**4.1.5. Recursos Lingüísticos:** La lingüística computacional depende de recursos lingüísticos, como corpus lingüísticos (conjuntos de textos) etiquetados, diccionarios y ontologías. Estos recursos son esenciales para entrenar y evaluar modelos lingüísticos.

**4.1.6 Aplicaciones Prácticas:** La lingüística computacional tiene numerosas aplicaciones prácticas, incluyendo motores de búsqueda basados en lenguaje natural, asistentes virtuales, sistemas de traducción automática, análisis de sentimientos en redes sociales y más. La lingüística computacional y los modelos lingüísticos desempeñan un papel crucial en la creación de tecnologías que permiten a las máquinas comprender y utilizar el lenguaje humano de manera efectiva. Estos avances tienen un impacto significativo en campos como la inteligencia artificial, la comunicación asistida por computadora y la mejora de la interacción hombre-máquina.

Hasta ahora, C y C++ han sido herramientas fundamentales en el ámbito de la programación. En el caso de la programación en WINDOWS, es común utilizar VISUAL BASIC, una extensión del antiguo BASIC, o VISUAL C++, una extensión de C++. Por otro lado, para desarrollar programas destinados a la ejecución en Internet, JAVA emerge como una solución sólida. Aunque guarda similitudes con C++, un programa en Java tiene la ventaja de poder ejecutarse sin modificaciones en la mayoría de las computadoras disponibles en el mercado, siempre y cuando cuenten con el intérprete Java instalado.

Tanto C++ como Java se utilizan principalmente para aplicaciones y productos finales. Sin embargo, en el ámbito de la investigación y el procesamiento interno de datos, se recurre con frecuencia a lenguajes de script como PERL. Para la creación y ejecución de programas en Perl y Java, se pueden obtener intérpretes de forma gratuita en Internet[9]. En el desarrollo de software con fines comerciales, es común emplear intérpretes y entornos de desarrollo más avanzados, como Visual C++ para aplicaciones en el entorno Windows. Es esencial realizar un curso en estructuras de datos para aquellos interesados en la programación. Textos de autores como Alfred Aho o Robert Sedgewick ofrecen una visión detallada de los contenidos de un curso de este tipo. En dicho curso, los estudiantes aprenden a traducir soluciones a problemas expresadas verbalmente en pasos específicos en algoritmos escritos en un lenguaje de programación como C. Además, adquieren habilidades para organizar datos básicos en estructuras más complejas y desarrollar algoritmos efectivos para clasificar y buscar dichas estructuras.

En la actualidad, el manejo de bases de datos es imperativo, con conocimientos sobre claves primarias, claves secundarias y la prevención de redundancia de información. SQL se erige como el lenguaje estándar para bancos de datos y resulta esencial para aquellos que desean dedicarse seriamente al desarrollo de programas. En el ámbito matemático, se requieren conocimientos de lógica declarativa y predicativa, así como de lógica modal y lógica difusa. Además, la familiaridad con la teoría de conjuntos, funciones, autómatas y la teoría de lenguajes formales resulta beneficioso. Conceptos básicos de teoría de grafos y álgebra lineal son útiles, especialmente para aquellos involucrados en la creación de programas complejos. Finalmente, la estadística y la teoría de probabilidades son áreas matemáticas que ofrecen una gran ayuda al científico en este campo. Un estudio en lingüística computacional generalmente concluye con una tesis que implica el desarrollo de software para el procesamiento del lenguaje. Dado que casi ningún programa complejo es realizado por un solo individuo, y es esencial comprender la organización del trabajo entre varios programadores, se recomienda llevar a cabo proyectos en equipos.

## 5. PROCESAMIENTO DE LENGUAJES NATURALES (PLN)

El Procesamiento de Lenguajes Naturales (PLN) emerge como un campo interdisciplinario en la intersección de la inteligencia artificial y la lingüística computacional. Su objetivo es facultar a las máquinas para comprender, interpretar y generar el lenguaje humano de manera significativa. En las últimas décadas, el PLN ha experimentado avances significativos, impulsados por el auge del aprendizaje automático y la disponibilidad masiva de datos lingüísticos. El reconocimiento del habla ha alcanzado niveles de precisión notables, permitiendo aplicaciones como asistentes virtuales y transcripciones automáticas. Los modelos de lenguaje basados en redes neuronales han mejorado la comprensión contextual, llevando a sistemas capaces de generar texto coherente y relevante.

[9] como el intérprete de Perl en http://www.perl.com y el JAVA TOOLKIT en www.sun.com, o la versión personal del JBUILDER de Borland en http://www.borland.com.

No obstante, a pesar de estos logros, los desafíos en PLN persisten. La ambigüedad inherente al lenguaje humano sigue siendo un obstáculo, ya que una sola frase puede tener múltiples interpretaciones. La comprensión del contexto, esencial para una interpretación precisa, plantea desafíos adicionales, especialmente en conversaciones extensas. La barrera idiomática en la traducción automática sigue siendo un terreno fértil para la mejora, y las expresiones idiomáticas presentan dificultades adicionales. Para abordar estos desafíos, se emplean diversas herramientas y métodos. La tokenización, que divide el texto en unidades más pequeñas, y el etiquetado de partes del discurso (POS Tagging) son fundamentales. El análisis sintáctico, que descompone las oraciones en estructuras gramaticales, contribuye a comprender las relaciones entre las palabras. Los modelos de aprendizaje automático, especialmente las redes neuronales, han demostrado ser eficaces en tareas específicas, desde la traducción automática hasta la generación de texto.

El futuro del PLN se vislumbra emocionante. La atención se centra en la comprensión más profunda del significado, superando la mera interpretación superficial. La integración de técnicas de aprendizaje profundo, junto con enfoques más avanzados de procesamiento contextual, promete abordar la ambigüedad y mejorar la comprensión del lenguaje en su totalidad. Además, la ética en el uso de estas tecnologías se vuelve crucial, y los investigadores están trabajando en enfoques responsables y transparentes.

**5.1. Pioneros del PLN**: En 1950, Alan Turing publicó un artículo titulado "Computing Machinery and Intelligence" ("Máquinas de Cálculo e Inteligencia"), en el que propuso lo que ahora se conoce como la Prueba de Turing. Esta prueba se centra en la capacidad de una máquina para exhibir un comportamiento indistinguible del de un ser humano en una conversación. El trabajo de Turing sentó las bases para la idea de que las máquinas podrían entender y procesar el lenguaje natural de manera similar a los humanos. Sin embargo, es importante destacar que el PLN ha evolucionado enormemente desde los días de Turing, y muchos investigadores y científicos han contribuido significativamente al campo a lo largo de los años. Otros nombres importantes en el desarrollo del PLN incluyen a Claude Shannon, quien trabajó en teoría de la información y comunicación; Noam Chomsky, cuyas contribuciones a la lingüística teórica influyeron en el desarrollo de algoritmos de procesamiento del lenguaje natural; y John McCarthy, un pionero en inteligencia artificial que también contribuyó al PLN.

**5.2. Alan Turing, Enigma y esteganografía:** Según Borowska, A. & Rzeszutko, E. (2014), en su artículo “The cryptanalysis of the enigma cipher. the plugboard and the cryptologic bomb’ La máquina Enigma fue un dispositivo de cifrado utilizado por las fuerzas armadas alemanas durante la Segunda Guerra Mundial. Fue diseñada para codificar mensajes de manera que solo aquellos con la clave correcta pudieran descifrarlos. Así mismo fue inventada por el alemán Arthur Scherbius a principios de la década de 1920. La Enigma utilizaba una serie de rotores que giraban y cambiaban la disposición de las letras en el teclado con cada pulsación, lo que hacía que el cifrado fuera extremadamente complejo.

Los estudios sobre la máquina Enigma son fundamentales para entender la historia de la criptografía y la importancia de la inteligencia en tiempo de guerra. Durante la Segunda Guerra Mundial, los Aliados, en especial los británicos, lograron descifrar los códigos Enigma, un logro crucial que ayudó significativamente en la lucha contra las fuerzas alemanas. El trabajo de Alan Turing y otros criptoanalistas de Bletchley Park, Reino Unido, fue esencial para romper los códigos Enigma. Sus esfuerzos y la máquina "Bomba", un precursor de los primeros ordenadores, fueron vitales para descifrar los mensajes enemigos y contribuir al éxito de los Aliados.

La Máquina Enigma funcionaba de la siguiente manera, en un resumen simplificado:

**Teclado:** El operador ingresaba el mensaje original presionando las teclas del teclado. Cada letra presionada activaba los rotores internos de la máquina.

**Rotores:** La Máquina Enigma tenía rotores, discos giratorios que contenían letras del alfabeto. Cada rotor representaba una sustitución de letras.

**Conexiones eléctricas:** Cuando el operador presionaba una tecla, se cerraba un circuito eléctrico a través de los rotores. Cada rotor tenía conexiones eléctricas que conectaban las letras del teclado con letras diferentes en el panel de lámparas indicadoras.

**Cambio de rotores:** Después de cada letra ingresada, los rotores giraban, cambiando las conexiones eléctricas y alterando el cifrado. Este cambio automático de la configuración de cifrado era una de las características clave de la Máquina Enigma.

**Lámparas indicadoras:** La Máquina Enigma tenía un panel de lámparas que mostraba la letra cifrada correspondiente a la letra ingresada. Esta letra cifrada era la que se transmitía como parte del mensaje cifrado.

**Transmisión:** El operador anotaba la letra cifrada y luego la transmitía como parte del mensaje cifrado.

Para descifrar el mensaje, el receptor utilizaba una Máquina Enigma configurada de la misma manera que la del emisor. Cuando ingresaba la letra cifrada, los rotores en la máquina receptora también giraban automáticamente, restableciendo las conexiones eléctricas y revelando la letra original en el panel de lámparas.

La complejidad del sistema residía en la variabilidad constante de las conexiones eléctricas debido al giro de los rotores, lo que hacía que el cifrado cambiará con cada letra ingresada. Este aspecto complicaba enormemente los esfuerzos para descifrar los mensajes sin conocer la configuración exacta de la Máquina Enigma utilizada por el emisor.

***Ejemplificación de Enigma en lenguajes modernos***

Simplificando el fundamento esteganográfico en Python:

```
import string

class EnigmaRotor:

    def __init__(self, wiring, offset):

        self.wiring = wiring

        self.offset = offset


    def forward(self, letter):

         return self.wiring[(ord(letter) - ord('A') + self.offset)
% 26]
```

```
    def backward(self, letter):
        return chr((self.wiring.index(letter) - self.offset + 26)
% 26 + ord('A'))


class EnigmaMachine:
    def __init__(self, rotors):
        self.rotors = rotors

    def encrypt_letter(self, letter):
        for rotor in self.rotors:
            letter = rotor.forward(letter)
        return letter

    def decrypt_letter(self, letter):
        for rotor in reversed(self.rotors):
            letter = rotor.backward(letter)
        return letter

# Configuración de los rotores
wiring_rotor1 = "EKMFLGDQVZNTOWYHXUSPAIBRCJ"
wiring_rotor2 = "AJDKSIRUXBLHWTMCQGZNPYFVOE"
wiring_rotor3 = "BDFHJLCPRTXVZNYEIWGAKMUSQO"

rotor1 = EnigmaRotor(wiring_rotor1, 0)
rotor2 = EnigmaRotor(wiring_rotor2, 0)
rotor3 = EnigmaRotor(wiring_rotor3, 0)

# Crear una máquina Enigma con tres rotores
```

```
enigma_machine = EnigmaMachine([rotor1, rotor2, rotor3])


# Mensaje a cifrar
message = "HELLO"


# Cifrado del mensaje
encrypted_message = ""
for letter in message:
    encrypted_message += enigma_machine.encrypt_letter(letter)


# Descifrado del mensaje
decrypted_message = ""
for letter in encrypted_message:
    decrypted_message += enigma_machine.decrypt_letter(letter)


# Mostrar resultados
print("Mensaje original:", message)
print("Mensaje cifrado:", encrypted_message)
print("Mensaje descifrado:", decrypted_message)
```

**Explicación de las variables:**

- **EnigmaRotor:** Esta clase representa un rotor de la Máquina Enigma. Cada rotor tiene un cableado (wiring) que determina cómo se mapean las letras al pasar a través de él, y un offset que indica la posición inicial del rotor.

- **forward(letter):** Método para cifrar una letra al pasarla a través del rotor en la dirección hacia adelante.

  **backward(letter):** Método para descifrar una letra al pasarla a través del rotor en la dirección hacia atrás.

- **EnigmaMachine:** Esta clase representa la Máquina Enigma completa, que está compuesta por varios rotores.

- **__init__(rotors):** Inicializa la máquina con una lista de rotores.
- **encrypt_letter(letter)**: Cifra una letra pasándola a través de todos los rotores en orden.
- **decrypt_letter(letter):** Descifra una letra pasándola a través de todos los rotores en orden inverso.
- **Configuración de rotores:** Se definen tres rotores (rotor1, rotor2, rotor3) con sus respectivos cableados y posiciones iniciales.
- **Creación de la máquina Enigma:** Se crea una instancia de la máquina Enigma con los tres rotores.
- **Mensaje a cifrar:** Se define un mensaje de ejemplo ("HELLO"). En lengua inglesa, el español tendría la variable consonántica de /ñ/
- **Cifrado y descifrado del mensaje:** Se recorre cada letra del mensaje original, se cifra utilizando la máquina Enigma y se almacena el resultado en encrypted_message. Luego, se descifra el mensaje cifrado y se almacena en decrypted_message.
- **Mostrar resultados:** Se imprime el mensaje original, el mensaje cifrado y el mensaje descifrado.

**5.3. Matemáticas en Enigma:** Según la descripción propuesta por Carter (2004) Enigma podía establecer cerca de 532,985,208,200,576 combinaciones posibles. El panel frontal de la Enigma tenia 26 tomas dobles (A-Z) donde se insertan cables para conectar pares de letras. Para calcular las posibles conexiones en el panel de enchufes, se consideran tres componentes clave: el número total de cables utilizados, grupos de tomas y las interconexiones entre grupos.

Así mismo la tesis de maestría en Matemáticas de Stephanie Faint apoya el complejo modelo matemático que luego establecería el desarrollo de “Bombe”. Primero, consideremos la selección de tomas. Dado que hay 26 tomas en el panel y cada cable tiene dos extremos, cada cable utiliza dos tomas. Con un total de 13 cables ($0 \leq p \leq 13$), tenemos 26 tomas en total. Dado que debemos elegir cables de los 26 disponibles, el número total de combinaciones posibles de tomas es:

$$({}^{26}C_{2p}) \text{ or } \binom{26}{2p}.$$

Para determinar de cuántas maneras diferentes se pueden conectar estos p cables con las 2p tomas seleccionadas, se sigue un patrón sistemático. El número total de formas posibles en que estos p cables podrían haberse conectado con esas 2p tomas es: $(2p-1)!!$. Por lo tanto, el número total de conexiones diferentes que podría haber realizado una máquina Enigma se expresa como:

$$(2p-1)(2p-3)(2p-5)...(3)(1) = (2p-1)!!$$

$$^{26}C_{2p} \times (2p-1)!! = \frac{26!}{(26-2p)! \times p! \times 2^p}$$

*Proof.* As L.H.S.

$$\begin{aligned}
^{26}C_{2p} \times ((2p-1)!!) &= \frac{26!}{(26-2p)! \times 2p!} \times (2p-1)(2p-3)...(3)(1) \\
&= \frac{26!}{(26-2p)! \times (2p)(2p-2)(2p-4)....(4)(2)} \\
&= \frac{26!}{(26-2p)! \times (2.p).2.(p-1).2.(p-2)...(2.2).(2.1)} \\
&= \frac{26!}{(2p-2p)! \times 2^p.(p-1).(p-2)....2.1} \\
&= \frac{26!}{(26-2p)! \times p! \times 2^p} \\
&= R.H.S
\end{aligned}$$

Así mismo se plantean las siguientes resultados:

| Values of p | Combinations | Values of p | Combinations |
|---|---|---|---|
| 0 | 1 | 7 | 1,305,093,289,500 |
| 1 | 325 | 8 | 10,767,019,638,375 |
| 2 | 44850 | 9 | 53,835,098,191,875 |
| 3 | 3,453,450 | 10 | 150,738,274,937,250 |
| 4 | 164,038,875 | 11 | 205,552,193,096,250 |
| 5 | 5,019,589,575 | 12 | 102,776,096,548,125 |
| 6 | 100,391,791,500 | 13 | 7,905,853,580,625 |

### 5.3.1 Posibles Combinaciones en el Panel de Enchufes

El número total de combinaciones posibles en el panel de enchufes se puede calcular sumando todas las combinaciones para cada valor de p. Esta suma se expresa como.

∑ =01326!(26−2)!×!×2∑ p= 013 (26−2p)!×p!×2 p 26! , dando como resultado un asombroso 532,985,208,200,576 combinaciones posibles.

**5.4 BOMBE:** Welchman en 2005 describe a “Bombe”[10] y su importancia durante la Segunda Guerra Mundial. Los criptologos británicos desarrollaron un dispositivo electromecánico de alta complejidad conocido como la "Bombe". Este ingenioso dispositivo tenía como objetivo principal descifrar mensajes secretos cifrados por la máquina Enigma alemana, un desafío criptográfico formidable de la época. La Marina y el Ejército de los Estados Unidos posteriormente fabricaron sus propias versiones de la máquina, manteniendo la misma especificación funcional pero implementando diseños distintos tanto entre sí como en comparación con las bombas polacas y británicas. La base conceptual de la bomba británica se originó a partir de un dispositivo previo denominado "bomba kryptologiczna" diseñado en Polonia por el criptólogo Marian Rejewski en el Biuro Szyfrów (Cipher Bureau). Rejewski, quien previamente había descifrado mensajes Enigma alemanes durante siete años, influyó en el desarrollo del diseño inicial de la bomba británica en 1939 en la Escuela de Código y Cifrado del Gobierno del Reino Unido (GC&CS) en Bletchley Park, donde Alan Turing desempeñó un papel fundamental en su creación.

### 5.4.1 Explicación código de “Bombe” en python básico

```
  from cryptography.fernet import Fernet

def genera_clave():
    clave = Fernet.generate_key()
    with open("clave.key", "wb") as archivo_clave:
        archivo_clave.write(clave)


def cargar_clave():
    return open("clave.key", "rb").read()


def encriptar(mensaje):
    clave = cargar_clave()
    f = Fernet(clave)
    mensaje_codificado = mensaje.encode()
    mensaje_encriptado = f.encrypt(mensaje_codificado)
    return mensaje_encriptado


def desencriptar(mensaje_encriptado):
    clave = cargar_clave()
```

[10] Simulador Java https://www.codesandciphers.org.uk/virtualbp/tbombe/bombesc.htm

```
    f = Fernet(clave)
    mensaje_codificado = f.decrypt(mensaje_encriptado)
    mensaje = mensaje_codificado.decode()
    return mensaje


genera_clave()
mensaje = "Hola, este es un mensaje secreto"
mensaje_encriptado = encriptar(mensaje)
print(mensaje_encriptado)
mensaje_desencriptado = desencriptar(mensaje_encriptado)
print(mensaje_desencriptado)
```

Este código utiliza la biblioteca “cryptography” para generar una clave de cifrado, encriptar un mensaje de texto simple y luego desencriptarlo. Aquí hay una breve explicación de cada variable y función en el código:

- ***Fernet*** es una clase en la biblioteca “cryptography” que se utiliza para cifrar y descifrar datos.
- ***genera_clave()*** es una función que genera una clave de cifrado y la guarda en un archivo llamado “clave.key”.
- ***cargar_clave()*** es una función que carga la clave de cifrado desde el archivo “clave.key”.
- ***encriptar(mensaje)*** es una función que encripta un mensaje de texto utilizando la clave de cifrado.
- ***desencriptar(mensaje_encriptado)*** es una función que desencripta un mensaje de texto encriptado utilizando la clave de cifrado.
- ***mensaje*** es una variable que contiene el mensaje de texto que se va a encriptar.
- ***mensaje_encriptado*** es una variable que contiene el mensaje de texto encriptado.
- ***mensaje_desencriptado*** es una variable que contiene el mensaje de texto desencriptado.

**5.5. Claude Shannon y su Contribución a la Teoría de la Información y Comunicación:** Claude Shannon, un matemático e ingeniero eléctrico estadounidense, es ampliamente reconocido como uno de los padres de la teoría de la información y comunicación. Sus contribuciones revolucionaron nuestra comprensión de cómo la información se transmite y procesa, sentando las bases para el desarrollo de la informática y las comunicaciones modernas.

### 5.5.1 Obras Destacadas

**"A Mathematical Theory of Communication" (Una Teoría Matemática de la Comunicación, 1948)**: En este artículo seminal, publicado en la revista Bell System Technical Journal, Shannon presentó la teoría de la información. Propuso medidas cuantitativas de la información y la incertidumbre, introduciendo conceptos como entropía y la capacidad de canal. Esta obra ha sido fundamental en el desarrollo de la teoría de la información y ha influido en campos tan diversos como la criptografía y la compresión de

datos.

*(Shannon, C. E. (1948). A Mathematical Theory of Communication. The Bell System Technical Journal, 27(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x)*

**"The Mathematical Theory of Communication" (La Teoría Matemática de la Comunicación, 1949, junto con Warren Weaver):** En colaboración con Weaver, Shannon expandió sus ideas originales en un libro que se convirtió en un texto clásico. Este trabajo presenta la teoría de la información de manera más accesible y detallada, explorando cómo la información se transmite eficientemente en sistemas de comunicación.

*(Shannon, C. E., & Weaver, W. (1949). The Mathematical Theory of Communication. University of Illinois Press.)*

**"Programming a Computer for Playing Chess" (Programando una Computadora para Jugar Ajedrez, 1950):** Shannon no solo se centró en la teoría de la información, sino que también se aventuró en la inteligencia artificial. En este artículo, describió cómo programó una computadora para jugar ajedrez, marcando el inicio de la exploración de la intersección entre la teoría de la información y la inteligencia artificial.

*(Shannon, C. E. (1950). Programming a Computer for Playing Chess. Philosophical Magazine, 41(314), 256–275. https://doi.org/10.1080/14786445008521796)*

El trabajo de Claude Shannon ha influido en innumerables disciplinas, desde las ciencias de la computación hasta las telecomunicaciones. Su enfoque matemático riguroso para entender la información ha sido fundamental para el desarrollo de tecnologías modernas, como la compresión de datos, la teoría de la codificación y la inteligencia artificial. Su legado perdura en cada rincón de nuestra sociedad digital, siendo un faro guía para futuras generaciones de científicos de la información y comunicación.

**5.6 Noam Chomsky: Pionero de la Lingüística Teórica y su Impacto en el Procesamiento del Lenguaje Natural:** Noam Chomsky ha dejado un impacto duradero en el entendimiento de la estructura del lenguaje humano. Su enfoque teórico, conocido como la gramática generativa, ha influido no solo en el campo de la lingüística, sino también en el desarrollo de algoritmos para el procesamiento del lenguaje natural (PLN).**5.6.1 Contribuciones a la Lingüística Teórica:**

**"Syntactic Structures" (Estructuras Sintácticas, 1957):** En esta obra seminal, Chomsky propuso la teoría de la gramática generativa, postulando la existencia de una gramática innata en la mente humana. Argumentó que la capacidad de producir y comprender estructuras gramaticales complejas es inherente al ser humano. Esta teoría revolucionó la lingüística al desplazar el enfoque del comportamiento observable hacia los mecanismos internos que rigen el lenguaje.

*(Chomsky, N. (1957). Syntactic Structures. Mouton & Co.)*

**"Aspects of the Theory of Syntax" (Aspectos de la Teoría de la Sintaxis, 1965):** Aquí, Chomsky desarrolló aún más su teoría, introduciendo conceptos como la gramática transformacional y la noción de profundo y superficial estructuras sintácticas. Estos conceptos han sido fundamentales para entender cómo las reglas gramaticales subyacentes generan las formas superficiales del lenguaje.

*(Chomsky, N. (1965). Aspects of the Theory of Syntax. MIT Press.)*

**5.6.2 Impacto en el Procesamiento del Lenguaje Natural:** Las teorías de Chomsky han influido directamente en el campo emergente del PLN. Aunque inicialmente sus ideas encontraron críticas en el ámbito de la inteligencia artificial, la noción de una gramática innata ha proporcionado fundamentos sólidos para desarrollar algoritmos que imitan procesos cognitivos humanos.

**5.6.3 Complejidades sintácticas de Chomsky:** La relación entre la teoría lingüística de Chomsky y la esteganografía reside en la capacidad de manipular estructuras lingüísticas de manera que la información se oculte de manera efectiva. Al aprovechar las complejidades sintácticas propuestas por Chomsky, es posible insertar información en el texto sin alterar significativamente su apariencia superficial.

***Jerarquía de la gramática**:* Chomsky propuso una jerarquía de gramáticas formales que clasifica los distintos tipos de lenguajes según su complejidad sintáctica. Esta jerarquía incluye gramáticas regulares, gramáticas libres de contexto, gramáticas sensibles al contexto y gramáticas recursivamente enumerables.

***Estructuras sintácticas profundas y superficiales**:* Chomsky distingue entre estructuras sintácticas profundas y superficiales. La estructura profunda representa el significado subyacente de una oración, mientras que la estructura superficial es la forma gramatical que observamos en la superficie.

***Transformaciones sintácticas:*** Chomsky introdujo el concepto de transformaciones sintácticas para explicar cómo una estructura sintáctica profunda se transforma en una estructura superficial. Estas transformaciones permiten explicar la variación en la superficie de las oraciones sin cambiar su significado subyacente.

***Gramática generativa:*** La idea central de Chomsky es que la capacidad de generar oraciones en un lenguaje es innata en los seres humanos. Propuso que existe una gramática universal compartida por todas las lenguas humanas, y que la variación observada entre las lenguas se debe a diferencias en los parámetros específicos que se activan en esa gramática universal.

**5.6.4 Ejemplo de gramática sensible en python**

La esteganografía tiene como hipótesis desarrollar gramáticas sensibles al contexto. Las gramáticas sensibles al contexto son un tipo de gramática formal que se encuentra en un nivel de la jerarquía de Chomsky que es más poderoso que las gramáticas libres de contexto pero menos poderoso que las gramáticas recursivamente enumerables. Estas gramáticas se definen por una cuádrupla $G = (N, \Sigma, P, S)$, donde N es un conjunto finito de símbolos no terminales, $\Sigma$ es un conjunto finito de símbolos terminales, P es un conjunto finito de reglas de producción, y S es el símbolo inicial. La característica clave de las gramáticas sensibles al contexto es que las reglas de producción pueden cambiar un símbolo no terminal A a otra cadena $\beta$ dependiendo del contexto en el que aparece A.

```
import re

class ContextSensitiveGrammar:
    def __init__(self, N, T, P, S):
        self.N = N
        self.T = T
        self.P = P
        self.S = S

    def generate(self, string):
        for rule in self.P:
            left, right = rule.split('->')
            left = left.strip()
            right = right.strip()

            if left in string:
                  context = re.findall(f'(?<={left}).*(?={right})',
string)
                if context:
                      string = string.replace(left + context[0] +
right, right + context[0] + left)

        return string

N = {'A', 'B', 'C'}
T = {'a', 'b'}
P = {'AaB -> BaA', 'BbC -> CbB', 'Ca -> aC'}
S = 'A'

grammar = ContextSensitiveGrammar(N, T, P, S)
string = 'aabbCC'
print(grammar.generate(string))
```

Este código define una clase ContextSensitiveGrammar que toma como entrada los conjuntos de símbolos no terminales N, símbolos terminales T, reglas de producción P y símbolo inicial S. La clase también tiene un método generate que toma una cadena como entrada y devuelve la cadena transformada según las reglas de producción de la gramática sensible al contexto.

En el ejemplo de código, se define una gramática sensible al contexto con los conjuntos de símbolos no terminales N = {'A', 'B', 'C'}, símbolos terminales T = {'a', 'b'}, reglas de producción P = {'AaB -> BaA', 'BbC -> CbB', 'Ca -> aC'} y símbolo inicial S = 'A'. Luego, se crea una instancia de la clase ContextSensitiveGrammar con estos parámetros. Finalmente, se genera una cadena de prueba string = 'aabbCC' y se aplica la gramática sensible al contexto a la cadena utilizando el método generate. El resultado se imprime en la consola.

### 5.6.5 Ejemplo de gramática sensible en Java

```
import java.util.regex.Matcher;
import java.util.regex.Pattern;

public class ContextSensitiveGrammar {
    private final String[] N;
    private final String[] T;
    private final String[] P;
    private final String S;

        public ContextSensitiveGrammar(String[] N, String[] T,
String[] P, String S) {
        this.N = N;
        this.T = T;
        this.P = P;
        this.S = S;
    }

    public String generate(String string) {
        for (String rule : P) {
            String[] parts = rule.split("->");
            String left = parts[0].trim();
            String right = parts[1].trim();

            if (string.contains(left)) {
                 Pattern pattern = Pattern.compile(left + "(.*?)" +
right);
                Matcher matcher = pattern.matcher(string);

                if (matcher.find()) {
                    String context = matcher.group(1);
                         string = string.replace(left + context +
right, right + context + left);
                }
            }
        }

        return string;
    }

    public static void main(String[] args) {
        String[] N = {"A", "B", "C"};
        String[] T = {"a", "b"};
        String[] P = {"AaB -> BaA", "BbC -> CbB", "Ca -> aC"};
        String S = "A";

                        ContextSensitiveGrammar grammar = new
ContextSensitiveGrammar(N, T, P, S);
        String string = "aabbCC";
```

```
System.out.println(grammar.generate(string));
```

La clase ContextSensitiveGrammar tiene un constructor que toma cuatro argumentos: N, T, P y S. N es una matriz de símbolos no terminales, T es una matriz de símbolos terminales, P es una matriz de reglas de producción y S es el símbolo de inicio. El método generate toma una cadena como entrada y aplica las reglas de producción hasta que no se puedan aplicar más reglas. El método main crea una instancia de la clase ContextSensitiveGrammar y llama a su método generate con la cadena "aabbCC".

## 6. ENTROPÍA Y ESTEGANOGRAFÍA

La entropía es un concepto clave en el campo de la informática. Según Softtek, compañía de IA en México, se refiere a la cantidad y calidad de la información aleatoria que utiliza un sistema operativo para crear claves criptográficas que sirven para encriptar datos. Cuanta más entropía tenga un sistema, más seguras serán sus claves y más difícil será descifrarlas. En los sistemas informáticos tradicionales, la entropía se puede obtener a partir de fuentes del mundo real, como el movimiento del ratón o el teclado que realizan los usuarios. Sin embargo, estas fuentes se vuelven escasas cuando la informática se traslada del mundo físico a la nube.

Los informes recientes indican que los mecanismos de protección actuales no son suficientes para enfrentar las amenazas existentes. El 66% de las violaciones de seguridad tardan meses o años en ser detectadas, mientras que el 84% de estos ataques se ejecutan en horas o menos. Por eso, es necesario buscar nuevas formas de generar entropía en la nube que garanticen la seguridad de los datos.La relación entre la entropía y la esteganografía se encuentra en el manejo de la información y la capacidad de ocultar mensajes de manera efectiva en un contexto de apariencia aleatoria.En el contexto de la esteganografía, la entropía es crucial. Cuando se oculta información en un medio, como una imagen o un archivo de audio, es esencial mantener la entropía aparente del medio para evitar levantar sospechas. Si la entropía del medio cambia significativamente debido a la inserción de información, podría ser detectada por técnicas de análisis estadístico.

En otras palabras, la esteganografía exitosa implica la capacidad de ocultar información de manera que la entropía del medio aparente se mantenga coherente con su naturaleza original. Esto puede implicar la manipulación cuidadosa de bits o la introducción de información de manera que no cause un aumento evidente en la complejidad del medio. Los algoritmos esteganográficos se diseñan teniendo en cuenta la entropía para garantizar que la información oculta no sea detectable fácilmente mediante técnicas de análisis estadístico o visual.

En el estudio propuesto por Venhuizen, N.J.; Crocker, M.W.; Brouwer, H. Semantic Entropy en Language Comprehension. Entropy (2019). En este artículo, titulado "Semantic Entropy in Language Comprehension", los autores exploran el procesamiento del lenguaje desde una perspectiva de entropía semántica. Argumentan que el procesamiento del lenguaje, palabra por palabra, se ve afectado por la experiencia lingüística previa y el conocimiento general del mundo. Se proponen las métricas de "surprisal" y "entropy reduction" como teorías vinculantes entre la dificultad de procesamiento de palabras y los modelos lingüísticos probabilísticos. El estudio se

basa en un modelo de comprensión que integra la experiencia lingüística y el conocimiento del mundo, proponiendo una métrica de "entropy reduction" centrada en la comprensión. Los resultados sugieren una clara disociación entre "surprisal" y "entropy reduction", reflejando diferentes aspectos del proceso cognitivo de la comprensión. El artículo presenta una perspectiva novedosa sobre la informatividad de las palabras, considerando tanto la experiencia lingüística como el conocimiento del mundo en la comprensión en línea.

| | $p_1$ | $p_2$ | $p_3$ | ... | $p_n$ |
|---|---|---|---|---|---|
| $M_1$ | 1 | 0 | 0 | ... | 1 |
| $M_2$ | 0 | 1 | 1 | ... | 1 |
| $M_3$ | 1 | 1 | 0 | ... | 0 |
| ... | . | . | . | ... | . |
| $M_m$ | 0 | 1 | 0 | ... | 0 |

**Ejemplo de un espacio de significado $\mathcal{SM}\times\mathcal{P}$:**

donde $\mathcal{M}=\{M1,\ldots,Mm\}$

define el conjunto de modelos y $\mathcal{P}=\{p1,\ldots,pn\}$

el conjunto de proposiciones. Las filas representan modelos como combinaciones de proposiciones y las columnas representan vectores de significado que se derivan de este espacio, de modo que: $\vec{v}$ $i(pj)=1$

si y solo si $Mi\models pj$.

Dado un conjunto bien definido de modelos $\mathcal{M}$ y proposiciones $\mathcal{P}$ (es decir, $\mathcal{P}$ describe completamente el conjunto de proposiciones que se pueden capturar en $\mathcal{M}$), el espacio vectorial resultante $\mathcal{SM}\times\mathcal{P}$ ofrece representaciones distribuidas que son composicionales y probabilísticas. Comenzando por lo primero, el espacio de significado $\mathcal{SM}\times\mathcal{P}$ no solo permite derivar los vectores de significado de proposiciones individuales en $\mathcal{P}$, sino también de sus combinaciones. Es decir, dada una definición de negación y conjunción sobre vectores de significado, se puede definir el significado de cualquier combinación lógica de proposiciones en el espacio semántico. El vector de significado $\vec{v}(p)$ de una proposición $p\in\mathcal{P}$ define sus valores de verdad en relación con $\mathcal{M}$, lo que

significa que podemos definir su negación $\vec{v}\ (\neg p)$ como el vector que asigna 0 a todos los $M \in \mathcal{M}$ tales que $p$ se satisface en $M$ y 1 en caso contrario:

$$\vec{v}\ i(\neg p)=1 \text{ si y solo si } Mi \not\models p \text{ para } 1 \leq i \leq |\mathcal{M}|.$$

La definición de la conjunción $p \wedge q$ , dadas $p,q \in \mathcal{P}$, se define como el vector $\vec{v}\ (p \wedge q)$ que asigna 1 a todos los $M \in \mathcal{M}$ tales que $M$ satisface tanto $p$ como $q$ y 0 en caso contrario:

$$\vec{v}\ i(p \wedge q)=1 \text{ si y solo si } Mi \models p \text{ y } Mi \models q \text{ para } 1 \leq i \leq |\mathcal{M}|.$$

La naturaleza probabilística del espacio de significado $\mathcal{SM} \times \mathcal{P}$ se deriva del hecho de que los vectores de significado para proposiciones individuales en $\mathcal{SM} \times \mathcal{P}$ codifican inherentemente su probabilidad. Dado un conjunto de modelos $\mathcal{M}$ que refleja la naturaleza probabilística del mundo, la probabilidad de cualquier fórmula $\varphi$ se puede definir por el número de modelos que satisfacen $\varphi$ , dividido por el total de modelos:

$$P(\varphi)=|\{M \in \mathcal{M} | M \models \varphi\}|/|\mathcal{M}|.$$

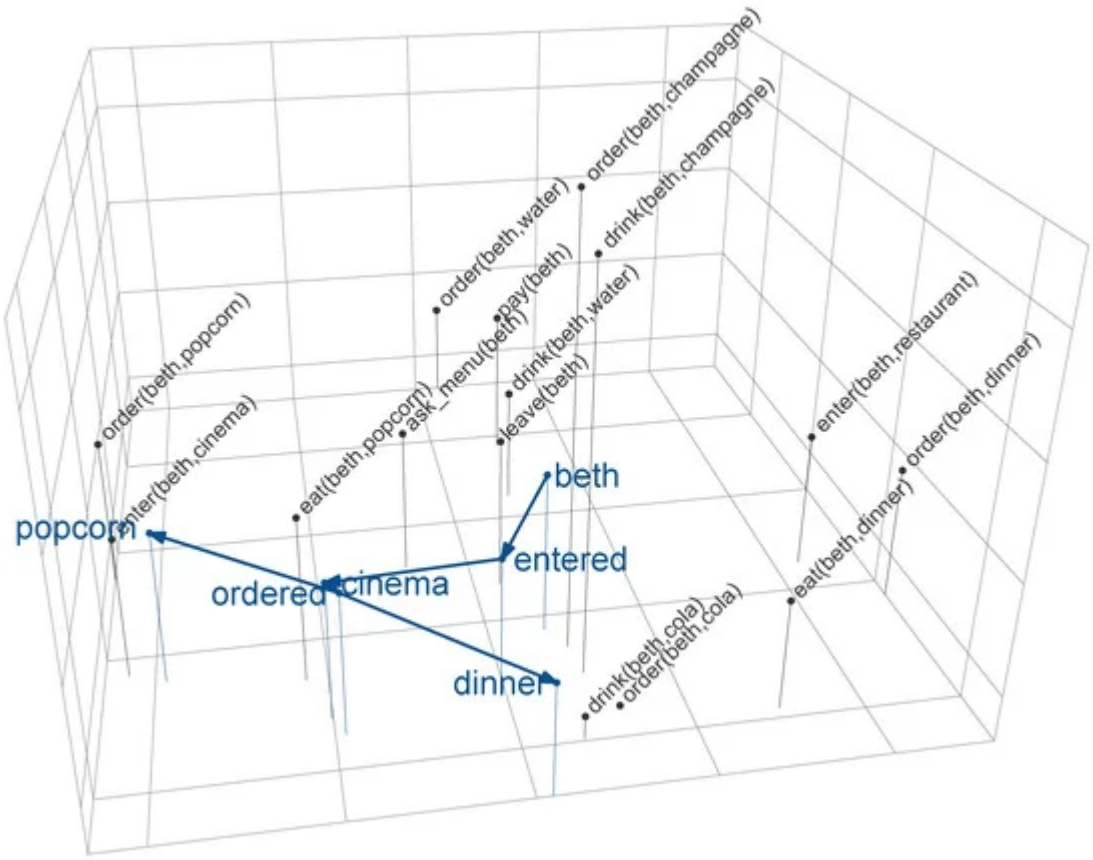
Así, las proposiciones (o combinaciones lógicas de proposiciones) que son verdaderas en un conjunto grande de modelos obtendrán una alta probabilidad y viceversa. Dado que esto permite la definición directa de la probabilidad conjuntiva de dos fórmulas, también podemos definir la probabilidad condicional de cualquier fórmula $\psi$ dada $\varphi$:

$$P(\psi|\varphi)=P(\varphi \wedge \psi)/P(\varphi).$$

Para obtener estimaciones de probabilidad sensatas sobre la co-ocurrencia proposicional en $\mathcal{SM} \times \mathcal{P}$ , el conjunto de modelos $\mathcal{M}$ debe reflejar la estructura probabilística del mundo real con respecto a las condiciones de verdad y co-ocurrencia de cada proposición $p \in \mathcal{P}$ . Llegar a dicho conjunto de modelos $\mathcal{M}$ es un ejercicio no trivial. Una estrategia posible sería deducir el espacio de significado a partir de corpora anotados o bases de conocimientos con inferencias basadas en el conocimiento del mundo o a partir de datos humanos recopilados por la multitud sobre la co-ocurrencia proposicional.

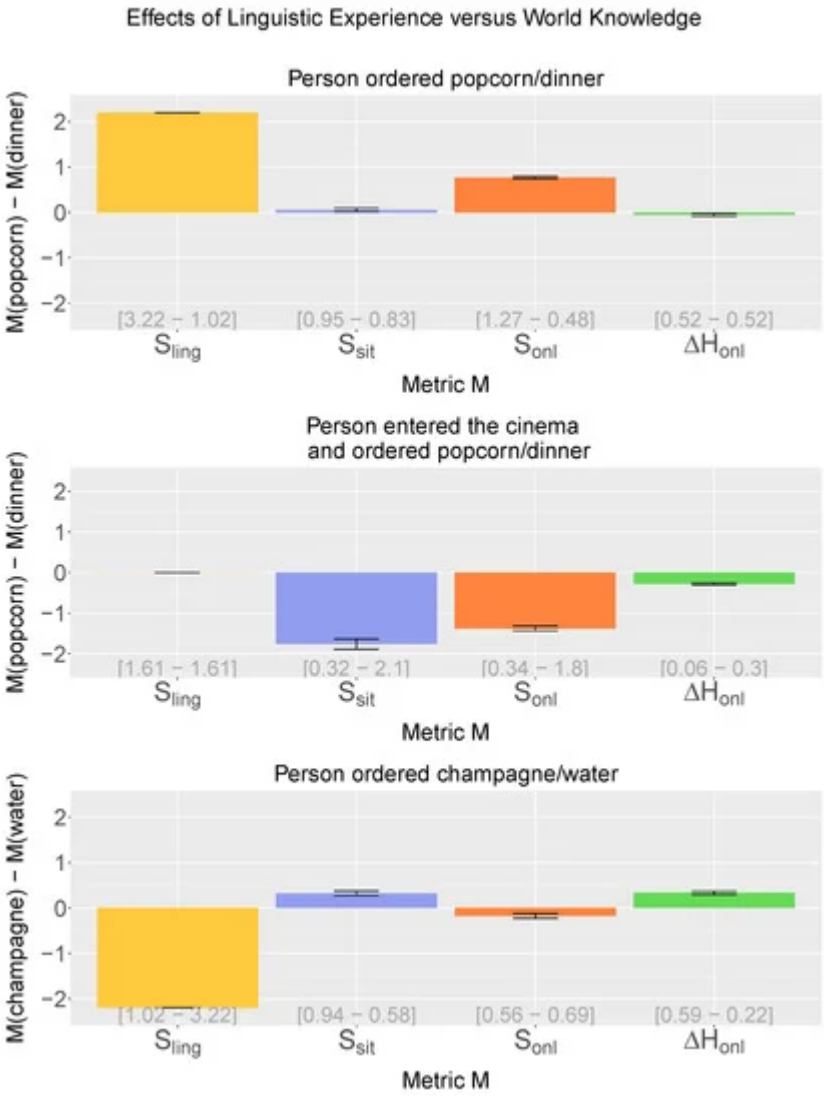
Basándose en su entrada lingüística, el modelo de comprensión construye incrementalmente un vector de significado en su capa de salida que captura el significado de la oración hasta ese momento; en otras palabras, el modelo navega eficazmente por el espacio de significado $\mathcal{SM} \times \mathcal{P}$ palabra por palabra. Es decir, cada palabra entrante $wt$ induce una transición desde un punto en el espacio de significado, $\vec{v}\ t-1$ , hacia el siguiente, $\vec{v}\ t$.

**Ejemplificación del estudio:** Visualización tridimensional del espacio de significado (mediante escalamiento multidimensional) para un subconjunto de las proposiciones atómicas (aquellas relacionadas con Beth). Puntos y flechas de colores muestran la trayectoria de navegación palabra por palabra del modelo para la oración "Beth entró al cine y ordenó [palomitas/cena]" (se omiten las palabras funcionales).



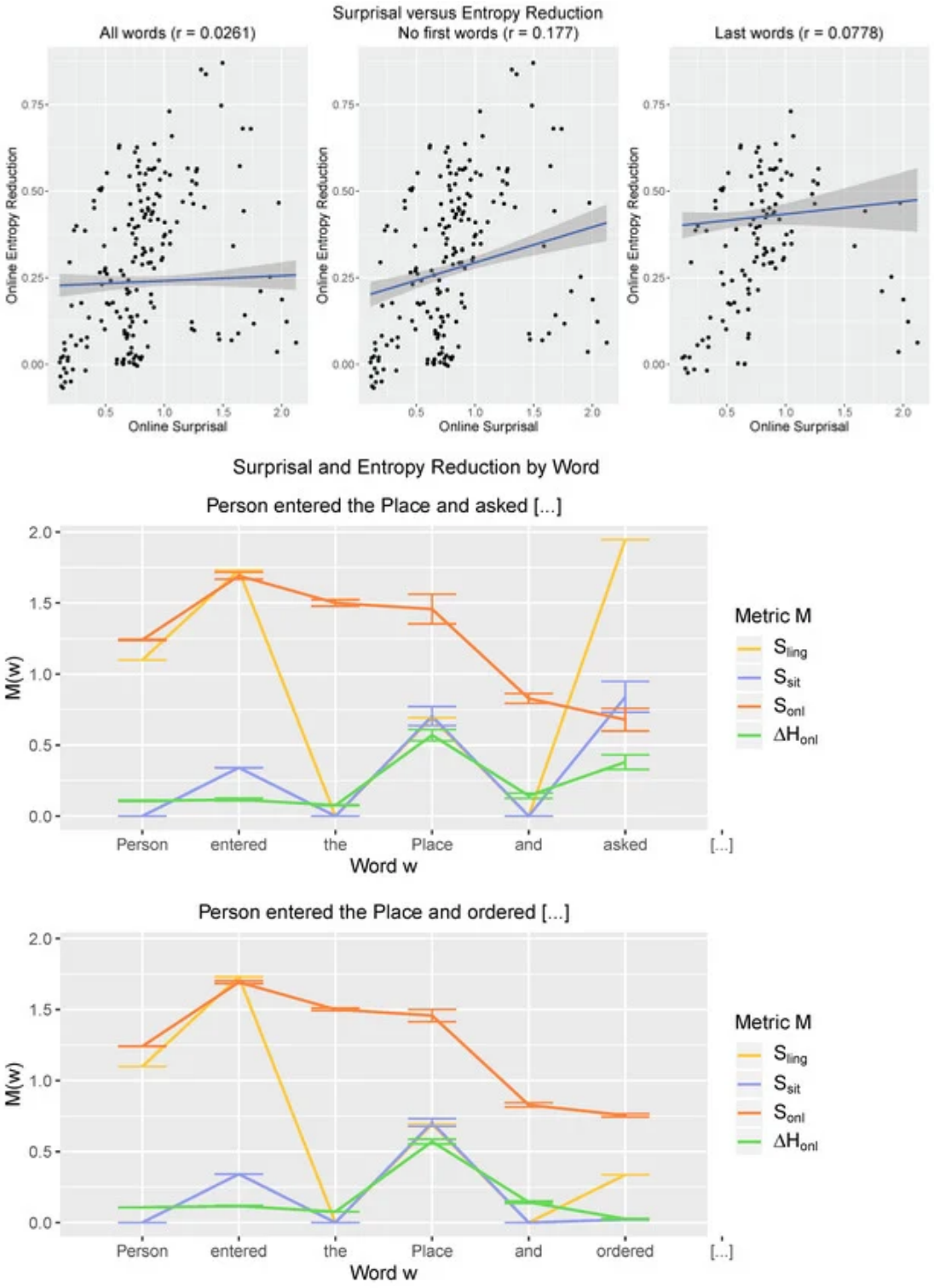

Muestra 1. Entropía lingüística

**Efectos de la Experiencia Lingüística frente al Conocimiento del Mundo:** El modelo de comprensión de VCB asigna oraciones a sus correspondientes vectores de significado probabilístico de manera incremental, palabra por palabra. Esencialmente, el modelo se expone más frecuentemente a ciertos pares de oraciones-semánticas durante el entrenamiento, dando forma a su experiencia lingüística. Además, dado que cada vector de significado lleva inherentemente su propia probabilidad en el espacio de significado, ciertas oraciones pueden asignarse a significados más probables que otros, proporcionando al modelo conocimiento del mundo. Estas fuentes individuales de conocimiento, que influyen en el comportamiento del modelo, se pueden cuantificar de manera independiente en los datos de entrenamiento utilizando la sorpresa. La experiencia lingüística a la que se expone el modelo se puede cuantificar utilizando la sorpresa lingüística fuera de línea, que se estima de manera sencilla a partir de las oraciones con las que se entrena el modelo.




Muestra 2. Entropía lingüística

Comparación entre las estimaciones en línea de reducción de entropía y sorpresa. Los gráficos de dispersión muestran la correlación entre las estimaciones de sorpresa y reducción de entropía para todas las palabras (izquierda), excluyendo las primeras palabras (centro) y solo para las últimas palabras (derecha). Las líneas azules sólidas representan las regresiones lineales correspondientes con sus intervalos de confianza del 95%. El coeficiente de correlación de Pearson (r) se muestra en la parte superior de cada gráfico.




Muestra 3. Entropía lingüística.

En conclusión, la comprensión de la entropía en el contexto de la lingüística computacional se revela como un elemento crucial para delinear futuras líneas de investigación. La relación entre sorpresa y reducción de entropía, especialmente en el marco de modelos de procesamiento del lenguaje, ofrece una perspectiva única sobre cómo las estructuras lingüísticas se desarrollan y se interpretan cognitivamente. Este entendimiento profundo no solo contribuye a la mejora de los modelos de lenguaje probabilísticos, sino que también proporciona valiosas percepciones sobre cómo los conocimientos lingüísticos y del mundo influyen en la dificultad de procesamiento. Así, explorar la entropía en la lingüística computacional se presenta como un camino esencial para avanzar en nuestra comprensión de la intersección entre el lenguaje, la cognición y la informática, abriendo nuevas oportunidades para investigaciones innovadoras y aplicaciones prácticas en este campo en constante evolución.

## 7. CRIPTOGRAFÍA CUÁNTICA Y LINGÜÍSTICA COMPUTACIONAL

En el panorama en constante evolución de las comunicaciones seguras, la criptografía cuántica ha surgido como una tecnología revolucionaria, captando la atención a escala mundial. El reciente compromiso de la Unión Europea de destinar 13 millones de dólares a la investigación y desarrollo de SECOQC (Comunicación Segura basada en Criptografía Cuántica)[11] subraya la importancia estratégica de este sistema de vanguardia. Posicionada como defensa contra el sistema de recopilación de inteligencia Echelon empleado por las principales naciones, la criptografía cuántica ofrece una solución única y robusta a los retos que plantean los criptosistemas tradicionales. Las motivaciones para invertir en criptografía cuántica, a pesar de la eficacia de los criptosistemas modernos considerados "intratables", plantean preguntas intrigantes. Esta exploración profundiza en las limitaciones de los enfoques criptográficos actuales, arrojando luz sobre las vulnerabilidades que impulsan la búsqueda de soluciones criptográficas cuánticas.

A simple vista, la criptografía y la lingüística pueden parecer campos totalmente distintos, pero al profundizar, se revela una conexión intrigante entre ambas disciplinas. Ambas se centran en el manejo de información y en cómo las estructuras subyacentes afectan el sistema en su totalidad. Así como la física cuántica utiliza herramientas matemáticas para comprender fenómenos complejos, podemos explorar cómo la criptografía y la lingüística se entrelazan en el razonamiento sobre la información.La criptografía de clave pública, aunque es una piedra angular en la seguridad de las comunicaciones, se enfrenta a limitaciones inherentes. Los intrincados cálculos que implican los esquemas de clave pública, como RSA y Diffie-Hellman, dan lugar a procesos más lentos, lo que los hace más adecuados para el intercambio de claves que para cifrar grandes volúmenes de datos. Para superar esto, las instituciones suelen adoptar un enfoque híbrido, combinando la velocidad de los sistemas de clave simétrica con la seguridad inicial de los sistemas de clave pública.

[11] Kalaivani, V. (2023). Enhanced BB84 quantum cryptography protocol for secure communication in wireless body sensor networks for medical applications. Personal and Ubiquitous Computing, 27(3), 875.

En el ámbito de la criptografía, consideremos el uso de algoritmos basados en teoría de números para cifrar mensajes. Aquí, las propiedades matemáticas subyacentes se convierten en la base para crear sistemas seguros de comunicación. De manera similar, en lingüística, el análisis gramatical y las estructuras sintácticas sirven como "códigos" para transmitir significado. La forma en que organizamos las palabras en una oración puede afectar drásticamente la interpretación final. Tomemos como ejemplo la aplicación de métodos criptográficos en la seguridad de la información lingüística. ¿Cómo podríamos garantizar la confidencialidad de una conversación textual? Aquí es donde la criptografía entra en juego, desarrollando técnicas para ocultar el significado real detrás de capas de complejidad lingüística. Podríamos imaginar un sistema que codifica significados específicos mediante la alteración de la estructura sintáctica, convirtiendo un mensaje claro en algo aparentemente aleatorio.

A su vez, la lingüística también puede influir en la criptografía, especialmente en la creación de algoritmos de generación de claves. La forma en que construimos frases y expresamos ideas puede inspirar métodos para generar secuencias aparentemente caóticas pero que, en realidad, siguen patrones lingüísticos específicos. Esto podría proporcionar una capa adicional de complejidad a los algoritmos criptográficos, dificultando aún más su descifrado. En términos de teorías interesantes y descriptivas, podríamos explorar cómo la gramática universal propuesta por Noam Chomsky podría influir en el diseño de algoritmos criptográficos. Sin embargo, la seguridad de los criptosistemas de clave pública más utilizados se basa en suposiciones y no en pruebas matemáticas concretas. Algoritmos como RSA y Diffie-Hellman se basan en la supuesta intratabilidad de la factorización de grandes números enteros, un proceso que, por ahora, sigue estando fuera del alcance de la potencia de cálculo contemporánea. La confianza en estos algoritmos se basa en años de escrutinio público, pero los riesgos se hacen patentes ante los avances tecnológicos, especialmente en computación cuántica.

Las vulnerabilidades de los criptosistemas de clave pública actuales se manifiestan en dos áreas clave: la amenaza potencial que supone la computación cuántica y la incertidumbre que rodea a la existencia de futuros teoremas de factorización. El auge de la computación cuántica, con su capacidad de procesamiento sin precedentes, plantea un desafío directo a la seguridad de sistemas como RSA. Además, la falta de una prueba definitiva que afirme la imposibilidad de desarrollar un teorema de factorización rápido introduce un elemento de incertidumbre. Esta incertidumbre se convierte en un riesgo potencial para la seguridad nacional y la propiedad intelectual, que exigen una seguridad sin concesiones.

**Secuencia Milanova, un “Enfoque Lingüístico Cuántico” para Codificar y Decodificar Información:** Este estudio propuesto por Pavel Cherkashin (2023) describe la Secuencia Milanova, creada por Denitsa Milanova e inspirada por Elena Milanova, básicamente es una estructura matemática revolucionaria que transforma elementos lingüísticos en estados cuánticos mediante la proporción áurea, la secuencia de Fibonacci y números surreales, dando lugar al Lenguaje Fractal Universal (LFU).La Secuencia Milanova es un sistema matemático innovador que, utilizando dos polinomios y un tensor, puede representar desde palabras individuales hasta páginas completas. Este concepto va más allá de lo matemático, abriendo aplicaciones potenciales en campos como la física cuántica, la química, la lingüística y la metafísica.La Secuencia Milanova

se forma mediante dos polinomios, f(x) y g(x), con coeficientes de números surreales, conectados a una matriz mediante un tensor, operación repetida para su construcción.

La asignación de significados se realiza a través de la repetición infinita de números naturales, por ejemplo, (12345678910111213...). Cada par de coeficientes se asigna a un cuadrante específico de la matriz, leída fila por fila, generando la Secuencia Milanova generalizada.

**Codificación de Elementos Lingüísticos en un Estado Cuántico:** La Secuencia Milanova codifica y decodifica elementos lingüísticos en estados cuánticos mediante cadenas binarias, utilizando la proporción áurea y la secuencia de Fibonacci.

La secuencia se caracteriza por un patrón específico, definido por la proporción áurea ($\phi \approx 1.618$), el n-ésimo número de Fibonacci ($Fn = Fn-1 + Fn-2$), y la fórmula de la cadena binaria ($Bn = \lfloor \phi Fn \rfloor 2/Fn$).

Ejemplo de las primeras 10 cadenas binarias:

B0 = 0

B1 = 1

B2 = 10

B3 = 101

B4 = 1000

B5 = 10001

B6 = 100010

B7 = 1000101

B8 = 10001000

B9 = 100010001

La Secuencia Milanova ofrece un enfoque completo y único para la transformación de componentes lingüísticos en estados cuánticos, avanzando hacia un futuro donde las máquinas comprendan y se comuniquen de manera natural. Este avance clave en lingüística cuántica y procesamiento del lenguaje natural abre nuevas posibilidades para la investigación y desarrollo, especialmente en el cambiante ámbito de la computación cuántica.La lingüística cuántica explora la comunicación desde la perspectiva del sistema nervioso humano y su naturaleza cuántica, fundamentada en la teoría de que nuestro sistema nervioso se basa en un holograma. Este holograma, un dispositivo cuántico, genera imágenes y proyecciones a través de nuestras conversaciones internas. Similar a la física cuántica, que aborda un campo cuántico como un lugar de múltiples posibilidades, la lingüística cuántica considera una matriz de opciones lingüísticas. Nuestros pensamientos y palabras, regidos por principios matemáticos, definen nuestra mente. Al

operar desde un campo cuántico, se generan numerosas posibilidades alternativas. La lingüística cuántica, vinculada a la Programación Neurolingüística (PNL), impacta el sistema nervioso humano, creando otra matriz con variadas posibilidades de comunicación. La elección consciente de palabras permite construir una atmósfera de cambio y mejorar la calidad de vida. La importancia de la lingüística cuántica radica en su capacidad para explorar científicamente la naturaleza lingüística, ofreciendo una perspectiva más profunda del lenguaje y la comunicación.

Utilizar la lingüística cuántica implica producir una comunicación más efectiva y un mejor uso del lenguaje, lo que puede transformar positivamente la vida cotidiana. Ayuda a comprender el modo de comunicación en diversas situaciones, permitiendo cambios y mejoras. Es una herramienta poderosa para resolver problemas, al identificar los componentes clave de las declaraciones y analizar posibilidades. La lingüística cuántica, al igual que la PNL, ofrece una comprensión profunda de los problemas, facilitando la construcción de relaciones positivas y la resolución efectiva de conflictos.En la práctica, la lingüística cuántica implica la disección de enunciados para identificar el núcleo del problema. Evalúa cómo se puede aplicar en la vida cotidiana, permitiendo una comprensión más profunda del lenguaje utilizado y su impacto en la comunicación. Al enfrentar problemas, se aprende a disecar oraciones en estados cuánticos, considerando todas las posibilidades y abordando la matriz de opciones lingüísticas. La lingüística cuántica es una herramienta transformadora que permite cambiar la perspectiva, comprender mejor la mente y las emociones, y contribuir al bienestar personal y colectivo.

## 8. **SOBRE ALGUNAS CONSIDERACIONES FILOSÓFICAS**

Partiendo del trabajo de  la Universidad Nacional de Colombia “ Aproximaciones filosóficas y sintácticas al significado desde la metafísica, la teoría gramatical y lógica” (2021). El autor, Daniel Rivera, emprende un viaje filosófico a través de las complejidades del lenguaje y su relación con la inteligencia artificial. Su interés radica en explorar alternativas a los enfoques estadísticos dominantes en la lingüística computacional, centrándose en el aspecto "esencial" del lenguaje que impulsa la capacidad creativa y la construcción de significado. Rivera plantea la conjetura de que comprender el propósito fundamental del lenguaje, es decir, el entendimiento y la generación irrestricta de mundos posibles, podría ofrecer perspectivas prometedoras para la inteligencia artificial. Sin embargo, se encuentra con un obstáculo aparentemente insuperable: la encarnación del qualia[12] en una máquina.

El "problema difícil", como lo llama Rivera, implica la dificultad de dotar a una inteligencia artificial no solo con capacidad de ejercicio autónomo, sino también con la capacidad de experimentar, sentir dolor y deseo de manera genuina. Si bien los mecanismos de procesamiento de información son relativamente claros (el "problema fácil"), la verdadera incógnita radica en cómo

[12] El término "qualia" se refiere a las experiencias subjetivas, individuales y conscientes asociadas con la percepción. Está relacionado con la calidad única de cómo experimentamos diferentes sensaciones, como el color rojo, el sabor dulce o el dolor. Los qualia son aspectos subjetivos de la experiencia que no pueden ser plenamente compartidos o comunicados con precisión a través del lenguaje. La noción de qualia es central en debates filosóficos sobre la naturaleza de la mente, la conciencia y la experiencia subjetiva.

incorporar experiencias concretas, el misterio del qualia. Rivera sugiere que la frontera del qualia no debe considerarse como una condena para proyectos de inteligencia artificial y lingüística computacional. Más bien, plantea que esta frontera presenta nuevos desafíos y oportunidades, proporcionando una nueva luz que ilumina de manera diferente el terreno conocido.

En la segunda parte del texto, se adentra en la gramática categorial como un formalismo que clasifica expresiones mediante tipos, ofreciendo una perspectiva lógica para entender la relación entre sintaxis y semántica en lenguas naturales. Destaca las ventajas de este enfoque, como la fidelidad al espíritu original de la propuesta, su carácter lexicalista y la conexión con la lógica lineal. Rivera enfatiza la importancia del manejo de recursos en la teoría lingüística y cómo la gramática categorial, al incorporar la lógica lineal, captura la idea del control de recursos. Además, señala que el isomorfismo Curry-Howard proporciona una caracterización computacional de la semántica directamente visible en la sintaxis.

La integración sintaxis-semántica propuesta no niega la tesis inicial del autor, que la sintaxis es computación y la semántica es experiencia. Más bien, sugiere que, aunque la computación aborda gran parte del procesamiento de información, existe un límite que se encuentra en la experiencia subjetiva, fenomenológica e intencional del significado. Por ello se explora la intersección entre el lenguaje, la inteligencia artificial y la filosofía, abordando desafíos fundamentales y proponiendo enfoques lógicos para comprender la dinámica del lenguaje.

La integración de la esteganografía en el marco propuesto por Daniel Rivera abre la puerta a una forma ingeniosa de manipular el lenguaje y ocultar capas adicionales de información dentro de textos aparentemente normales. En lugar de depender únicamente de métodos convencionales de codificación o cifrado, la esteganografía lingüística podría aprovechar las características intrínsecas de la lógica y la estructura sintáctica del lenguaje. Al ocultar información específica en los entresijos de la gramática y la semántica, se crea un método de comunicación sutil y no evidente. Esto implica que el significado real o la información clave se enmascaran de manera efectiva entre las capas superficiales del texto, ofreciendo una forma de comunicación segura y discreta. La lógica lineal, al proporcionar un marco preciso para la manipulación de recursos y la representación de relaciones sintácticas, se convierte en una herramienta valiosa para estructurar la información oculta. En este sentido, la esteganografía lingüística no solo se convierte en una técnica de ocultamiento, sino también en una manifestación de cómo la estructura profunda del lenguaje puede ser aprovechada de manera creativa para propósitos comunicativos más allá de la expresión superficial.

**8.1 Aspectos éticos y legales de la esteganografía:** En Colombia, el derecho a la intimidad está protegido por la Constitución Política de 1991 que establece:

> *"Artículo 15: Todas las personas tienen derecho a su intimidad personal y familiar y a su buen nombre, y el Estado debe respetarlos y hacerlos respetar. De igual modo, tienen derecho a conocer, actualizar y rectificar las informaciones que se hayan recogido sobre ellas en bancos de datos y en archivos de entidades públicas y privadas."*

Este artículo consagra el derecho fundamental a la intimidad, reconociendo la importancia de proteger la esfera personal y familiar de los individuos, así como su buen nombre. La garantía incluye el derecho a controlar y corregir la información recopilada sobre ellos en bases de datos y archivos, lo que refleja la relevancia que se otorga a la autodeterminación informativa de los ciudadanos. La esteganografía, como técnica para ocultar información dentro de otros datos aparentemente inofensivos, ha sido objeto de creciente interés en el ámbito digital. Aunque puede tener aplicaciones legítimas, su uso ético ha sido objeto de debate, especialmente en contextos donde se han documentado abusos y violaciones a la privacidad.

Desde una perspectiva legal en Colombia y Latinoamérica, la esteganografía plantea desafíos significativos en términos de protección de la privacidad y los derechos individuales. Colombia, en particular, ha sido testigo de casos alarmantes de violación de la privacidad, como las "chuzadas", en las cuales se emplearon tácticas de espionaje masivo que comprometieron la intimidad de ciudadanos, periodistas y defensores de derechos humanos. En Colombia, la Constitución garantiza el derecho a la intimidad y la protección de datos personales. Sin embargo, la implementación efectiva de estas protecciones enfrenta obstáculos, como se evidenció en los casos de "chuzadas" donde agencias gubernamentales se involucraron en la interceptación ilegal de comunicaciones y el uso de esteganografía para ocultar sus actividades. Estos incidentes resaltan la necesidad urgente de fortalecer las leyes y regulaciones que salvaguardan la privacidad en la era digital.

A nivel latinoamericano, la falta de armonización en las leyes de privacidad y la ausencia de marcos regulatorios sólidos permiten lagunas legales que pueden ser explotadas mediante el uso indebido de la esteganografía. La necesidad de una colaboración regional para abordar estos problemas y establecer estándares comunes se hace evidente, ya que las tecnologías digitales no conocen fronteras y las amenazas a la privacidad pueden trascender las jurisdicciones nacionales. El uso ético de la esteganografía implica la necesidad de equilibrar la seguridad nacional con la protección de los derechos individuales. La transparencia en las prácticas de vigilancia, controles judiciales efectivos y la rendición de cuentas son elementos clave para garantizar un uso ético de esta tecnología. Los casos de "chuzadas" en Colombia subrayan la importancia de establecer salvaguardias sólidas y mecanismos de supervisión independientes para prevenir abusos futuros.

**8.1.1 Antecedentes de “Chuzadas” por parte del DAS:** Como se establece en la tesis de Díaz (2022) de la universidad Javeriana el Departamento Administrativo de Seguridad (DAS) fue creado y adscrito al ejecutivo en 1960, con el propósito de mantener el orden y la seguridad de la Nación. Sin embargo, a lo largo de su existencia, se vio envuelto en diversos escándalos, como el asesinato del exlíder del M-19 Carlos Pizarro (1990), el homicidio del periodista Jaime Garzón (1999) y su participación en los falsos positivos (2008) (García, 2014).

Durante la década del 90, cuando el narcotráfico declaró la guerra al Estado, se estableció una conexión directa entre el Cartel de Cali y el DAS. "Los Rodríguez Orejuela además tenían una línea directa con el DAS y la Policía, entidades con las que se comunicaban bajo el seudónimo de los Canarios y a las que entregaban información sobre Escobar y sus aliados" (CEV, 2022, p. 256). Estos vínculos propiciaron la creación de alianzas perversas como los llamados Pepes. "Los

organismos de seguridad, en particular el DAS y la Policía, se unieron a los enemigos de Escobar y, de esta manera, los convirtieron en un aparato paramilitar sin precedentes en Colombia" (p. 264).

En el gobierno de César Gaviria, el DAS experimentó varias reformas destinadas a fortalecer la inteligencia civil y separarla de la inteligencia militar. Se nombró a un civil en su dirección con el objetivo de "darle un nuevo aire a la entidad" (CEV, 2022, p. 272). A pesar de estos esfuerzos, los vínculos entre paramilitares y el DAS persistieron y fueron documentados por investigaciones periodísticas y documentos judiciales en la década de los 90. Salvatore Mancuso afirmó en una entrevista recogida por la CEV (2022): "el Estado sabía lo que estábamos haciendo, íbamos de la mano de ellos, con la bendición de ellos…Entonces las autodefensas realizaban acciones que el Ejército, Policía, DAS no podían ejecutar" (p. 311).

Es importante señalar que, durante los gobiernos de Álvaro Uribe Vélez y la implementación de una política contrainsurgente, se iniciaron campañas de estigmatización y hostigamiento contra opositores políticos. "Uribe utilizó una narrativa en la que la tarea de acabar con las insurgencias incumbía a toda la sociedad en un juego de suma cero: quien no está del lado del Estado, está en el bando enemigo" (p. 425). Así, las organizaciones de inteligencia civiles y militares, centradas en la lucha antiterrorista y en la identificación de los "enemigos" del Estado, llevaron al DAS a consolidarse en torno a la Política de Seguridad Democrática. "Alcanzó su máximo nivel de corrupción con el espionaje ilegal y el hostigamiento a miembros de la Corte Suprema de Justicia que investigaban a los políticos de la coalición de gobierno involucrados con paramilitares" (CEV, 2022, p. 426).

Y es que este trabajo nos da indicios de los pocos estándares éticos que tienen las instituciones del Gobierno Colombiano y en las cuales debe hacerse una profunda reforma que contribuya a la seguridad de los ciudadanos. Sin embargo, la ambivalencia de establecer métodos seguros de detección de atentados y acciones delictivas criminales ponen en jaque sin duda a nivel judicial y político la forma en que se deben establecer estos límites.  La nueva tecnología militar debe establecer parámetros en una sociedad multi conectada y multi comunicada en medio de la saturación de contenido y potenciales “portadores criminales” pero así mismo rastros invisibles para interceptaciones ilegales.

# 9. TÉCNICAS DE ESTEGANOGRAFÍA

**Técnicas aplicadas a la lingüística computacional y esteganografía:** La lingüística computacional, un campo que fusiona lingüística y ciencias de la computación, se sumerge en el análisis y procesamiento de lenguajes naturales mediante algoritmos y modelos. Avances recientes han llevado a la creación de sistemas capaces de entender contextos, emociones e incluso matices culturales en el lenguaje. Estas herramientas no sólo simplifican la traducción automática, sino que también abren la puerta a la comprensión profunda de la comunicación humana en el ciberespacio.

La fusión de estas disciplinas apunta hacia un nuevo horizonte en la seguridad de la información. Imaginen un sistema que no solo "traduce" lenguas, sino que también identifique automáticamente mensajes codificados en el lenguaje natural. La capacidad de reconocer patrones lingüísticos que esquivan las técnicas tradicionales de detección se vuelve crucial en un mundo donde la información sensible está en constante riesgo.Las aplicaciones prácticas de esta convergencia son vastas. En el ámbito de la seguridad nacional, la capacidad de descifrar mensajes ocultos en el flujo de información en línea podría ser la clave para anticipar amenazas.

A nivel empresarial, la protección de datos confidenciales se refuerza al integrar técnicas esteganográficas en la comunicación digital cotidiana. En la era digital, la esteganografía emerge como un arte esencial de ocultar información. Este estudio se sumerge en las técnicas empleadas por potencias clave: Estados Unidos, China, Rusia e Israel. A medida que la información se convierte en el activo más valioso, comprender cómo estos actores estratégicos utilizan la esteganografía no solo revela sus capacidades tecnológicas, sino también su impacto en la seguridad global. Este análisis no solo desentraña el sigilo detrás de estas prácticas, sino que también arroja luz sobre los desafíos cruciales de la ciberseguridad en un mundo cada vez más interconectado.

## 9.1 ESTUDIOS RELACIONADOS EN ESTADOS UNIDOS

**9.1.1 Manuscrito Voynich:** Sin duda uno de los misterios más interesantes en la lingüística es el que se relaciona con el Manuscrito Voynich. La obra enigmática es un volumen ilustrado de contenido desconocido, ostenta una relevancia histórica sustancial. Adscrito a un autor anónimo y plasmado en un alfabeto inidentificable junto a una lengua incomprensible denominado voynichés, el manuscrito ha capturado la atención de eruditos y entusiastas por igual. Nombrado en honor al comerciante de libros antiguos Wilfrid M. Voynich, quien lo adquirió en 1912, el manuscrito se encuentra resguardado actualmente en la Biblioteca Beinecke de libros raros y manuscritos de la Universidad de Yale bajo la designación MS 408.

El proceso de datación mediante carbono 14 ha revelado que el pergamino que alberga este enigma literario fue manufacturado entre los años 1404 y 1438. El análisis estilístico, abordando tanto la escritura como las ilustraciones, ha confirmado su origen en el siglo XV, con indicios que apuntan hacia Europa Central, posiblemente Alemania o el norte de Italia. La historia de su posesión por el emperador Rodolfo II del Sacro Imperio, aunque carece de registros concretos, sugiere que el monarca invirtió 600 ducados de oro en la adquisición, siendo posteriormente custodiado por su

consejero, el médico y farmacéutico Jacobus de Tepenec, cuya firma sutilmente se distingue en el primer folio. Tras pasar por diversas manos, la obra encontró su lugar en la biblioteca del Colegio Romano antes de ser adquirida nuevamente por Wilfrid Voynich en 1912.

En la actualidad, el manuscrito Voynich consta de aproximadamente 240 páginas, algunas de ellas desplegables y de diversas dimensiones. Sus páginas albergan ilustraciones que abarcan desde representaciones de plantas y hierbas hasta contenedores farmacéuticos, diagramas astronómicos y zodiacales, y sistemas inusuales de tuberías y bañeras poblados por figuras femeninas desnudas. El "idioma" voynichés, sujeto a la Ley de Zipf, presenta una entropía inusualmente baja, revelando regularidades que hacen que la combinación de caracteres sea altamente predecible.

A pesar de los esfuerzos de criptógrafos profesionales, lingüistas y entusiastas, incluidos descifradores de códigos de las guerras mundiales, el manuscrito Voynich persiste en resistir cualquier intento de desciframiento. Las hipótesis abordadas, que van desde un idioma desconocido hasta cifrados sofisticados y engaños sin sentido, han quedado sin verificar de manera independiente. Este enigma continúa siendo objeto de estudio y especulación, estimulando la imaginación popular y perpetuando el misterio que envuelve su significado y origen.

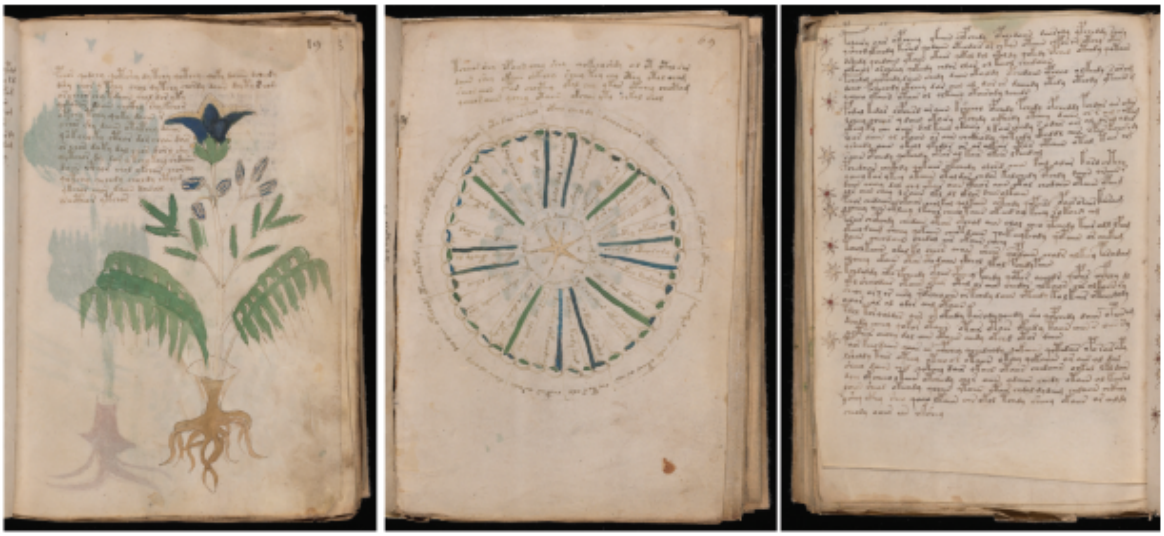


Figura 1
Folios 19r, 69v y 105 v del manuscrito, ilustrando los temas relacionados con la herbología, la alquimia y la astrología. Imágenes cortesía de la Biblioteca de Libros y Manuscritos Raros Beinecke, Universidad de Yale.

Bowern, C. L., & Lindemann, L. (2021). The linguistics of the Voynich manuscript. *Annual Review of Linguistics*, *7*, 285-308.

**Estudios Universidad de Yale, departamento de lingüística**

El estudio de Bowern, C. L., y Lindemann, L. (2021), realiza un análisis detallado de las dificultades lingüísticas para comprender este tipo de código. Inicialmente, este trabajo busca establecer una línea de trabajo sobre cómo este código, de no ser un "engaño" de un lenguaje natural, sería la ejemplificación de la capacidad de ocultamiento que tendría la esteganografía.

### 9.1.2. Lengua Natural o Mentira elaborada

La primera cuestión que debemos abordar es si la "lengua" del Manuscrito Voynich es realmente un lenguaje natural. Es decir, podría ser un disparate de algún tipo, un engaño diseñado para parecer un cifrado. Los defensores de la hipótesis del disparate señalan la regularidad anormal y el alto grado de similitud en algunas palabras adyacentes, lo que lo hace diferente de muchos otros tipos de texto (cf. Barlow 1986). El estatus lingüístico del Manuscrito Voynich ha sido objeto de controversia. Mientras autores como Reddy & Knight (2011) y Hauer & Kondrak (2016) han trabajado desde principios de desciframiento, otros han afirmado demostrar que el lenguaje del Voynich no es en absoluto un lenguaje. Rugg (2004) y Rugg & Taylor (2016) sugieren que el lenguaje podría haber sido generado por técnicas criptográficas del siglo XV, que podrían producir texto cifrado o disparates. Rugg (2004) sugiere que el manuscrito puede haber sido codificado con la llamada técnica de rejilla de Cardan, mientras que Rugg & Taylor (2016) sugieren que el material es más compatible con un engaño que con un lenguaje cifrado. Daruka (2020) llega a la misma conclusión de que el Manuscrito Voynich es un engaño y contiene disparates, aunque creado por medios diferentes a los sugeridos por Rugg. Sin embargo, como muestran Landini (2001), Montemurro et al. (2013), Reddy & Knight (2011) y Sterneck & Bowern (2020), la explicación del disparate no explica la estructura documental de nivel superior, que discutimos más a fondo en la Sección 4. Es decir, es poco plausible que un documento falso basado en disparates tenga una estructura interna de este tipo, y es más probable que el Manuscrito Voynich sea un cifrado u otra codificación de un lenguaje natural.

En esta etapa, no estamos convencidos por ninguno de los argumentos a favor de que el texto Voynich sea un disparate. Dado que el disparate es, por naturaleza, aleatorio, no debería mostrar ninguna de las propiedades organizativas de nivel superior que muestra el Manuscrito Voynich (como se resume en las Secciones 3.3 y 4). El Manuscrito Voynich es altamente inusual y no se asemeja a un lenguaje a nivel de caracteres. Si miramos más allá de las palabras individuales hasta los niveles de línea y párrafo, así como la distribución de palabras en todo el manuscrito, el texto del manuscrito parece un lenguaje natural codificado en lugar de un disparate. Dado que las medidas utilizadas para rastrear la estructura del párrafo es poco probable que sean manipuladas directamente, son un buen indicador de una estructura real. Si descartamos el disparate, la pregunta entonces se convierte en qué tipo de codificación representa el sistema de escritura del Voynich. Aunque esto aún es desconocido, algunos tipos de códigos se pueden descartar. Por ejemplo, probablemente no sea un cifrado de sustitución simple porque, de ser así, probablemente se habría descifrado a estas alturas. Por otro lado, los cifrados polialfabéticos son poco probables tanto porque probablemente no se usaron en el siglo XV como porque un cifrado polialfabético llevaría a que palabras idénticas se codificaran de manera diferente en diferentes partes del manuscrito. No veríamos las mismas (o similares) palabras siendo las más frecuentes en cada página (consulte la Sección 3).

Finalmente, el manuscrito podría estar en un lenguaje construido (conlang). Hasta donde sabemos, el conlang premoderno más extenso es el lenguaje Balaibalan, basado en turco, persa y árabe (Häberl 2015, Koç 2005). Balaibalan, un lenguaje esotérico sufí, está representado en tres manuscritos que datan aproximadamente de 1580, pero probablemente fueron esfuerzos colectivos de composición a lo largo de muchos años. Hay otros dos conlangs antiguos bien atestiguados: el Lingua Ignota, basado en latín y alemán, creado por Hildegard von Bingen (Higley 2007) en el siglo XII; y el Enochian, un lenguaje "angelical" similar al inglés inventado en el siglo XVI por Edward Kelly y John Dee (Laycock 2001). Los tres conlangs se basan en gran medida en lenguajes

naturales y consisten en raíces inventadas incrustadas en la morfología y sintaxis de lenguajes naturales. Por lo tanto, si el Voynichés es un conlang, podríamos esperar que presente un patrón morfosintáctico con otros lenguajes naturales pero que sea anómalo a nivel de raíces.

### 9.1.3 . Antecedentes del Manuscrito

En la discusión que sigue, los autores asumen algunos puntos que aún son objeto de debate en los estudios del Voynich, considerados poco productivos. En primer lugar, basándose en las fechas de carbono del pergamino (así como en su apariencia general y las opiniones de expertos en estudios medievales), se asume que el manuscrito es un objeto medieval genuino (Clemens 2016, Clement 1997) y no una falsificación moderna. Aquellos que argumentan que el manuscrito es una falsificación moderna deben asumir que Voynich (u otra persona) obtuvo una gran cantidad de pergamino medieval intacto e hizo tinta altamente consistente con las prácticas medievales, en una época antes de que se desarrollaran métodos para datar con precisión el pergamino. Es decir, deben asumir anacrónicamente que un falsificador moderno intentaba evitar la detección al eludir pruebas que aún no se habían inventado.

También se considera poco probable que el Manuscrito Voynich sea una falsificación medieval. El costo asociado con la producción de un manuscrito así y la cantidad de personas involucradas hacen poco probable que se haya creado puramente para engañar. Un engaño mucho más pequeño habría servido al mismo propósito con mucho menos gasto. Además, quienes asumen que el manuscrito es un engaño medieval subestiman masivamente la cantidad de esfuerzo requerido para producir un sinsentido sostenido similar al lenguaje. El códice físico data del período 1404-1438 (Clemens 2016) según la datación por carbono, pero no se sabe si el códice físico existente fue copiado de alguna fuente anterior. Por esta razón, no se asume que el lenguaje debe ser medieval. Siguiendo a Davis (2020), se asumen múltiples escribas. Davis aporta pruebas de la forma de los glifos que indican que participaron al menos cuatro (y probablemente cinco) manos diferentes en la producción del manuscrito.

Siguiendo la observación inicial de Currier (1976) y el trabajo posterior de Davis (2020) y Reddy & Knight (2011) (entre otros), se asume que hay dos "lenguas" en el manuscrito, etiquetados por Bowern, C. L., & Lindemann, L. (2021) por conveniencia como Voynich A y Voynich B. Más precisamente, hay dos métodos de codificación de al menos un lenguaje natural. Aunque se utiliza el término lenguaje aquí (siguiendo la convención), no está claro que las diferencias entre Voynich A y Voynich B se deban a diferentes idiomas subyacentes o variedades (aunque esto es posible). Voynich A se utiliza en parte de las secciones herbarias y farmacéuticas, mientras que Voynich B se utiliza en la sección balneológica, en algunos folios de las secciones medicinales y herbarias, y en la sección astrológica. Las principales diferencias entre los dos idiomas están en la frecuencia de las palabras (que se discuten más adelante en la Sección 3). Ciertas secuencias de caracteres son muy comunes en Voynich A (ol ol y or or) y relativamente poco comunes en Voynich B, y viceversa (dy dy). No se aborda extensamente la conexión entre los dos idiomas aquí, pero se tratan las dos variedades por separado con fines analíticos. Cabe destacar que hay isomorfismo entre las manos de Davis y los dos idiomas de Currier, con el escriba 1 escribiendo en Voynich A y los escribas 2, 3, 4 y 5 en Voynich B.

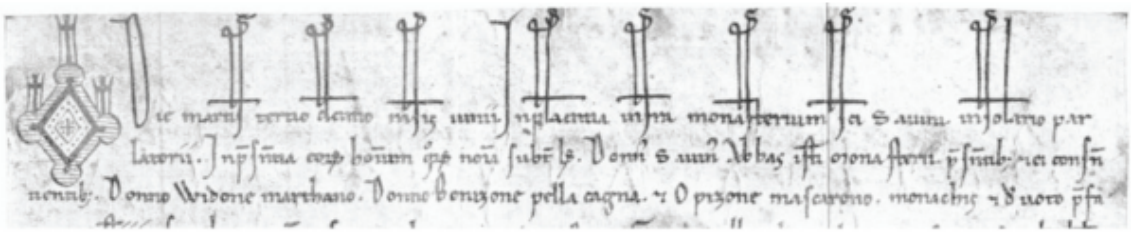
En el trabajo de Bowern, C. L., & Lindemann, L. (2021) sobre la lingüística del Manuscrito Voynich, se aborda la cuestión de la fonología y escritura, centrándose en el escrito Voynichese del manuscrito. Se destaca que el manuscrito no se limita únicamente a Voynichese, ya que presenta otras ortografías, como los nombres de los meses en occitano en las páginas del zodiaco, la frase

parcialmente oscurecida en latín al final del manuscrito y la firma ahora invisible de Jacobus Horcický de Tepenec en la primera página.

Para analizar las propiedades de Voynichese, se utiliza una transcripción convencionalizada 1:1 entre los caracteres Voynich y los caracteres ASCII, conocida como EVA (Alfabeto Voynich Extensible, anteriormente Alfabeto Voynich Europeo). Esta transliteración tiene como objetivo facilitar investigaciones basadas en métodos estadísticos del texto. Los investigadores reconocen que esta correspondencia es una simplificación y que no pueden sacar conclusiones adecuadas sobre la estructura fonológica de Voynichese sin comprender la estructura de la ortografía. Sin embargo, hasta la fecha, la transcripción de Takahashi ofrece la única posibilidad para análisis textuales basados en métodos computacionales, a pesar de las limitaciones que esto implica.

| Transliteration | Glyph | Transliteration | Glyph |
|---|---|---|---|
| ' | | l | |
| a | | m | |
| c | | n | |
| d | | o | |
| e | | p | |
| f | | q | |
| g | | r | |
| h | | s | |
| i | | t | |
| k | | v | |
| x | | y | |
| Combined characters | | | |
| ch | | sh | |
| cfh | | cth | |
| cph | | ckh | |

En cuanto a la escritura, se señala que incluye caracteres latinos de varias tradiciones, como carolingia y beneventana, así como números y caracteres que no tienen contrapartida en otros manuscritos, salvo quizás como adornos ornamentales. Se destaca que algunos caracteres Voynich son claramente parte del alfabeto latino, mientras que otros se asemejan a números y algunos no tienen paralelos en otras escrituras. Se introduce la distinción entre los caracteres "gallows" (horca) y los caracteres "bench" (banco), siendo los primeros llamados así por su apariencia superficial a horcas y los segundos por su semejanza a un banco. La investigación subraya la complejidad de la escritura Voynich y la necesidad de abordar sus particularidades para comprender mejor la naturaleza lingüística del manuscrito.



Además de los caracteres discutidos anteriormente, existen otras formas que aparecen solo una vez, dos veces o unas pocas veces en el manuscrito (ver Davis 2020, p. 170). La más común de estas es el símbolo alquímico x.

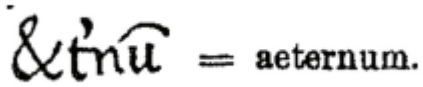
Hay 21 o 22 glifos comunes, además de aproximadamente la misma cantidad de formas de caracteres más raras. El número total de glifos en el conjunto de datos depende en gran medida de cómo se clasifican las diversas formas variables de los grafemas. Por ejemplo, hay dos transliteraciones comunes en uso, EVA y v101, que utilizan caracteres diferentes. Utilizamos EVA en esta revisión porque es la más ampliamente utilizada y es la base de las transliteraciones legibles por máquina en las que se basa nuestro trabajo estadístico en las secciones siguientes.

EVA representa cada letra Voynich con un carácter ASCII. Fue diseñado para producir versiones legibles por máquina del texto y ahora se usa ampliamente. Los mapeos de las letras Voynich a caracteres ASCII se basaron en las suposiciones del criptógrafo seudónimo Glen Claston sobre qué caracteres del alfabeto latino podrían ser más cercanos a los caracteres Voynich (ver http://www.voynich.nu/transcr.html). Por ejemplo, d se translitera en EVA como d. Los equivalentes de transliteración están fuertemente influenciados por consideraciones europeas o latinas. Por ejemplo, la secuencia qo se representa como qo en gran parte porque q nunca aparece en el texto excepto antes de o, un patrón reminiscente de la distribución de q con u en latín. Sin embargo, no hay otra razón independiente para asumir que q en Voynich es la q del alfabeto latino.

En nuestra opinión, EVA probablemente subdiferencia caracteres (agrupando dos variantes de d, por ejemplo). También crea diacríticos a partir de caracteres que podrían entenderse mejor como glifos individuales. EVA ch, por ejemplo, comprende dos componentes distintos: c y h (representados en ASCII como ch), aunque h nunca se encuentra por separado y c es idéntico a e. Sin embargo, EVA es el sistema de transcripción que se utiliza en la versión legible por máquina del Manuscrito Voynich, por lo que lo adoptamos aquí en espera de una revisión exhaustiva del sistema de transcripción (para otros problemas relacionados con la transcripción, consulte la barra lateral titulada Algunas Dificultades en la Creación de una Transcripción de Voynichese).

Dificultades:

Las divisiones de palabras son posiblemente poco confiables.
Se desconocen las relaciones entre los caracteres.
La relación entre los grafemas y la fonología subyacente es desconocida




En cuanto a la presencia de caracteres en el manuscrito, se observa que algunos caracteres aparecen tanto en combinación como de forma independiente en el manuscrito. El carácter de banco (ch) se combina con los caracteres de horca para formar cTh, cKh, cFh y cPh. Los caracteres ch y Sh parecen distinguirse solo por la pluma sobre el banco. Esta ligadura recuerda al carácter utilizado para abreviaturas en muchos manuscritos medievales, como se ilustra en la Figura 3. Esta misma ligadura también puede distinguir entre las letras e y s. Las letras d y g parecen distinguirse por un trazo descendente. Sin embargo, se desconoce si estas similitudes son casuales (comparar con la relación entre las letras latinas o, b y d) o reflejan algún principio subyacente de regularidad en el script.

En cuanto a la transliteración EVA, aunque facilita la pronunciación del material Voynich, carece de fundamentos más allá de cierta similitud con algunas letras del script Voynich. Por ejemplo, transcribir los caracteres de horca como f, k, t y p es simplemente una convención. En la sección de Fonología y Ortografía, se examina el sistema lingüístico subyacente al Manuscrito Voynich. Cabe destacar que esta evaluación no aborda la fonología en el sentido puramente lingüístico, ya que no se tiene conocimiento de la organización abstracta del sonido representado por la ortografía ni de la relación entre la ortografía y la fonología (propiamente dicha). Sin embargo, se pueden extraer algunas conclusiones sobre la representación del sistema lingüístico en comparación con representaciones ortográficas de otros idiomas.

En relación con vocales versus consonantes, tanto Reddy & Knight (2011) como Guy (1991) discuten si el sistema de escritura Voynich es un alfabeto (que contiene tanto vocales como consonantes) o un abjad (que representa solo consonantes). Un test para esta clasificación utiliza el algoritmo de Sukhotin (Guy 1991), que parte de la premisa de que en la mayoría (si no en todos) los idiomas naturales, las vocales tienen más probabilidad de estar adyacentes a consonantes que a otras vocales. Es decir, las sílabas de la forma CV o CVC son más comunes que aquellas de la forma V o VC. El algoritmo calcula una matriz de adyacencia para todos los caracteres, suma sucesivamente las filas, asume que el segmento más frecuente es una vocal y luego elimina el doble del número de ocurrencias de la matriz de adyacencia. Se continúa identificando posibles vocales hasta que no quedan sumas positivas. Para Voynichese en general, el algoritmo de Sukhotin identifica cinco vocales: a, o, e, h e i (es decir, a, o, e, h e i). Vale la pena señalar que tres de estos caracteres (a, o e i) son similares a caracteres que representan vocales en el alfabeto latino. La letra e es ligeramente más común en Voynich B, y la letra h es ligeramente más común en Voynich A.

Algunos problemas adicionales persisten. En primer lugar, el algoritmo de Sukhotin es sensible a si se incluyen o no las divisiones de palabras en los cálculos. Cuando se excluyen las divisiones de palabras, los caracteres identificados como vocales son a, g, e, n, o y y. Es decir, i ya no se identifica como una vocal y se identifican tres caracteres que son casi exclusivamente finales de palabra (g, n e y). En segundo lugar, se identifican caracteres diferentes entre los dos idiomas Voynich, Voynich A y Voynich B.

Reddy & Knight (2011) argumentan que Voynichese muestra más propiedades de un abjad que de un script alfabético. Su argumento se basa en parte en la inducción de clases de caracteres a partir del comportamiento de agrupación de los caracteres. Utilizan un Modelo Oculto de Markov bigrama de dos estados sobre letras en Voynichese e inducen dos clases. Para alfabetos, estas clases suelen corresponder a consonantes y vocales. Para Voynichese, sin embargo, las dos clases corresponden al carácter final en una palabra versus el resto de la palabra. Este resultado podría ser impulsado principalmente por los caracteres que se encuentran solo al final de una palabra (como m, n y g). Es decir, el resultado puede indicar no que el script Voynich sea un abjad, sino más bien que hay variantes posicionales para formas de caracteres. Además, aunque el algoritmo de Sukhotin destaca tanto consonantes como vocales en los abjads, en los scripts abjad no todas las palabras contienen uno de los caracteres vocálicos (aunque todas las palabras contienen un fonema vocal). En cambio, casi todas las palabras en el Manuscrito Voynich tienen al menos uno de los caracteres que el algoritmo de Sukhotin destaca como vocales.

### 9.1.4 Entropía

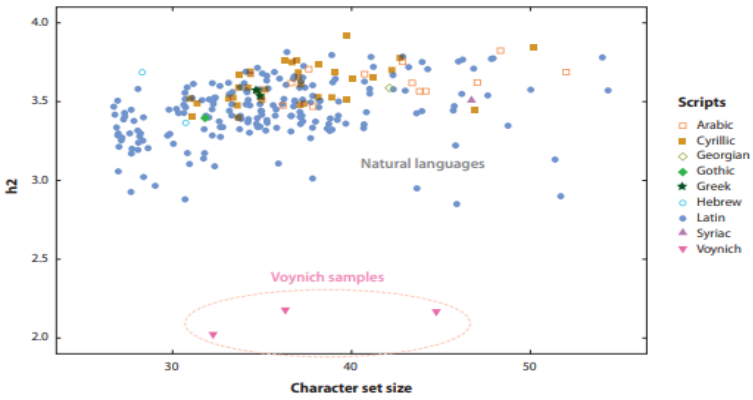
La entropía de caracteres, entendida como la cantidad de imprevisibilidad o desorden presente en un texto, se mide a través de la entropía de caracteres a nivel individual. Esta métrica define la cantidad promedio de información transportada por un solo carácter, generalmente expresada en bits o "shannons". La noción fue introducida por Shannon (1949) y encuentra su aplicación en la teoría de la información, donde es esencial para medir la tasa teórica de transmisión de información. La entropía de caracteres ofrece otra medida para comparar idiomas entre sí y con el Voynichese.

Bennett (1976) destaca la naturaleza inusual del script Voynich en su discusión sobre la entropía condicional de caracteres, también conocida como entropía de segundo orden o h2. La entropía condicional de caracteres se refiere a la predictibilidad global de una letra dada la letra anterior. Por ejemplo, en textos en inglés, la letra q casi invariablemente va seguida de la u. La probabilidad condicional del bigrama qu (la probabilidad de u dado que la letra anterior es q) es cercana a 1. La entropía condicional global se calcula a partir de las probabilidades condicionales de cada bigrama, ponderadas por su ocurrencia general en el texto, como se muestra en la siguiente ecuación:

$$H(X|Y) = \sum_{i,j} P(x_i, y_j) \log_2 \frac{P(y_j)}{P(x_i, y_j)}.$$

En el análisis de Bennett (1976), se compara el Manuscrito Voynich con textos en cuatro lenguas europeas modernas, revelando que la entropía condicional en Voynichese es considerablemente menor. Esto implica que las secuencias de caracteres dentro de las palabras en el texto Voynich son inusualmente predecibles en comparación con las lenguas europeas. Los caracteres de Voynich aparecen en secuencias sorprendentemente predecibles, con ciertos caracteres ubicados al principio o al final de la palabra o solo después de ciertos caracteres específicos. Bennett concluye que la entropía de caracteres en Voynich es comparable a la de lenguas polinesias como el hawaiano, que tienen inventarios de fonemas reducidos y formas de sílabas limitadas. Sin embargo, Stallings (1998) señala que el texto hawaiano utilizado por Bennett emplea una ortografía simplificada que no distingue entre vocales largas o paradas glotales, lo que disminuiría la entropía.

En efecto, la entropía condicional de caracteres en la transcripción EVA del texto Voynich es significativamente más baja que en cualquier otra lengua de la muestra. La Figura 4 ilustra la entropía condicional de caracteres (h2) y el tamaño del conjunto de caracteres para 250 lenguas en la muestra, codificados por tipo de escritura. En la figura, la entropía condicional de caracteres de Voynichese (Voynich A, Voynich B y la muestra completa con caracteres raros incluidos) es mucho más baja que la de cualquier muestra de lengua natural. La entropía de Voynichese difiere de la de cualquier otra lengua o escritura. Se han explorado manipulaciones plausibles del sescrituras, como diversas abreviaturas y eliminación de vocales. Estas manipulaciones afectan la entropía de caracteres, pero no al mismo grado.




Entropía condicional de caracteres (h2) versus número de caracteres para el Voynichese y otras lenguas. Este gráfico se limita a lenguas escritas con alfabetos (latino, cirílico, gótico, georgiano) o abjads (hebreo, árabe, siriaco). Las alfasilabarios como el devanagari tienen un mayor tamaño de conjunto de caracteres pero un h2 similar. Los logogramas como el chino tienen un tamaño de conjunto de caracteres mucho mayor y un h2 mucho más alto. Todas las lenguas con los valores más bajos de h2 están escritas con alfabetos.

### 9.1.5 **Reflexiones sobre el manuscrito Voynich y la esteganografía**

.

La aplicación de métodos esteganográficos en el estudio del enigmático Manuscrito Voynich promete arrojar nuevas perspectivas significativas sobre tanto su contenido como su propósito intrínseco. La hipótesis que planteamos sostiene que este manuscrito, en apariencia indescifrable, podría albergar un sistema esteganográfico sofisticado, indicativo de una capacidad lingüística excepcional por parte de su creador. La comprensión profunda de este potencial sistema esteganográfico no solo se limitaría a descifrar el contenido del manuscrito, sino que podría tener ramificaciones revolucionarias en el ámbito de la lingüística computacional. Postulamos que el análisis detallado del manuscrito, respaldado por comparaciones meticulosas con otros sistemas lingüísticos y esteganográficos conocidos, podría allanar el camino para el desarrollo y mejora de métodos avanzados de codificación y descifrado. Esta investigación no solo desvelaría los secretos del Manuscrito Voynich, sino que también abriría nuevas fronteras en la aplicación práctica de la esteganografía en la lingüística computacional, promoviendo avances significativos en la codificación de información y la seguridad de datos.

## 9.2. **Illegals Program**[13]

El Programa de Ilegales (nombre dado por el Departamento de Justicia de los Estados Unidos) fue una red de agentes rusos durmientes bajo cobertura no oficial. Una investigación del Buró Federal de Investigación (FBI) culminó con el arresto de diez agentes el 27 de junio de 2010, y un intercambio de prisioneros entre Rusia y Estados Unidos el 9 de julio de 2010. Los espías arrestados eran ciudadanos rusos que habían sido infiltrados en los Estados Unidos por el Servicio de Inteligencia Exterior de Rusia (conocido por su abreviatura rusa, SVR), la mayoría de ellos utilizando identidades falsas. Haciéndose pasar por ciudadanos estadounidenses comunes, intentaron establecer contactos con académicos, industriales y responsables políticos para obtener

[13]The "Illegals" Program. # 9". June 28, 2010 recuperado de https://www.justice.gov/sites/default/files/opa/legacy/2010/06/28/062810complaint1.pdf

acceso a información confidencial. Fueron el objetivo de una investigación de varios años por parte del FBI, denominada Operación Cuentos de Fantasmas, que culminó a finales de junio de 2010 con el arresto de diez personas en Estados Unidos y una undécima en Chipre.

Los diez agentes durmientes fueron acusados de "llevar a cabo misiones a largo plazo y encubiertas en los Estados Unidos en nombre de la Federación Rusa". El sospechoso arrestado en Chipre huyó bajo fianza al día siguiente de su detención. Una duodécima persona, un ciudadano ruso que trabajaba para Microsoft, también fue detenida alrededor de la misma época y deportada el 13 de julio de 2010. Documentos judiciales de Moscú hechos públicos el 27 de junio de 2011 revelaron que otros dos agentes rusos, alegadamente conocidos por el FBI, lograron escapar de los Estados Unidos sin ser arrestados. El 9 de julio de 2010, los diez agentes fueron trasladados a Viena después de declararse culpables de no registrarse como representantes de un gobierno extranjero. Ese mismo día, los agentes fueron intercambiados por cuatro ciudadanos rusos, tres de los cuales habían sido condenados e encarcelados por espionaje (alta traición) en Rusia en nombre de Estados Unidos y el Reino Unido. El 31 de octubre de 2011, el FBI hizo públicas varias docenas de imágenes fijas, fragmentos de videos de vigilancia y documentos relacionados con su investigación en respuesta a solicitudes de la Ley de Libertad de Información. Según el archivo oficial del departamento de justicia de Estados Unidos, dentro de la acusación de 18 Páginas se establecen los siguientes métodos esteganográficos.

11. To further the aims of the conspiracy, Moscow Center has arranged for the defendants clandestinely to communicate with the Russian Federation. In particular, the conspirators have used a number of methods of secret communications. See Complaint III.A. As set forth below, these include covert communications by means of private wireless networks.

12. In general terms, covert communication via a private wireless network is a form of electronic communication through paired laptop computers. Such covert communication utilizes temporary wireless networks that spring up between two computers and can be used to transmit data between them. The way this system typically works is as follows: a laptop computer ("LAPTOP A") is pre-configured to create its own private wireless local area network. This wireless network is programmed to only communicate with another specific laptop ("LAPTOP B"), based on LAPTOP B's Media Access Control ("MAC") address.[2] Once LAPTOP A transmits the signal to establish its own private wireless network, it will be "joined" by LAPTOP B when LAPTOP B comes within a certain physical distance of LAPTOP A. Once the two laptop computers are both on the private wireless network, they can communicate with each other by

---

[2] A MAC address is a unique identifying number assigned by a manufacturer to electronic communications devices, such as laptop computers and Wi-Fi network cards. All computers that access the internet through a wireless network do so by publicly broadcasting their MAC addresses via radio transmission in order to locate a device connected to the internet.

4

Adjunto FBI. 1.1 *Archivo departamento de justicia de los Estados Unidos*

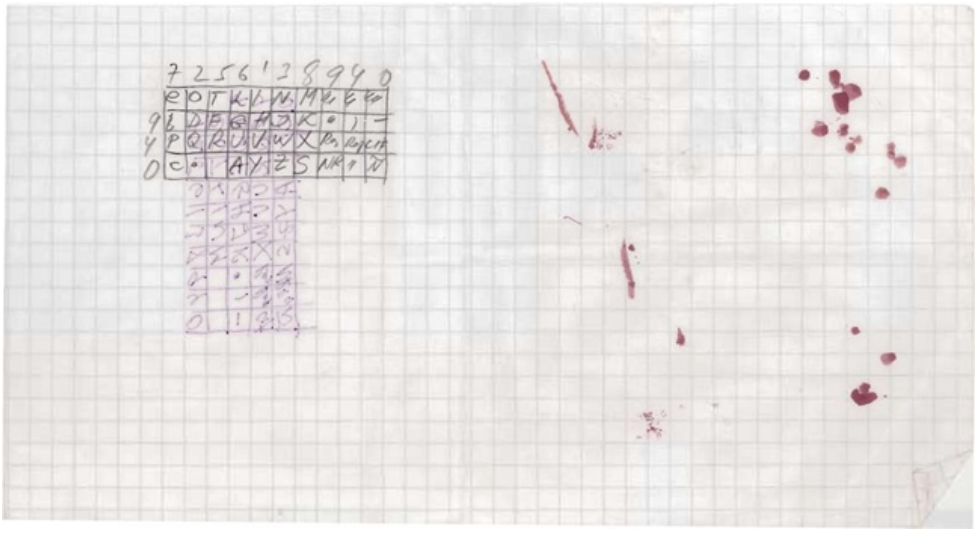
**Esteganografía en "Illegal programs":** Uno de los métodos más sofisticados utilizados fue el empleo de software esteganográfico personalizado desarrollado en Moscú. Este software permitía ocultar mensajes en archivos aparentemente inofensivos. La iniciación de este programa requería la combinación de teclas Control-Alt-E y la introducción de una contraseña de 27 caracteres, detalle que el FBI descubrió al encontrarla anotada. Estos mensajes codificados se camuflaban hábilmente en documentos comunes, proporcionando una capa adicional de secreto a la información intercambiada.

Además de la tecnología, los agentes recurrieron a tácticas físicas como el uso de una radio de onda corta para enviar ráfagas codificadas de datos. Esta forma de comunicación, que se basa en la transmisión de información a través de ondas electromagnéticas, añadía una dimensión analógica a sus estrategias digitales. Según el FBI (2015) En enero de 2010, Anna Chapman, uno de los agentes implicados, llevó a cabo una transferencia electrónica de datos desde su laptop en una cafetería de Nueva York. Dos meses después, utilizó una red Wi-Fi privada, posiblemente en una tienda Barnes & Noble en Greenwich Street, para comunicarse con el mismo oficial ruso que se encontraba cerca. El uso de un extensor de rango para su laptop demostró la adaptabilidad y precaución empleadas en estas operaciones encubiertas. La esteganografía, al fusionar tecnología avanzada con métodos tradicionales, se erigió como la columna vertebral de estas comunicaciones clandestinas, permitiendo a los agentes rusos operar en las sombras y eludir la detección durante un tiempo considerable.

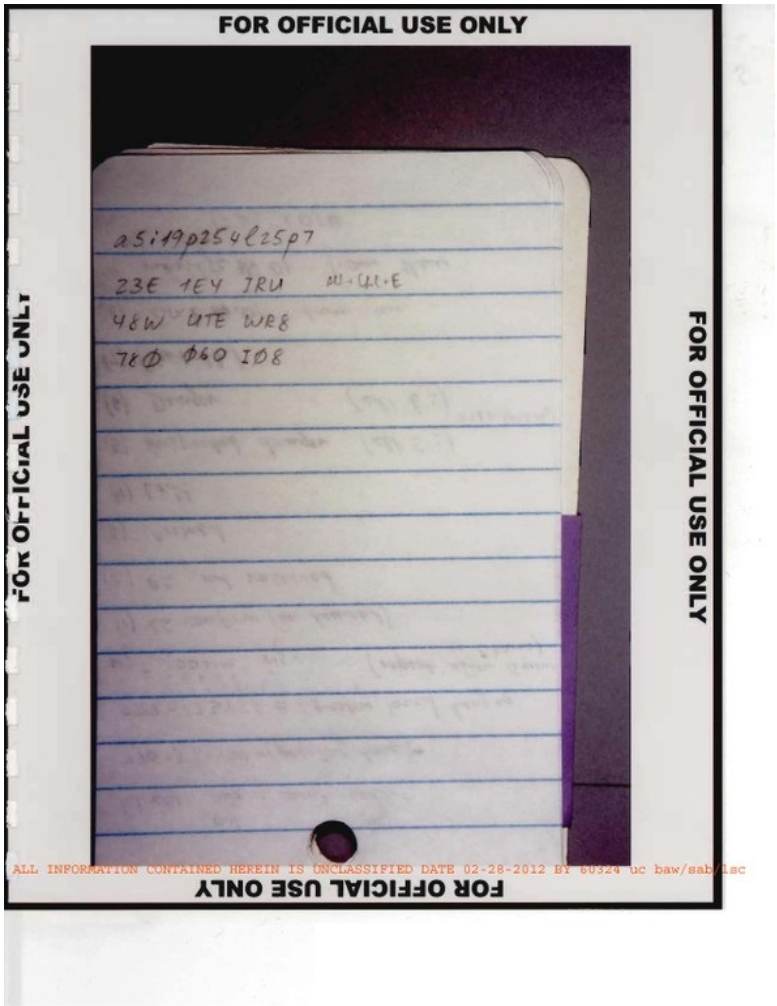
exchanging data. The data can be encrypted so that it can only be read with the aid of specialized decryption software, similar to that used to decrypt messages hidden through steganography, as described in the Complaint attached hereto.

1. ANNA CHAPMAN'S USE OF PRIVATE WIRELESS NETWORKS

13. Since in or about January 2010, law-enforcement agents, acting pursuant to judicial orders, conducted surveillance of ANNA CHAPMAN, the defendant, at various locations in New York City. On approximately ten Wednesdays between January 2010 and June 2010, law-enforcement agents observed CHAPMAN in the physical vicinity of an individual ("Russian Government Official #1") who has, on multiple occasions, been observed entering the Russian Mission to the United Nations in Manhattan. Based on my training, experience, and participation in this investigation, I believe that on each of these ten occasions, CHAPMAN and Russian Government Official #1 covertly exchanged electronic communications via a private wireless network, as described above.



Prueba 1. Caso Lazaro. Departamento de justicia de los Estados Unidos.




Prueba 2. Caso Lazaro. Departamento de Justicia de los Estados Unidos.

La historia de Juan Lázaro, en el segundo caso de la corte, utilizando un código de tablero de ajedrez para cifrar mensajes, y la mención de la red wifi privada de Anna Chapman subrayan la diversidad de estrategias empleadas. Cada espía tenía su propio método, adaptándose a las circunstancias y desafíos específicos de su misión. A pesar de estos "trucos del siglo XXI", la red de espionaje fue descubierta en 2010. Sin embargo, la naturaleza siempre evolutiva de la inteligencia rusa sugiere que más capítulos de esta intrincada saga podrían revelarse en el futuro. Mientras tanto, la historia de estos ilegales del siglo XXI destaca la persistencia y adaptabilidad de las operaciones de espionaje en un mundo cambiante. La evolución de la inteligencia rusa tras la desintegración de la Unión Soviética ha revelado un enfoque constante en la infiltración encubierta, demostrado por la fragmentación de la KGB en el FSB y el SVR, junto con la persistencia del GRU. Contrariamente a las expectativas post Guerra Fría, Rusia continuó sus operaciones de espionaje, dando vida a un nuevo "programa de ilegales" en Estados Unidos.

En este contexto, un "ilegal" se define como un espía con identidad falsa, capacitado meticulosamente por el SVR antes de su despliegue en suelo estadounidense. La formación abarca desde idiomas extranjeros hasta técnicas avanzadas de comunicación, incluyendo la operación de cifrado y onda corta, así como el envío de mensajes codificados en código Morse. Esta red de espionaje, descubierta en 2010, utilizó métodos innovadores y herramientas tecnológicas inusuales. La esteganografía se destacó como un pilar clave en sus comunicaciones. Se empleó software taquigráfico, no disponible comercialmente, que permitía la inserción de datos cifrados en imágenes almacenadas en sitios web públicos. Estos datos eran invisibles a simple vista y solo podían eliminarse mediante un software específico llamado "Stenography". El FBI descubrió este método durante búsquedas, encontrando discos protegidos con contraseña que albergaban tanto el software como los datos necesarios. Además de estas tácticas digitales, los espías rusos recurrieron a métodos analógicos como el uso de radios y "radiogramas", ráfagas codificadas de datos enviadas y recibidas a través de transmisores de radio. La evidencia recopilada durante las operaciones de vigilancia auditiva reveló el uso de estaciones numéricas, una técnica clásica de comunicación encubierta. El caso terminó con el intercambio de presos políticos por parte de los dos países.

## 9.3. ESTUDIOS RELACIONADOS EN RUSIA

Blas (2020) describe el desarrollo tecnológico de la guerra por parte del Kremlin. En el curso de la historia, la guerra ha trascendido tres ámbitos fundamentales: tierra, mar y aire. Sin embargo, a finales del siglo pasado, surgió un cuarto ámbito, el ciberespacio, donde los medios militares tradicionales se sometieron a una transformación radical. Este nuevo tipo de guerra se basa en flujos de información, hardware y software, planteando la pregunta de si representa una esfera novedosa para los conflictos entre naciones. El ciberespacio implica un cambio drástico en la forma de concebir la guerra, dando lugar a estrategias que impactan directamente en la sociedad civil. Mientras algunos estrategas destacan la importancia humana, argumentando que lo cibernético es simplemente una herramienta bajo el control humano, otros sugieren, de manera más fantasiosa, que la rápida evolución tecnológica nos acerca inevitablemente a la creación de entidades como Skynet. La realidad es que las armas cibernéticas son una constante en los conflictos actuales, constituyendo una parte esencial de la estrategia militar estatal.

Estas armas cibernéticas no solo son una nueva forma de conflicto, sino que representan una amenaza significativa para la paz mundial, alterando radicalmente el panorama de la guerra contemporánea. Los ciberataques pueden originarse entre estados, actores privados o incluso de actores privados contra estados, dificultando la identificación de su origen. Este riesgo se

manifiesta en la capacidad de un grupo de adolescentes para paralizar el sistema financiero global desde el desván de su hogar. En el ámbito de la defensa, el ciberespacio plantea una amenaza invisible capaz de afectar instituciones, infraestructuras públicas y compañías privadas fundamentales para la economía de un estado. Para la Unión Europea, que carece de una política exterior o de defensa común, la amenaza de ciberataques es real y palpable. Las intervenciones de Rusia en elecciones de diversos estados amenazan la estabilidad y credibilidad de las democracias occidentales.

Este artículo se adentrará en dos áreas conceptuales cruciales: la guerra cibernética y la estrategia militar rusa en el ciberespacio. Explorará la guerra híbrida, que combina ataques cibernéticos con acciones presenciales, y analizará los motivos y estrategias detrás de los ciberataques rusos en Estonia, Georgia, Ucrania y Lituania. El objetivo es trazar conexiones y responder a las preguntas planteadas en las secciones subsiguientes.

### 9.3.1. Malwares en la ciberguerra Kaspersky[14]

**Monty Three: Espionaje industrial con esteganografía(2020)**

Esta empresa de ciberseguridad describe el estudio de un malware avanzado. En el verano de 2020, Kaspersky hizo el descubrimiento de un conjunto de herramientas C++ de varios módulos previamente desconocido, utilizado en ataques de espionaje industrial altamente dirigidos que datan desde 2018. La atención inicial hacia este malware se debió a su rareza, la clara orientación de la campaña y la falta de similitudes evidentes con campañas conocidas en términos de código, infraestructura o tácticas, técnicas y procedimientos (TTP). Hasta la fecha, se considera que este conjunto de herramientas y el actor detrás de él son nuevos. Los creadores del malware lo denominaron "MT3"; siguiendo esta abreviatura, se ha nombrado al conjunto de herramientas como "MontysThree".

Este malware consta de varios módulos C++ utilizados para la persistencia, la obtención de datos de un mapa de bits mediante esteganografía, el descifrado de tareas de configuración (capturas de pantalla, toma de huellas digitales del objetivo, obtención de archivos, etc.) y su ejecución, así como comunicaciones de red con servicios en la nube importantes como Google, Microsoft y Dropbox. MontysThree está configurado para buscar documentos específicos de Microsoft Office y Adobe Acrobat almacenados en directorios de documentos actuales y en medios extraíbles. Además del cifrado personalizado basado en XOR, el malware utiliza esquemas de cifrado 3DES y RSA para el descifrado de la configuración y las comunicaciones.

Se observan artefactos de lenguaje natural en MontysThree, con configuración de lengua rusa, que buscan directorios presentes solo en versiones localizadas en cirílico de Windows. Aunque la mayoría de las comunicaciones externas en la nube pública utilizan autorización basada en tokens, algunas muestras contienen cuentas de correo electrónico que pretenden ser similares a las chinas, consideradas como señuelos. Estos indicios sugieren que el malware fue desarrollado por un actor de habla rusa y está dirigido específicamente a versiones cirílicas de Windows.

[14] Kaspersky Industrial CyberSecurity (KICS) es una plataforma integral diseñada para salvaguardar las capas de tecnología operativa (TO) en entornos industriales. Ofrece productos integrados y servicios que protegen elementos como sistemas DCS, arquitecturas SCADA, HMI, controladores como PLC e IED, así como redes y estaciones de trabajo. La innovación de Kaspersky radica en su enfoque hacia la convergencia de la seguridad de la información (TI) y la tecnología operativa (TO), asegurando la integración fluida de soluciones empresariales y especializadas en su cartera.

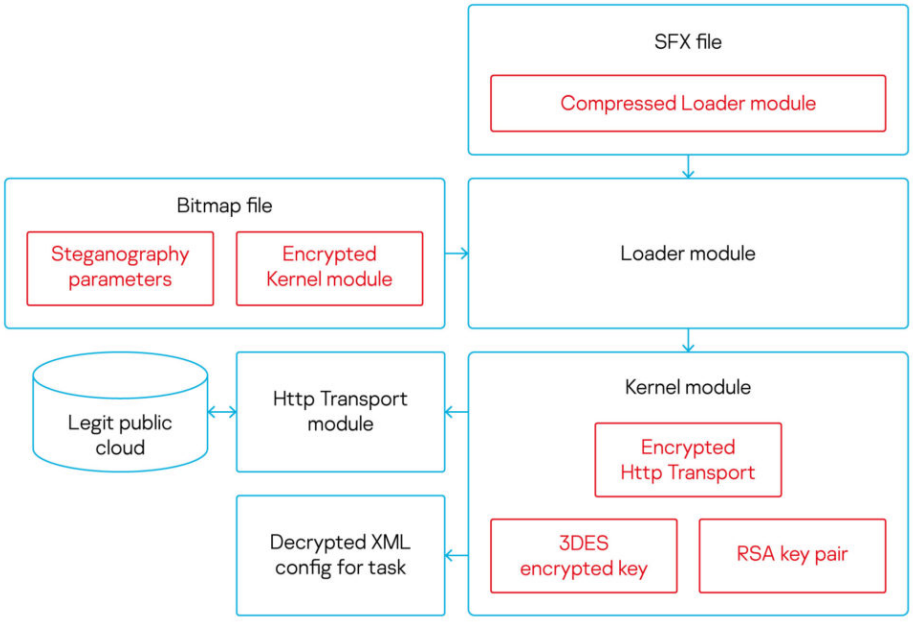
**Funcionamiento del Malware**

Según la descripción de Kaspersky el módulo de carga inicial se distribuye dentro de archivos autoextraíbles (SFX) RAR con nombres relacionados con la lista de teléfonos de los empleados, documentación técnica y resultados de pruebas médicas. No hay señuelos, sólo archivos PE (disfrazados de un archivo .pdf o .doc), pero estos títulos son ahora un truco típico utilizado en el phishing: "actualización de información corporativa" o "resultados de análisis médicos". Uno de los cargadores (MD5 da49fea229dd2dedab2b909f24fb24ab) tiene el nombre "Список телефонов сотрудников 2019.doc" ("Lista de teléfonos de empleados", en ruso). Otros cargadores tienen los nombres "Tech task.pdf" e "invitro-106650152-1.pdf". Este último es el nombre de un laboratorio médico en Rusia. Todos ellos parecen trucos típicos de phishing. El guión SFX es el siguiente:

```
Path=%TEMP%\
SavePath
Setup=rundll32.exe "invitro-106650152-1.pdf",Open
Silent=1
Overwrite=1
Update=U
Delete=invitro-106650152-1.pdf
```

Durante la ejecución, el script SFX llama a la función Open() (volveremos a este nombre exportado) del ejecutable del cargador descomprimido en el directorio %TEMP% y lo elimina. A juzgar por el nombre del archivo, lo más probable es que imite resultados de análisis médicos, dado que "Invitro" es un laboratorio médico destacado en Rusia. Este PE32 inicial es el primer módulo de carga.

***Funcionamiento de los módulos***



El diagrama anterior muestra el flujo de ejecución general de los módulos MontysThree. En la siguiente tabla se enumeran cuatro módulos y sus características. Los módulos comparten convenciones de comunicación comunes. Cuando se trata de datos compartidos, como la configuración y el registro de ejecución detallado, el malware inicializa la estructura en el almacenamiento local de subprocesos (TLS), que a su vez hace referencia a estructuras de montón. Curiosamente, además de la RAM, el registro de ejecución se almacena en el disco en un archivo cifrado con un XOR de un byte.

El punto de entrada DllEntryPoint() funciona como un constructor, que asigna la estructura con TlsAlloc() y la guarda en una variable global. Los módulos deben exportar una función llamada Open(), que no toma parámetros (pero puede analizar la línea de comando) y devuelve un código de error de cuatro bytes.

| Nombre del módulo | Características |
|---|---|
| **Cargador** | Este módulo antidetección está a cargo de la esteganografía personalizada y el descifrado del módulo del kernel. |
| **Núcleo** | Este módulo del kernel (principal) se encarga de descifrar el XML de configuración y luego analizar y ejecutar las tareas correspondientes en él. |

| | |
|---|---|
| **HttpTransporte** | Módulo de red para comunicarse con servicios legítimos de nube pública de Google, Microsoft, Dropbox, así como con fuentes WebDAV. El módulo es capaz de realizar solicitudes a través de RDP y Citrix de forma ingenua utilizando clientes legítimos. |
| **Actualización de enlace** | El módulo de persistencia es un modificador .lnk de inicio rápido de Windows. Con este ingenuo método de persistencia, los usuarios ejecutarían el módulo Loader por sí mismos cada vez junto con los navegadores desde la barra de herramientas de Inicio rápido de Windows. |

Ahora echemos un vistazo a cómo los desarrolladores combinaron fuertes estándares de criptografía modernos con estándares personalizados basados en XOR.

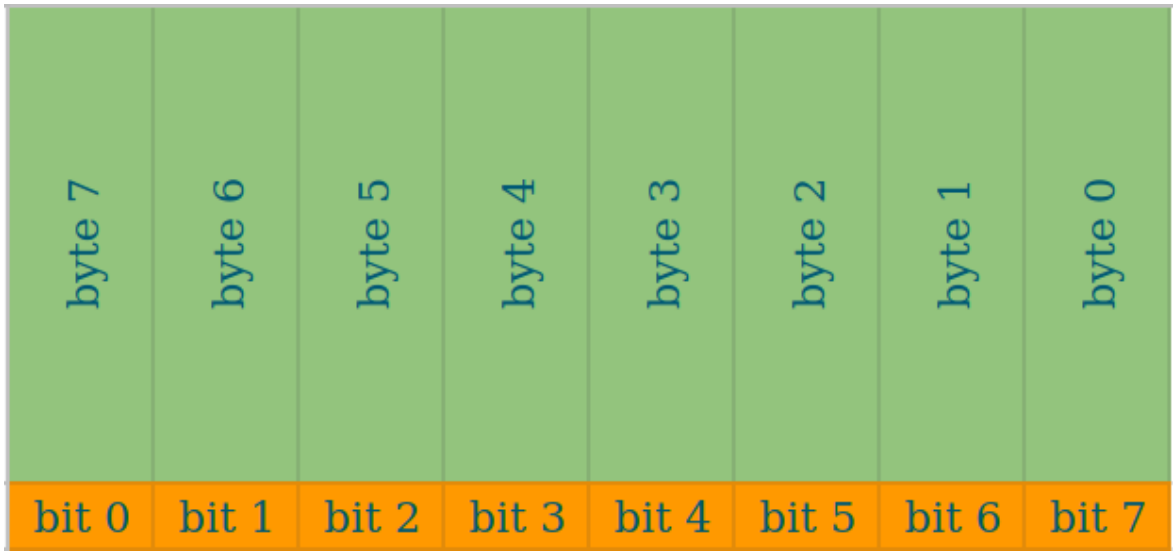
| Tarea | Cifrado en uso |
|---|---|

| | |
|---|---|
| esteganografía | Para descifrar el módulo del kernel, el cargador inicial utiliza un algoritmo personalizado. |
| Cifrado de registros | Los registros de malware existen al mismo tiempo en la memoria y en archivos cifrados en el disco. En la RAM, los desarrolladores almacenan los registros en texto plano, en el disco utilizan XOR de un byte. |
| Cifrado de configuración | El módulo kernel utiliza potentes algoritmos de cifrado. Los datos de configuración se cifran con 3DES y la clave se cifra mediante RSA. Todas las claves (RSA pública/privada y 3DES cifradas) se almacenan dentro de la sección .data del módulo. |
| Cifrado del módulo de red | HttpTransport inicialmente cifrado está formado por cuatro blobs binarios almacenados en el módulo del kernel. El kernel los concatena y los descifra con un algoritmo personalizado basado en XOR. Se utiliza una clave redonda de cuatro bytes de longitud. |

| | |
|---|---|
| Cifrado de comunicaciones | El algoritmo de cifrado es RSA y utiliza las mismas claves públicas y privadas almacenadas dentro de la sección .data del módulo del kernel. |

**Módulo de carga: descifrador de mapas de bits y lanzador de la siguiente etapa**

Kaspersky Lab describe que si el nombre de archivo del mapa de bits que contiene los datos cifrados con esteganografía se proporciona al cargador como argumento, el cargador descifra el siguiente paso de la matriz de píxeles. En la primera iteración, extrae los datos de los parámetros de esteganografía. Para ello, el algoritmo toma los últimos bits de los bytes.




Los parámetros IID, IParam e ISize se mantienen en los primeros 384 bytes de la matriz de píxeles, lo que significa que sólo se necesita el último bit de cada byte de la matriz de píxeles. Como resultado, el módulo reúne 48 bytes de estructura de configuración de esteganografía con los campos, determinando las siguientes etapas de descifrado.

| Campo | **Compensar** | **Características** |
|---|---|---|
| IID | 0x00 | Determina que una o dos capas de descifrado se aplicarían a la siguiente matriz de píxeles. |
| IParam | 0x04 | Determina qué bits de bytes de matrices de píxeles formarían el siguiente módulo del kernel. |
| ITamaño | 0x28 | El tamaño resultante del módulo del kernel descifrado. |

Después de extraer los parámetros de esteganografía, el siguiente paso se descifra utilizando un algoritmo de dos pasos. En primer lugar, el algoritmo IParam elige los bits de los bytes de la matriz de píxeles. Luego, si IID es igual a 2, se aplica una operación de dexorización personalizada utilizando una clave redonda de cuatro bytes en los bytes recopilados. La clave inicial para el primer descifrado de cuatro bytes tiene el valor codificado 0x23041920. Entonces la fórmula para la clave XOR redonda para los siguientes bytes es:

clave ^= 8 * ( clave ^ ( clave << 20 ))

Consideramos que este algoritmo de esteganografía está hecho a medida y no tomado de ningún repositorio de código abierto de terceros. Sorprendentemente, el resultado del descifrado no se inyecta en la memoria de algún proceso, sino que se guarda en el disco como un archivo llamado msgslang32.dll. Luego, el cargador simplemente usa las funciones API de Windows LoadLibraryW() y GetProcAddress() para ejecutar la función Open() del siguiente stager, como vimos anteriormente con el módulo del cargador.

**Módulo kernel: descifrador de configuración y despachador de tareas**

El módulo del kernel contiene tres claves de cifrado utilizadas para el descifrado de la configuración y las comunicaciones C2. Las claves RSA públicas y privadas se almacenan en la sección .data como PUBLICKEYBLOB y PRIVATEKEYBLOB respectivamente. Se utilizan para cifrar las comunicaciones C2 y también para descifrar la clave 3DES. La tercera clave 3DES también se almacena en la sección .data en su forma cifrada; esta clave se utiliza para descifrar un archivo .cab incrustado que contiene la configuración XML. Para descomprimir el archivo .cab, el módulo utiliza la utilidad estándar del sistema de Windows, “expand.exe”. Veremos otro uso de software común en el módulo HttpTransport.

La configuración XML contiene datos valiosos que nos ayudan a comprender el interés del operador de la campaña. Está estructurado utilizando varias "tareas" para el malware, como tomar huellas digitales del objetivo utilizando su versión del sistema operativo, lista de procesos y capturas de pantalla; pero también toma la lista de documentos recientes de los usuarios con cualquiera de las extensiones .doc, .docx, .xls, .xlsx, .rtf, .pdf, .odt, .psw, .pwd de varios directorios de documentos recientes en %USERPROFILE % y %APPDATA%, incluidos %APPDATA%\Microsoft\Office\Последние файлы. El nombre de esta carpeta se traduce como “Archivos recientes” en ruso, lo que sugiere que el malware está dirigido a versiones localizadas en cirílico de Windows.

```
        <cp>sunT23:40:34</cp>
        <cp>sunT23:48:01</cp>
        <cp>sunT23:57:01</cp>
</SCHEDULE>
  <Dropbox>
    <SENDRECV chunk="500000" access_token="<edited>"/>
  </Dropbox>
</TASK>.MT3D..}X|..........R.e.c.e.n.t. .[.d.o.c. .d.o.c.x. .x.l.s. .x.l.s.x. .r.t.f. .p.d.f. .o.d.t. .p.s.w. .p.w.d.]...x.m.l.....<?xml version
="1.0" encoding="utf-8"?>
<!--Recent doc docx xls xlsx rtf odt psw pwd-->
<TASK xsi:noNamespaceSchemaLocation="../../../../Task.xsd" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" persistent="0">
  <File>
    <FIND mask="doc;docx;xls;xlsx;rtf;pdf;odt;psw;pwd" subdir="1" get="1" path="\\?\%USERPROFILE%\Recent" maxsize="1000000" ignore_temp="1" older=
"14" />
    <FIND mask="doc;docx;xls;xlsx;rtf;pdf;odt;psw;pwd" subdir="1" get="1" path="\\?\%APPDATA%\Microsoft\Office\Recent" maxsize="1000000" ignore_te
mp="1" older="14" />
    <FIND mask="doc;docx;xls;xlsx;rtf;pdf;odt;psw;pwd" subdir="1" get="1" path="\\?\%APPDATA%\Microsoft\Office\Последние файлы" maxsize="1000000"
ignore_temp="1" older="14" />
    <FIND mask="doc;docx;xls;xlsx;rtf;pdf;odt;psw;pwd" subdir="1" get="1" path="\\?\%APPDATA%\Microsoft\Windows\Recent" maxsize="1000000" ignore_t
emp="1" older="14" />
```


*La configuración contiene la programación de tareas (captura de pantalla superior), tokens de acceso (aquí Dropbox, redactado), directorios y extensiones de interés. Sólo existe un directorio en las versiones localizadas de Windows en cirílico*

Observamos varias cadenas de texto cirílico como “Снимок рабочего стола” (instantánea de escritorio), “Системная информация” (información del sistema), “Время выхода” (hora de salida).

```
</TASK>.MT3D..}XH..........!.8.A.B.5.<.=.0.0. .8.=.D.>.@.<.0.F.8.0...x.m.l......<?xml version="1.0" encoding="utf-8"?>
<!--Системная информация-->
<TASK xsi:noNamespaceSchemaLocation="../../../../Task.xsd" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" persistent="0">
  <SysInfo>
    <INFO />
  </SysInfo>
</TASK>.MT3D..}X@..........!.?.8.A.>.:. .?.@.>.F.5.A.A.>.2...x.m.l......<?xml version="1.0" encoding="utf-8"?>
<!--Список процессов-->
<TASK xsi:noNamespaceSchemaLocation="../../../../Task.xsd" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" persistent="0">
  <SysInfo>
    <TASKLIST />
  </SysInfo>
</TASK>.MT3D..}X^...A.......!.B.0.@.B.>.2.K.9. .A.=.8.<.>.:. .@.0.1.>.G.5.3.>. .A.B.>.;.0...x.m.l......<?xml version="1.0" encoding="utf-8"?>
<!--Снимок рабочего стола-->
<TASK xsi:noNamespaceSchemaLocation="../../../../Task.xsd" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" persistent="0">
  <SysInfo>
    <SCREENSHOT type="image/jpeg" />
  </SysInfo>
  <SCHEDULE pause="0" />
</TASK>.
```

*La descripción de las tareas de configuración comienza con MT3D y contiene frases cortas adecuadas en ruso.*

**La estructura de configuración descifrada es la siguiente:**

| Campo | **Tamaño** | **Contenido** |
|---|---|---|
| magia | 4 bytes | MT3D. Todos los archivos analizados deben tener esto como prefijo para ser válidos. |
| Tiempo de creación | 4 bytes | Marca de tiempo, hora de creación de configuración de tarea almacenada como hora de época |

| | | |
|---|---|---|
| Tamaño del encabezado | 4 bytes | El tamaño del encabezado debe ser mayor que 18. El valor observado es, por ejemplo, 0x7E. |
| Tamaño XML | 4 bytes | La descripción de la tarea XML debe ser mayor que cero. El valor observado es, por ejemplo, 0x662D |
| cuerpo XML | Tamaño XML | La descripción y programación de la tarea en formato XML. |

Si bien las muestras que analizamos no contenían información RTTI, los registros de ejecución nos permitieron recuperar los nombres de las clases de C++. Después de que el módulo del kernel analiza las tareas de la configuración en la memoria, la clase principal que procesa la instrucción es CTask. El método IoControl() de CTask se encarga de manejar las tareas correspondientes y a su vez ejecuta los siguientes métodos:

| Método CTask | **Características** |
|---|---|

| | |
|---|---|
| PrincipalIoControl() | Manejador de tarea “Principal” en XML. En caso de un comando RESET, el archivo, que actúa como "tubería", se eliminará. Cualquier otro comando aquí se registrará, pero no se ejecutará. |
| ArchivoIoControl() | Controlador de la tarea "Archivo" con los subcomandos PUT, DEL, FIND, WATCH, WATCH_REMOVABLE, RUN y LOGS |
| SysInfoIoControl() | Manejador de la tarea “SysInfo” con los subcomandos SCREENSHOT, INFO y TASKLIST |
| HttpIoControl() | Manejador de la tarea “Http” con el subcomando SENDRECV |

| | |
|---|---|
| GDriveIoControl() | Manejador de la tarea “GDrive” con el subcomando SENDRECV |
| DropboxIoControl() | Manejador de la tarea “Dropbox” con el subcomando SENDRECV |

Todos los métodos utilizados para las comunicaciones externas primero descifran el módulo HttpTransport y lo utilizan para transmitir los datos correspondientes cifrados con RSA. Las claves RSA en uso son las mismas claves antes mencionadas que se utilizan para descifrar la clave de configuración 3DES. En un procedimiento de ventana separado, el malware monitorea si hay un dispositivo USB conectado y busca archivos de interés.

### 9.3.2. Módulo HttpTransport: tareas de red

El módulo HttpTransport existe como cuatro fragmentos de datos cifrados dentro de la sección .text del módulo del kernel. Cuando el kernel necesita comunicarse, descifra este módulo y, como es habitual en MontysThree, ejecuta la función Open(), pasando argumentos de línea de comando. Dependiendo de los argumentos transmitidos desde el módulo del kernel, el módulo puede cargar o descargar contenido utilizando los protocolos RDP, WebDAV, Citrix y HTTP. La descarga de datos de los servicios públicos de Google y Dropbox mediante tokens de usuario también se implementa en HttpTransport. En el caso de solicitudes HTTP GET/POST, el malware recibiría una imagen de mapa de bits de esteganografía utilizando funciones relacionadas con HTTP de la API de Windows desde una URL correspondiente.

Los protocolos de comunicación antes mencionados no están implementados dentro del módulo. Los autores de malware utilizan programas legítimos de Windows como RDP, clientes Citrix e Internet Explorer ya instalados en la máquina del objetivo. Por ejemplo, el módulo ejecuta una tarea para enviar algunos datos a una URL y recibir la respuesta a través de una conexión RDP de la siguiente manera: editar el archivo .rdp para ejecutar Internet Explorer de forma silenciosa en la máquina remota; pegue la URL en el navegador a través del portapapeles; espere y pegue también el contenido en la página web abierta a través del portapapeles; espere y reciba el resultado a través del portapapeles nuevamente.

Para copiar datos, el malware envía literalmente Ctrl+C, Ctrl+V y Ctrl+A. Quizás sea la primera vez que vemos un método de “comunicación RDP” de este tipo. La comunicación con Citrix se realiza mediante un procedimiento similar: el malware no implementa el protocolo sino que busca Windows Quick Launch .lnk para XenApp pnagent.exe, ejecuta Internet Explorer de forma remota y se comunica con él a través del portapapeles mediante atajos de teclado especiales. La carga y descarga de datos de Dropbox y Google se basa en otro principio: su implementación utiliza la clase personalizada CSimpleHttp para autenticar y enviar solicitudes HTTP. Para la comunicación WebDAV, los desarrolladores simplemente usan el comando de Windows "net use".

**Actualización de enlace:** Este módulo auxiliar se encarga de lograr la persistencia en el host. Cambia los archivos .lnk en el panel de inicio rápido de Windows para ejecutar el cargador junto con aplicaciones legítimas, como navegadores, cuando el usuario los ejecuta utilizando el enlace modificado.

### 9.3.3 ¿Quién está detrás de este malware?

Como mencionamos al principio, hasta la fecha no hemos observado similitudes ni superposiciones con campañas conocidas en términos de TTP, infraestructura o código malicioso. Hasta ahora, atribuimos esta actividad y el uso de MontysThree a un nuevo actor. Algunas muestras contienen detalles de cuentas utilizados para comunicarse con servicios de nube pública, que pretenden ser de origen chino. Teniendo en cuenta todos los artefactos cirílicos antes mencionados, consideramos que estos nombres de cuentas son señales de alerta falsas. Suponemos que el actor detrás de MontysThree habla ruso y persigue objetivos de habla rusa. Algunos de los nombres de los archivos RAR SFX utilizados para difundir el malware estaban escritos en ruso y hacían referencia a un laboratorio médico ruso, utilizado para incitar al usuario a abrir el archivo. La configuración XML mostró campos de datos y títulos de Windows escritos en ruso, así como rutas de carpetas específicas que existen en versiones localizadas en cirílico de Windows. También vimos algunos errores gramaticales en las cadenas de mensajes de registro en inglés del malware.

Normalmente vemos malware dirigido principalmente a entidades gubernamentales, diplomáticos y operadores de telecomunicaciones, lo que resulta fructífero para los actores patrocinados por el estado. Los casos de espionaje industrial como el de MontysThree son mucho más raros.La sofisticación general de la campaña no se compara con la de los actores APT de primer nivel en términos de método de difusión y persistencia. Y algunos aspectos del malware (iniciar sesión en RAM y archivos al mismo tiempo, mantener las claves de cifrado en el mismo archivo, ejecutar un navegador invisible en el host RDP remoto) parecen inmaduros y amateurs en términos de desarrollo de malware. Por otro lado, la cantidad de código y por tanto de esfuerzo invertido, en MontysThree es significativa. El conjunto de herramientas demuestra algunas decisiones tecnológicas: almacenar la clave 3DES bajo cifrado RSA, esteganografía personalizada para evitar IDS y el uso de proveedores legítimos de almacenamiento en la nube para ocultar el tráfico C2.

**Hashes de archivos**

**Cargador**

1B0EE014DD2D29476DF31BA078A3FF48
0976C442A06D2D8A34E9B6D38D45AE42
A2AA414B30934893864A961B71F91D98

**Núcleo**

A221671ED8C3956E0B9AF2A5E04BDEE3
3A885062DAA36AE3227F16718A5B2BDB
3AFA43E1BC578460BE002EB58FA7C2DE

**HttpTransporte**

017539B3D744F7B6C62C94CE4BCA444F
501E91BA1CE1532D9790FCD1229CBBDA
D6FB78D16DFE73E6DD416483A32E1D72

Dominios e IP

autosport-club.tekcities[.]com
dl10-web-stock[.]ru
dl16-web-etiquetado[.]ru
dl166-web-etiquetado[.]ru
dl55-web-reserva de yates[.]xyz

## 9.4. ESTUDIOS RELACIONADOS EN CHINA

En el marco del APSIPA "Asia-Pacific Signal and Information Processing Association" La universidad de Beijing se hizo presente con el trabajo “Study of Chinese Text Steganography using Typos” (2019) presentado por Linna Zhou y Derui Liao. En este se mostró la problemática actual , con la Explosión de la Información y el rápido desarrollo de la tecnología de la información, se generan cantidades enormes de datos constantemente cada día en Internet. Sin embargo, la mayoría de los textos proporcionados en línea suelen contener muchos errores tipográficos, lo cual es muy común entre los usuarios individuales, los medios de comunicación personal, etc. A pesar de esto, la desambiguación es un talento humano, por lo que estos errores tipográficos a menudo no dificultan la comprensión del texto por parte de las personas, e incluso a veces es difícil reconocer algunos errores.

Este fenómeno aparece tanto en inglés como en chino, por lo que parece ser interlingüístico. Por lo tanto, no sorprende que en tales textos se pueda realizar el ocultamiento de información al inyectar cuidadosamente errores tipográficos.Las autoras han investigado los errores tipográficos en chino en el contenido de texto en Weibo o WeChat, y proporcionan un método de esteganografía de texto basado en errores tipográficos en chino con la ayuda de Procesamiento de Lenguaje Natural (NLP), que puede incrustar información secreta mediante errores tipográficos cuidadosamente inyectados y garantizar la seguridad del secreto y la legibilidad de los textos. A diferencia de los algoritmos de esteganografía basados en formato, su algoritmo puede resistir ajustes de formato, reintroducción de OCR, etc. Además, las plataformas Weibo y WeChat contienen muchos tipos de medios, por lo que al combinar otros algoritmos, la esteganografía de información Cross-Media o incluso Cross-Social Network es práctica.

**9.4.1 Errores tipográficos en el chino:** En el ámbito de los errores tipográficos chinos, la búsqueda de minimizar los errores tipográficos tiene una larga historia, principalmente en la educación y la publicación. El enfoque ha sido reducir la ocurrencia de errores tipográficos, esos deslices que surgen al ingresar texto con un teclado. La exploración de Geng indaga en la interacción entre el método de entrada Pinyin y los errores de red, mientras que tanto Geng como Chen desentrañan las causas detrás de estos errores. El surgimiento de modelos de representación del lenguaje marcó un punto de inflexión en 2018, con la presentación de BERT (Bidirectional Encoder Representation from Transformers) por parte de Google, un modelo preentrenado aclamado como de última generación. Sin embargo, surgió una limitación notable en el Modelo de Lenguaje Enmascarado (MLE) de BERT, que modelaba principalmente palabras en inglés y caracteres chinos. Esta discrepancia lingüística planteó desafíos para la comprensión del modelo de representaciones semánticas completas, especialmente en el contexto del chino. ERNIE (Enhanced Representation through Knowledge Integration) surgió como una solución, superando este obstáculo y demostrando un rendimiento superior en chino.

En el ámbito de la esteganografía de texto, ocultar información en texto plano toma diversas formas, incluidas modificaciones a caracteres no semánticos como caracteres invisibles o signos de puntuación. Gan y Shirali-Shahreza emplean sinónimos y abreviaturas, respectivamente, para incrustar información. Basándose en el uso de redes neuronales para ocultar información, Luo et al. generan ingeniosamente poemas Tang que llevan mensajes ocultos. El trabajo de Wayner pionero en ocultar información al alterar el orden de los caracteres en palabras en inglés. Liu utiliza la codificación de matrices para ocultar información en texto en inglés, mientras que Topkara aprovecha transformaciones asimétricas computacionalmente (CAT) para resistir ataques de reemplazo de sinónimos y utiliza errores tipográficos, otra palabra existente, para evitar la revisión ortográfica.

Tomando inspiración del método de Liu (2014) , la integración de errores tipográficos en el resumen se hace evidente. El texto reflexiona sobre la Explosión de Información y el rápido desarrollo de la tecnología de la información, destacando la prevalencia de errores tipográficos en el contenido en línea. A pesar de la ubicuidad de los errores tipográficos, la inherente capacidad humana para la desambiguación a menudo mitiga el impacto en la comprensión. Reconociendo este fenómeno tanto en textos en inglés como en chino, el texto sugiere un potencial de cruce lingüístico para realizar ocultamiento de información mediante la inyección estratégica de errores tipográficos.

El enfoque luego se desplaza a una exploración específica de errores tipográficos chinos en plataformas como Weibo y WeChat. Al proponer un método de esteganografía de texto basado en errores tipográficos chinos con la ayuda de Procesamiento del Lenguaje Natural (NLP), el objetivo es incrustar información secreta mediante errores tipográficos cuidadosamente inyectados, garantizando al mismo tiempo la seguridad y la legibilidad del texto. Notablemente, este algoritmo se mantiene resiliente frente a ajustes de formato y reintroducciones de OCR. Además, con la abundancia de medios en plataformas como Weibo y WeChat, emerge la perspectiva de combinar algoritmos para ocultar información de manera práctica en medios cruzados o incluso en redes sociales cruzadas. Si bien la eficacia de los métodos basados en errores tipográficos es bien conocida en el texto en inglés, la investigación en el ámbito del texto en chino sigue siendo relativamente escasa. Este estudio colma esa brecha, arrojando luz sobre el potencial de la esteganografía de texto fundamentada en errores tipográficos chinos y sentando las bases para una exploración adicional en este dominio cruzado de lenguajes.

**9.4.2. Textos y redes sociales en China:** Cada día, se produce una enorme cantidad de información en sitios web de redes sociales, incluyendo imágenes, audio, video y mucho texto. Al mismo tiempo, con el rápido desarrollo de la tecnología y el Internet en China, el número de usuarios de Internet chinos ha aumentado rápidamente. Por lo tanto, los mensajes chinos en Internet deben ser incontables. Weibo (Sina) y WeChat (Tencent) pueden considerarse como los dos gigantes en las redes sociales chinas, con más de 462 millones de usuarios activos mensuales en Weibo y más de mil millones en WeChat. La gran cantidad de usuarios inevitablemente genera grandes cantidades de información todos los días, y encontramos que los errores tipográficos suelen aparecer en el texto publicado en Weibo o WeChat.

La existencia de errores tipográficos es tan extensa que a veces pueden aparecer en textos publicados por cuentas oficiales en tiempo real. Los errores también aparecen en algunos artículos largos publicados en las Cuentas Oficiales de Weibo o WeChat, pero esto generalmente no afecta la comprensión del lector. Los contenidos publicados por cuentas personales suelen ser más informales y con más errores tipográficos, e incluso algunos blogueros de Weibo hacen de los errores tipográficos su estilo de blogs. A veces, los errores tipográficos pueden agregar humor en las redes sociales. Además, cada vez más personas navegan por Internet en teléfonos inteligentes, lo que implica la dificultad de ingresar texto a través del teclado virtual en la pantalla. Aunque las pantallas de los teléfonos inteligentes son cada vez más grandes, el teclado virtual es mucho más pequeño que el real y no hay espacio entre las teclas. Por lo tanto, es más probable que ocurran errores tipográficos al escribir en teléfonos inteligentes. Los errores tipográficos no pueden evitarse por completo incluso cuando algunos métodos de entrada se esfuerzan en la corrección ortográfica. En inglés, las palabras están compuestas por letras y separadas por espacios. Pero el conjunto de caracteres chinos es demasiado grande para organizarse en el teclado, por lo que tenemos varios métodos de entrada chinos como Pinyin, Wubi, Bihua, Escritura, entre otros.

Para el teclado, tenemos el teclado de 9 teclas (T9) y el teclado de 26 teclas (T26). Diferentes métodos de entrada chinos tienen que resolver diferentes problemas, como la asignación de uno a muchos y el reconocimiento de la escritura a mano. Por lo tanto, se puede inferir que los errores tipográficos en chino incluyen los siguientes tipos: seleccionar la palabra equivocada, presionar la tecla equivocada o más/menos teclas, caracteres/palabras duplicados o faltantes, entre otros. Seleccionar el homónimo, que significa seleccionar la palabra incorrecta con la misma secuencia de pinyin o una secuencia de pinyin similar en la lista de candidatos, es común al seleccionar la palabra equivocada. Al presionar la tecla equivocada, el homónimo podría ocurrir si se presiona otra tecla final. Pero una tecla inicial incorrecta lleva a una palabra completamente incorrecta, a veces puede ser confusa. Por ejemplo, escribir 我好羡慕啊 en el teclado T9 podría presionar 964269426682 en orden o 96 para abreviar (con la ayuda de la sugerencia admitida por el método de entrada). Pero si se presiona 99, el resultado podría ser otra palabra: 真相大白天, una palabra completamente incorrecta. Presionar más/menos teclas podría llevar a un resultado similar a presionar la tecla equivocada. Dado que los errores tipográficos en chino son completamente diferentes de los de inglés, los algoritmos de esteganografía de texto en inglés basados en errores de entrada disponibles a veces son difíciles de aplicar al chino.

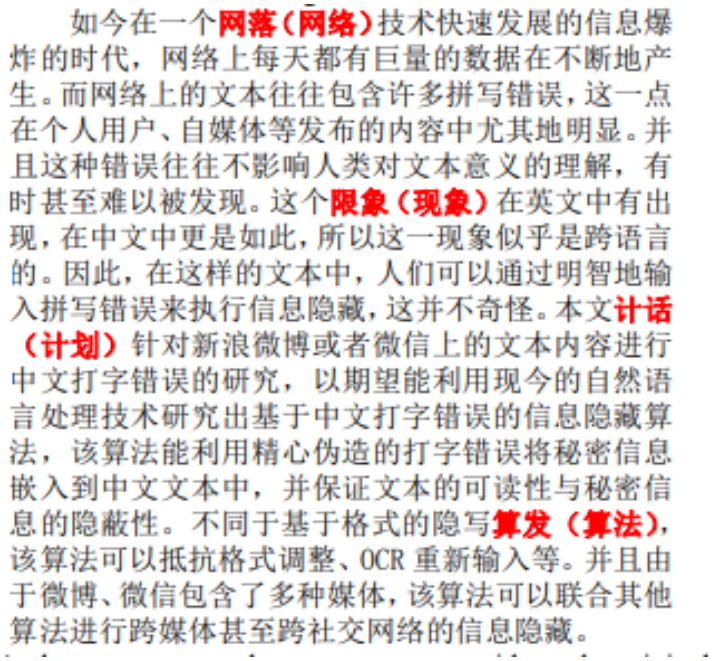
**9.4.3. Método esteganográfico:** El método de esteganografía utilizado se centra en la inserción de información en textos chinos mediante la modificación controlada de palabras. Para ello, se emplea el modelo BERT, capaz de predecir el token enmascarado en un contexto y su probabilidad. Este modelo se utiliza para predecir cada palabra en el contexto y determinar si coincide con la palabra original, permitiendo así la recuperación de palabras faltantes o incorrectas. Sin embargo, la estructura bidireccional de BERT dificulta la recuperación del token original al modificar más de un token a la vez. A diferencia del inglés, que utiliza separadores entre palabras, el chino separa las

palabras por su semántica, lo que requiere la segmentación de palabras chinas para obtener palabras significativas.

Se introduce el error tipográfico seleccionando el homónimo incorrecto, reemplazando un carácter por otro con una pronunciación similar. Se elige una palabra con más de dos caracteres para transportar información, utilizando el índice del carácter modificado como el bit secreto. El proceso de inserción implica tokenizar el texto, aplicar la Segmentación de Palabras Chinas, y seleccionar palabras según criterios específicos. Se evita la modificación de palabras clave para evitar confusiones. Luego, se selecciona una palabra en cada instancia para transformarla y embeber un bit secreto.

La extracción de la información se realiza mediante la tokenización del texto, manteniendo solo los tokens que no fueron modificados durante la inserción. Se aplican filtros basados en la probabilidad de predicción, la no pertenencia a una lista de palabras detenidas y la correspondencia con el homónimo predicho. Se aborda la problemática de las palabras intocables mediante la identificación de palabras de alta probabilidad y la utilización de una lista de palabras detenidas. Además, se propone limitar el alcance de uso del método a situaciones informales, como en Weibo y WeChat, donde la interacción frecuente permite la desambiguación a través de la comunicación.

**9.4.4. Ejemplificación y experimentación:** Su método implica ejecutar el modelo BERT en la versión de PyTorch en Python, con parámetros configurados en 12 capas, 768 unidades ocultas, 12 cabezas y 110 millones de parámetros para los modelos chinos. Utilizando un punto (。) como separador de instancias, la versión china de su resumen se divide en 5 instancias, cada una capaz de transportar hasta 5 bits. El resumen chino oculta el mensaje secreto 0b1101 y se presenta de la siguiente manera:


如今在一个**网落（网络）**技术快速发展的信息爆炸的时代，网络上每天都有巨量的数据在不断地产生。而网络上的文本往往包含许多拼写错误，这一点在个人用户、自媒体等发布的内容中尤其地明显。并且这种错误往往不影响人类对文本意义的理解，有时甚至难以被发现。这个**限象（现象）**在英文中有出现，在中文中更是如此，所以这一现象似乎是跨语言的。因此，在这样的文本中，人们可以通过明智地输入拼写错误来执行信息隐藏，这并不奇怪。本文**计话（计划）**针对新浪微博或者微信上的文本内容进行中文打字错误的研究，以期望能利用现今的自然语言处理技术研究出基于中文打字错误的信息隐藏算法，该算法能利用精心伪造的打字错误将秘密信息嵌入到中文文本中，并保证文本的可读性与秘密信息的隐蔽性。不同于基于格式的隐写**算发（算法）**，该算法可以抵抗格式调整、OCR 重新输入等。并且由于微博、微信包含了多种媒体，该算法可以联合其他算法进行跨媒体甚至跨社交网络的信息隐藏。

Y se puede extraer de esta manera el texto original. El uso de tipos de errores especiales puede imitar un texto de baja calidad escrito por un humano descuidado, lo cual podría generar menos sospechas.

```
high prob unhit: ------------------
6 落 络 0.9979057312011719 0
118 限 现 0.9992528557777405 0
194 话 划 0.9971093535423279 0
318 发 法 0.9993120431900024 0
groups:-----------------------------
如今在一个网落技术快速发展的信息爆炸的时代，网络上每天都有巨量的数据在不断
如今在一个网络技术快速发展的信息爆炸的时代，网络上每天都有巨量的数据在不断
typo at [6] : 网络 5 6 1
这个限象在英文中有出现，在中文中更是如此，所以这一现象似乎是跨语言的。因此，
这个现象在英文中有出现，在中文中更是如此，所以这一现象似乎是跨语言的。因此，
typo at [2] : 现象 2 2 0
本文计话针对新浪微博或者微信上的文本内容进行中文打字错误的研究，以期望能利
本文计划针对新浪微博或者微信上的文本内容进行中文打字错误的研究，以期望能利
typo at [3] : 计划 2 3 1
不同于基于格式的隐写算发，该算法可以抵抗格式调整、OCR重新输入等。
不同于基于格式的隐写算法，该算法可以抵抗格式调整、OCR重新输入等。
typo at [11] : 算法 10 11 1
并且由于微博、微信包含了多种媒体，该算法可以联合其他算法进行跨媒体甚至跨社
并且由于微博、微信包含了多种媒体，该算法可以联合其他算法进行跨媒体甚至跨社
No typo
Msg: 0b1101

Process finished with exit code 0
```

**Ejemplo**

Utilizamos el resumen en versión china como el texto para ilustrar los pasos de inserción y extracción de ocultar un mensaje secreto de 0b1101. En resumen, la inserción implica dividir, predecir, filtrar y reemplazar.

- **División:** Primero dividimos el texto en segmentos mediante un separador de instancias: "。". Luego, combinamos esos segmentos en instancias no mayores a una longitud dada .

Aquí hay algunos ejemplos:

Table 1 Examples of instances after dividing and masking

| Dividing examples |
|---|
| [CLS][MASK]今在一个网络技术快速发展的信息爆炸的时代，……有时甚至难以被发现。[SEP] |
| [CLS]如[MASK]在一个网络技术快速发展的信息爆炸的时代，……有时甚至难以被发现。[SEP] |
| [CLS]如今[MASK]一个网络技术快速发展的信息爆炸的时代，……有时甚至难以被发现。[SEP] |
| … |
| [CLS][MASK]个现象在英文中有出现，……这并不奇怪。[SEP] |
| [CLS]这[MASK]现象在英文中有出现，……这并不奇怪。[SEP] |
| … |

b) Predicting

Then we feed the instances to BERT to get the prediction of each *MASK*.

Table 2 Examples of prediction during embedding

| Real | Prediction | Probability | Hit |
|---|---|---|---|
| 如 | 如 | 0.9307 | ✓ |
| 今 | 今 | 0.8824 | ✓ |
| 在 | 是 | 0.9621 | × |
| …… | …… | …… | …… |
| 网 | 网 | 1.0000 | ✓ |
| 络 | 络 | 0.9979 | ✓ |
| 技 | 技 | 0.9999 | ✓ |
| …… | …… | …… | …… |

c) Filtering

We apply Chinese Word Segmentation to the instances, keep the valid tokens, and get the embeddable words. We set the $P_{embed_{word}} = 0.94$ and $P_{embed_{word}} = 0.92$ here.

| Instance | Index | Word | Prediction | Hit | Chinese | Probability | Embeddable | Why | Homonyms |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 如今 | 如今 | ✓ | ✓ | 0.906548 | × | low prob | - |
| 0 | 1 | 在 | 是 | × | ✓ | 0.962071 | × | not word or unhit | - |
| 0 | 2 | 一个 | 这个 | × | ✓ | 0.896691 | × | not word or unhit | - |
| 0 | 3 | 网络 | 网络 | ✓ | ✓ | 0.998935 | ✓ | - | 望、往、王……; 落、罗、洛…… |
| 0 | 4 | 技术 | 技术 | ✓ | ✓ | 0.999660 | ✓ | - | 机、计、级……; 数、书、属…… |
| …… | …… | …… | …… | …… | …… | …… | …… | …… | …… |

**Reemplazo:** Finalmente, seleccionamos una palabra al azar en una instancia para insertar el bit secreto hasta ocultar todos los bits en el texto. En este caso, optamos por la primera palabra para simplificar.

Table 4 Examples of replacing during embedding

| Instance | Index | Original | Typo |
|---|---|---|---|
| 0 | 5 | 网络 | 网**落** |
| 1 | 118 | 现象 | **限**象 |
| 2 | 193 | 计划 | 计**话** |
| 3 | 317 | 算法 | 算**发** |

La extracción implica dividir, predecir, filtrar, recuperar y extraer.

**a) División y Predicción:** Al igual que en los pasos de inserción, podemos obtener la probabilidad de cada token, algunos de ellos se muestran a continuación.

| Real | Prediction | Probability | Hit |
|---|---|---|---|
| 如 | 如 | 0.9395 | ✓ |
| 今 | 今 | 0.8917 | ✓ |
| 在 | 是 | 0.9576 | × |
| …… | …… | …… | …… |
| 网 | 网 | 0.3542 | ✓ |
| **落** | **络** | 0.9979 | × |
| …… | …… | …… | …… |

**b) Filtrado:** Algunos tokens tienen una alta probabilidad de coincidencia pero no lo hacen, lo que podría deberse a los errores tipográficos que inyectamos.

**c) Recuperación y Extracción:** Recuperamos esos errores tipográficos y registramos sus ubicaciones. Luego aplicamos la Segmentación de Palabras Chinas a las instancias y determinamos el índice del carácter homónimo en esas palabras con errores tipográficos. De esta manera, podemos extraer los bits del mensaje secreto.

| Index | Typo | Correction |
|---|---|---|
| 6 | 落 | 络 |
| 118 | 限 | 现 |
| 194 | 话 | 划 |
| 318 | 发 | 法 |

| Index | Real | Prediction | Probability | Hit | stop word | Homonym | Recover |
|---|---|---|---|---|---|---|---|
| 0 | 如 | 如 | 0.939501 | ✓ | - | - | × |
| 1 | 今 | 今 | 0.891668 | ✓ | - | - | × |
| 2 | 在 | 是 | 0.957569 | × | × | × | × |
| ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... |
| 5 | 网 | 网 | 0.354176 | ✓ | - | - | × |
| 6 | 落 | 络 | 0.997906 | × | × | ✓ | ✓ |
| 7 | 技 | 技 | 0.999677 | ✓ | - | - | × |
| ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... |

| Injected | Recovered | Chinese Word Segmentation | Typo at word | Typo index |
|---|---|---|---|---|
| 如今在一个网**落**技术快速发展的信息爆炸的时代，.....有时甚至难以被发现。 | 如今在一个网**络**技术快速发展的信息爆炸的时代，.....有时甚至难以被发现。 | 如今\|在\|一个\|网**络**\|技术\|快速\|发展\|的\|信息\|爆炸\|的\|时代\|，.....有时\|甚至\|难以\|被\|发现\|。 | 6 | 1 |
| 这个**限**象在英文中有出现，.....这并不奇怪。 | 这个**现**象在英文中有出现，.....这并不奇怪。 | 这个\|**现**象\|在\|英文\|中有\|出现\|，.....这\|并\|不\|奇怪\|。 | 2 | 0 |
| 本文计**话**针对新浪微博或者微信.....并保证文本的可读性与秘密信息的隐蔽性。 | 本文计**划**针对新浪微博或者微信.....并保证文本的可读性与秘密信息的隐蔽性。 | 本文\|计**划**\|针对\|新浪\|微博\|或者\|微信.....并\|保证\|文本\|的\|可读性\|与\|秘密\|信息\|的\|隐蔽性\|。 | 3 | 1 |
| 不同于基于格式的隐写算**发**，该算法可以抵抗格式调整、OCR 重新输入等。 | 不同于基于格式的隐写算**法**，该算法可以抵抗格式调整、OCR 重新输入等。 | 不同于\|基于\|格式\|的\|隐写\|算法\|，\|该\|算**法**\|可以\|抵抗\|格式\|调整\|、\|OCR\|重新\|输入\|等\|。 | 10 | 1 |
| 并且由于微博、微信..... | No typo | - | - | - |

Este método de esteganografía de textos propone una novedosa aproximación utilizando errores tipográficos con la ayuda de BERT. La técnica consiste en inyectar mensajes secretos en textos chinos creando deliberadamente errores tipográficos y luego recuperar esos errores para revelar el mensaje secreto mediante el modelo de procesamiento de lenguaje natural, BERT, sin necesidad del texto original. Esta metodología es especialmente aplicable en contextos informales como redes sociales tipo Weibo y WeChat, donde es común encontrar errores tipográficos.

La tasa de inserción de información no es constante y depende del portador utilizado. Se exploran distintos separadores de instancias, como el punto chino ("。") y caracteres no chinos (NCC), mostrando variabilidad en la tasa de inserción. La indetectabilidad se vincula con la longitud de la instancia, sugiriendo que modificar solo una palabra en una instancia no suele ser fácil de notar. Aunque los errores tipográficos pueden parecer extraños cuando se insertan de manera densa, es menos probable que provoquen sospechas, especialmente en situaciones informales como redes sociales. En cuanto a las contribuciones, se destaca la adaptación del método de Liu para aplicar la esteganografía en textos chinos, con la inserción de un mensaje secreto de 0b11010000 00000000. Aunque este método presenta una tasa de inserción más alta que el propuesto en este trabajo, la inyección mediante el intercambio de caracteres adyacentes, exitosa en inglés, resulta extraña en chino.

La evaluación revela que el índice de inserción no es estable y puede variar. Se plantea mejorar la tasa de inserción mediante modelos más potentes de procesamiento de lenguaje natural, la aplicación de esquemas de codificación y la exploración de diferentes tipos de errores tipográficos. Se reconocen limitaciones en cuanto a la potencia computacional requerida, la tasa de inserción limitada y la visibilidad de los errores tipográficos generados. Se proyecta futuras investigaciones centradas en mejorar la tasa de inserción, la invisibilidad de los errores tipográficos y evaluación de la robustez del método.Este trabajo cuenta con el respaldo de la Fundación Nacional de Ciencias Naturales de China (NSFC) bajo el número de concesión U1536207, el Programa Nacional de Investigación y Desarrollo Clave de China bajo el número 2016QY08D1600 y el Proyecto Nacional de Investigación y Desarrollo Clave de China bajo el número 2016YFB0801405.

## 9.5. ESTUDIOS RELACIONADOS EN ISRAEL

La inteligencia israelí, conocida por su agencia principal, el Mossad, es reconocida a nivel mundial por su eficacia y sofisticación en operaciones de inteligencia y seguridad. Fundada en 1949, el Mossad se dedica a la recopilación de información, operaciones encubiertas, contraespionaje y la protección de los intereses de Israel en el ámbito nacional e internacional. Las estructuras criptográficas y esteganográficas desempeñan un papel crucial en el ámbito de la inteligencia y la seguridad militar en Israel. La criptografía se utiliza para cifrar y proteger la información sensible contra accesos no autorizados, garantizando la confidencialidad de las comunicaciones. Israel ha desarrollado sistemas criptográficos avanzados para proteger sus comunicaciones y datos estratégicos, manteniendo un alto nivel de seguridad en sus operaciones.

En el entorno de los sistemas de seguridad militares, la inteligencia israelí ha demostrado una capacidad excepcional para adaptarse a las amenazas cambiantes y utilizar tecnologías avanzadas para mantener la ventaja estratégica. El enfoque integrado de la criptografía y la esteganografía es fundamental para salvaguardar la información vital en operaciones militares y de inteligencia, contribuyendo al éxito continuo de las agencias de seguridad de Israel en la protección de sus intereses nacionales.

**9.5.1. Universidad de Tel Aviv:** En el trabajo del año 2010 el Dr. Ofer Drori plantea el trabajo titulado "¿Toda la inteligencia de Israel se conoce entre sí? El desafío: Mejorar la cooperación de inteligencia entre las autoridades en el país" **כל שירותי המודיעין בישראל, האם הם מכירים זה את זה? האתגר: לשפר את שיתוף הפעולה המודיעיני בין הרשויות במדינה.**

Este documento fue preparado durante el tiempo de Drori como Investigador Senior en la Fundación de Investigación Goldman de la Escuela Hartog de Política y Gobierno de Tel Aviv en 2009-2010. La investigación se centra en asuntos de cooperación entre agencias de inteligencia en Israel, específicamente agencias civiles de inteligencia que se ocupan de la aplicación de la ley, la investigación y la inteligencia. El documento tiene como objetivo abordar el desafío de la cooperación eficiente, explorando la situación existente, las dificultades que surgen y proponiendo soluciones potenciales.

La situación actual se caracteriza por una cooperación limitada entre las agencias, lo que lleva al documento a presentar dos soluciones. Una es una solución organizativa que implica compartir información de manera eficiente mediante la definición de las necesidades de cada agencia. La otra es una solución tecnológica que propone la posibilidad de esfuerzos cooperativos a través de ubicaciones geográficas comunes, la circunvención legal de barreras y el uso de personal conjunto. El documento también sugiere diferentes niveles de cooperación y plantea que la transferencia de información puede ocurrir sin pasar advertencias físicas, ofreciendo un enfoque matizado. Recomienda la adopción de la solución tecnológica, considerando sus ventajas y la confianza que se puede construir con el tiempo.

La solución tecnológica, en esencia, combina el uso de la tecnología con la cooperación organizativa e interpersonal, convirtiéndola en la solución óptima. El documento enfatiza la necesidad de precaución en su implementación gradual, con la esperanza de que estas recomendaciones sean operativas para mejorar la cooperación de inteligencia entre las agencias en Israel.

Describir la situación actual en cada uno de los organismos y definir los problemas es fundamental para el éxito de las misiones de las organizaciones involucradas en inteligencia, investigaciones y aplicación de la ley. La creación de una imagen completa de inteligencia para los organismos que se ocupan de la inteligencia, investigaciones y aplicación de la ley es fundamental para el éxito de las tareas en el campo. La colaboración en la transferencia de información de inteligencia puede permitir la creación de una imagen completa y amplia para cualquier organización, algo casi imposible si cada organización intenta obtener la información por sí misma y basándose solo en sus propias fuentes. Además de la colaboración, existe una realidad diferente en la que no se cumplen las colaboraciones en el campo. Hay casos en los que se evitan colaboraciones en la transferencia de información de inteligencia de manera intencionada, y hay situaciones en las que la falta de transferencia de información se debe a defectos o incluso negligencia. Es precisamente por estas situaciones que considero de suma importancia basar la colaboración de inteligencia no solo en la conciencia de las personas, sino también en la tecnología que puede prevenir muchas irregularidades creadas por la naturaleza humana.

**9.5.2 Israel y la cooperación en medio de la crisis**: En los últimos tiempos, hemos sido testigos de situaciones extremas relacionadas, incluso, con la posible pérdida de vidas humanas como resultado de la falta o defecto en la colaboración de inteligencia entre organismos que deberían estar involucrados. En este contexto, mencionaré información breve sobre tres eventos diferentes en Israel y en el mundo que demuestran cuánto la falta de colaboración afecta a los países y sus ciudadanos.

El primer ejemplo se refiere al trabajo de los organismos de inteligencia y seguridad en Israel y a la falta de colaboración entre ellos que debería proteger la seguridad del país. En julio de 2009, el periódico Haaretz publicó un artículo titulado "Disputa entre el Mossad y el Shin Bet sobre áreas de responsabilidad", que presentaba hallazgos de un estudio realizado por un equipo dirigido por el exjefe del Shin Bet, el general Aharon Ze'evi-Farkash, unos meses antes. El estudio señaló que a pesar de la aprobación de las conclusiones del comité en 2004 y las directrices del primer ministro, no hubo una mejora significativa en la colaboración entre las agencias de inteligencia. El equipo, que operaba en el marco del Instituto de Estudios de Seguridad Nacional de la Universidad de Tel Aviv, agregó y señaló que "en ausencia de un líder para la comunidad de inteligencia y dada la limitada colaboración entre sus líderes, no hubo cambios significativos en el funcionamiento del sistema".

El segundo ejemplo se refiere a la actividad de la Autoridad Tributaria frente a los países del mundo, incluidas las empresas en la Organización para la Cooperación y el Desarrollo Económicos (OCDE), y todo de acuerdo con acuerdos firmados entre ellas y el Estado de Israel. En este contexto, se intercambia información entre los países miembros de la organización que han firmado acuerdos también en relación con las entradas de personas obligadas al impuesto en sus países de destino. La transferencia de información es requerida por los países de diversas maneras. Dado que Israel no transfiere información de manera proactiva, la mayoría de los países (especialmente Estados Unidos) no están dispuestos a transferir información a Israel debido a la falta de reciprocidad en la transferencia de materiales. El tema está bajo discusión con la intención de encontrar una solución, pero al momento de la publicación de este trabajo, la situación persiste.

El tercer ejemplo se refiere a un evento casi desastroso de la caída de un avión Northwest Airlines en un vuelo de Ámsterdam a Detroit durante la Navidad de 2010, llevado a cabo por un terrorista suicida nigeriano. Según informes de los medios de comunicación y también de la conferencia de prensa organizada por el expresidente estadounidense Obama, surge un fracaso sistémico entre las

autoridades de inteligencia y la aplicación de la ley en los Estados Unidos. Según la información publicada, la CIA tenía información sobre la intención del terrorista que estaba en manos del CIA desde abril y fue actualizada por el representante de la CIA unas semanas antes del evento. La información "reposó" en la agencia durante unas cinco semanas, pero no se transmitió a las autoridades responsables de la seguridad de los vuelos a Estados Unidos. Además, la información sobre otro terrorista que planeaba volar un avión, llamado "el nigeriano", no se conectó con la información del "nigeriano" por parte de su padre. También se descubrió que otros fallos en los sistemas de alerta (algunos de ellos probablemente israelíes) causaron que, a pesar de las indicaciones de alerta, como la ruta de vuelo no explicada del terrorista, la compra de un boleto caro poco antes de la salida, tener solo un bolso pequeño sin maleta y más, la información de alerta se perdió y desapareció de la vista de los responsables del área y no se transfirió al terreno para prevenir el evento. Para más detalles sobre el tema, consulte el Anexo 3.

En cuanto a la construcción de una imagen de inteligencia verdadera, diferentes tipos y formas de información contribuyen. En este trabajo, todos se referirán como "información de inteligencia", aunque algunos de ellos no sean "inteligencia" según la definición convencional utilizada por algunas organizaciones. Por ejemplo, en la Policía de Israel, la información de inteligencia en su sentido simple es la información recopilada por los centros de inteligencia, mientras que la información que subió al archivo de una investigación es información de investigación. Existen también otros tipos de información, como la información derivada de escuchas telefónicas o la información obtenida a través de datos de comunicación. Estos tipos de información se definen como archivos de información puntual que son esencialmente archivos de información de inteligencia y, cuando se combinan con investigaciones, se consideran archivos de investigación.

Se puede ver que los límites no son muy claros y, a veces, cambian a lo largo del tiempo. En relación con los organismos externos a la policía, la información sobre datos de comunicación en relación con el factor investigado por ellos es información de inteligencia en todos los aspectos. Para no ocuparnos de definiciones y semántica que no aclaran, en este trabajo se refiere a todos los archivos como archivos de inteligencia en cuanto a la necesidad de cooperación entre las agencias de inteligencia y las investigaciones en el país. Este trabajo se centra en la esencia de la cooperación y la colaboración en la transferencia de información y las formas de lograrlas de manera óptima. Los problemas principales en la situación actual son los siguientes:

- Problema de "falta de conocimiento": la existencia de información de inteligencia en manos de una entidad no se comunica en absoluto a la otra entidad.

- Desperdicio significativo de recursos debido a la recopilación redundante de la misma información por diferentes entidades.

- Ocurrencia de daños graves, tanto financieros como relacionados con la vida humana, derivados de la falta de transferencia de información vital entre las organizaciones.

A continuación, se presenta una visión general de la situación actual en las autoridades a las que se refiere este informe:

**Autoridad Tributaria:**La autoridad tributaria se ve impedida por ley de proporcionar información a cualquier entidad a menos que se obtenga un permiso específico del Ministro de Hacienda o del director de la autoridad tributaria.

**Policía:** La policía no está legalmente impedida de proporcionar información de inteligencia a entidades externas. Se aplican consideraciones especiales a ciertos bancos de datos definidos por la ley, como la base de datos del sistema legal penal o la base de datos de comunicaciones. La policía proporciona información a varias entidades, incluyendo la autoridad tributaria, el Instituto Nacional de Seguros, la Autoridad de Prohibición del Lavado de Dinero, la Autoridad de Instrumentos Financieros y el Servicio de Seguridad General.

**Autoridad de Instrumentos Financieros:** La Autoridad de Instrumentos Financieros no tiene restricciones legales para proporcionar información a entidades externas. Su única obligación es proporcionar información autorizada por el director de la Autoridad de Instrumentos Financieros.

**Autoridad de Prohibición del Lavado de Dinero:** Según la ley, la Autoridad de Prohibición del Lavado de Dinero puede proporcionar información a la policía y al Servicio de Seguridad General para prevenir el terrorismo. Además, puede proporcionar información a la policía relacionada con la propiedad intelectual. La autoridad recibe información de la autoridad tributaria y de cualquier entidad a través de solicitudes de información de la policía o el Servicio de Seguridad General.

**Autoridad de Restricciones Comerciales:** No hay restricciones legales para la Autoridad de Restricciones Comerciales al proporcionar información. La autoridad comparte información con la policía, la autoridad tributaria y la Autoridad de Instrumentos Financieros.

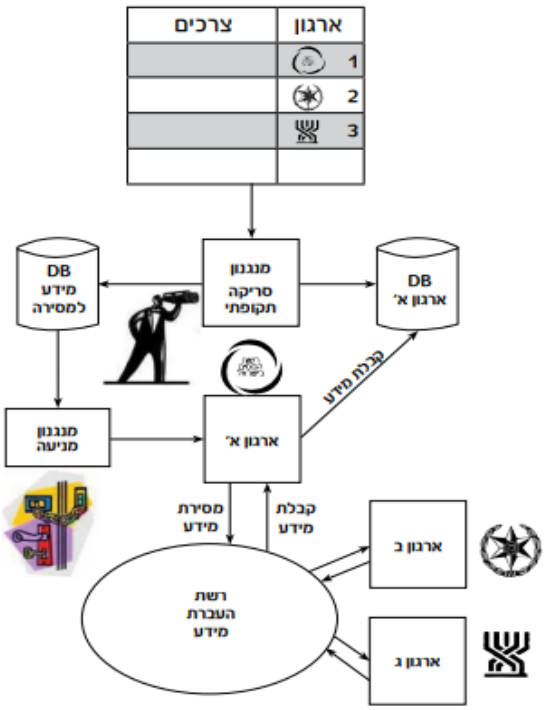
**Instituto Nacional de Seguros:** El Instituto Nacional de Seguros no tiene restricciones legales para proporcionar información. Proporciona información a la policía, la autoridad tributaria, la Autoridad de Instrumentos Financieros y el Servicio de Seguridad General.

**Servicio de Seguridad General:** El Servicio de Seguridad General puede recibir información de cualquier entidad y proporcionar información a la policía, la autoridad tributaria, la Autoridad de Instrumentos Financieros y la Autoridad de Prohibición del Lavado de Dinero. El informe destaca que la falta de cooperación efectiva entre las entidades gubernamentales crea problemas significativos y aboga por un equilibrio entre las necesidades de las autoridades y los derechos individuales.

**9.5.3 La transferencia de información entre entidades:** está sujeta a varias limitaciones, algunas de las cuales se mencionan anteriormente, y que podrían obstaculizar la transferencia automática de información entre estas entidades. Por otro lado, la verificación sistemática de las necesidades de la entidad A por parte de la entidad B, junto con la emisión de notificaciones correspondientes al organismo evaluador (es decir, la entidad B), permite que la transferencia de información a la entidad A se realice de manera controlada, cumpliendo con los principios éticos y organizativos. Aunque la transferencia de información puede ser altamente beneficiosa, los administradores de las entidades deben recordar que la transferencia de información sujeta a aprobación humana tiene limitaciones:

Es raro encontrar un único actor en grandes organizaciones de inteligencia que pueda tomar decisiones sobre la naturaleza de la transferencia de información debido a su incapacidad para ser experto en todos los componentes de la información organizativa. La distribución de la aprobación entre diferentes actores en la organización generará complejidades y, en su mayoría, retrasará la transferencia de información. En organizaciones grandes con un flujo sustancial de información, la dependencia de la transferencia manual de información por parte de una persona causará demoras significativas, incluso siendo impracticable. Incluso un experto humano puede cometer errores debido a la carga de información o errores humanos, lo que significa que incluso la aprobación

"manual" de la transferencia de información no está exenta de errores.

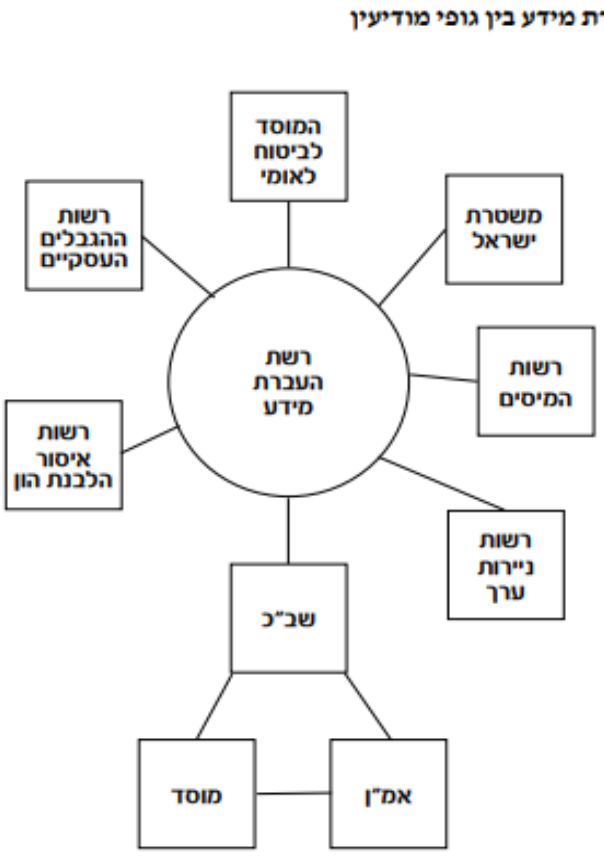



***Principales problemas organizacionales del Ministerio de Defensa Isaraelí.***

**Propuesta de solución:** Dadas las limitaciones mencionadas, la solución recomendada sería un enfoque tecnológico automatizado para la transferencia de información entre las entidades. Cada entidad que participa en el sistema debería gestionar los requisitos de las demás, creando y operando un mecanismo que escanee la información temporalmente según lo determine (ya sea de manera continua o diaria). Los resultados del escaneo generarán un conjunto de información o datos destinados a transferirse al organismo o entidades asociadas en el proyecto. El componente automatizado del mecanismo puede ser completamente controlado y monitoreado, desde un estado completamente automático hasta un estado completamente manual. En el modo "automático", la información recopilada en la organización emisora se transferirá después de un filtro automático a la organización receptora a través de una red de comunicación segura y un mecanismo financiero que se explicará a continuación. En el modo "manual", la información filtrada en la organización emisora se detendrá y sólo se transferirá después de la aprobación manual por parte de un representante de la organización emisora a la organización receptora.

Este concepto se puede implementar tecnológicamente utilizando diversas herramientas, siendo una opción la llamada "caja fuerte", donde de hecho existe un intermediario entre las computadoras de las organizaciones, y la información que se mueve de una organización a otra se realiza a través de esta "caja fuerte", que se puede configurar con diferentes reglas de comportamiento. El intermediario puede basarse en la red de Internet y también en líneas de comunicación privadas y encriptadas entre las organizaciones para satisfacer la necesidad de un nivel de seguridad más alto.

El mecanismo de intercambio de inteligencia entre las organizaciones propuestas se basa en una red de información dedicada que incluirá a las entidades colaboradoras, como la Policía de Israel, la Autoridad Tributaria, la Autoridad de Valores, la Unidad de Inteligencia Financiera, la Autoridad de Restricciones Comerciales, el Instituto Nacional de Seguros y el Shin Bet (Agencia de Seguridad de Israel). El Shin Bet también puede salir de esta red hacia una red de otras agencias de seguridad, como el Mossad y la Unidad de Inteligencia Militar del IDF.




***Croquis 1. Estrategia por parte de la mossad y la IDF.***

Cabe destacar que este proyecto implica la colaboración de siete entidades con un interés común en la transferencia de información (al menos en parte). Estas entidades son la Policía de Israel, la Autoridad Tributaria, la Autoridad de Valores, la Autoridad de Restricciones Comerciales, el Instituto Nacional de Seguros, la Unidad de Inteligencia Financiera y el Shin Bet. En ciertos casos, puede haber interés en agregar otras entidades, ya sean civiles o de seguridad. En resumen, la operación de un sistema como este requiere la gestión de varios componentes, como se detalla anteriormente: la definición de las necesidades de las organizaciones asociadas al proyecto. junto a un registro de las necesidades, escaneo de las necesidades frente a la información disponible, establecimiento de una estructura de base de datos para la entrega de información. Esta solución técnica proporciona una transferencia de información más eficiente y segura entre las organizaciones involucradas.

## 10. ESTEGANOGRAFÍA DIGITAL

En las últimas tres décadas, con la llegada de la digitalización asociada a las tecnologías de la información y de la comunicación, las posibilidades de la esteganografía se han multiplicado exponencialmente. Los contenidos multimedia digitales, tales como las imágenes, los vídeos o los archivos de audio, rápidamente se identificaron como portadores ideales para ocultar mensajes secretos que pudiesen pasar inadvertidos a los ojos de curiosos. Y no solo se usan contenidos multimedia como portadores de mensajes ocultos, sino que también los archivos de texto, el código fuente de software o los propios protocolos de Internet permiten crear canales esteganográficos encubiertos para establecer comunicaciones privadas sin que nadie repare en ello.

Sin embargo, los contenidos multimedia, por el elevado volumen de información que poseen, por su ubicuidad en toda la Red y por ser un tipo de archivos que pueden intercambiarse libremente entre usuarios sin despertar sospecha alguna, son el medio preferido para este tipo de aplicaciones. ¿Quién podría sospechar que las inocentes fotos de las vacaciones que alguien ha publicado en su cuenta de Instagram ocultan, en realidad, información clasificada que se está haciendo llegar de forma encubierta a un destinatario de un país lejano?

Más allá de las aplicaciones militares o de inteligencia, las técnicas esteganográficas permiten otros usos más o menos obvios. Por un lado, los disidentes en regímenes autoritarios donde se practica la censura o la persecución política, pueden usar la esteganografía para establecer comunicaciones encubiertas, evitando así el escrutinio de las autoridades. Por otro lado, con fines menos loables, la esteganografía se relaciona también con usos criminales o, incluso, terroristas. Comunicarse cuando se está sometido a una estrecha vigilancia es un reto muy complicado. Las autoridades tienen recursos y herramientas legales a su alcance para intervenir las comunicaciones, ya sean telefónicas, postales o telemáticas. Cuando un grupo de delincuentes o de terroristas sabe que los vigilan de cerca, la esteganografía se les presenta como una alternativa muy apetecible para proteger sus comunicaciones más delicadas.

Según Megias (2019) un caso relativamente reciente de este uso se registró en Berlín, en mayo de 2011, cuando un sospechoso de pertenecer a la banda terrorista Al Qaeda fue detenido por las autoridades alemanas. Al presunto terrorista se le incautó una tarjeta de memoria que contenía una carpeta protegida mediante contraseña. La policía científica alemana consiguió acceder a los contenidos de la carpeta y, para su sorpresa, solo hallaron en ella un vídeo con material pornográfico. Que tal archivo estuviese protegido por contraseña despertó las sospechas de las autoridades, que decidieron analizarlo con mayor detalle. De ese vídeo se extrajeron 141 archivos de texto ocultos que contenían información relevante sobre las operaciones de Al Qaeda y sus planes de futuro, bajo títulos tan inequívocos como “Trabajos futuros”, “Lecciones aprendidas” e “Informe de operaciones”. La esteganografía va más allá que la criptografía y pretende ocultar que la propia comunicación se está produciendo. La lista de amenazas conocidas no termina ahí. Al margen de los grupos de criminales o terroristas que usan la esteganografía como canal de comunicación encubierto también existen colectivos de ciberdelincuentes para los que estas herramientas son el mecanismo perfecto a través del cual desplegar sus ataques.

Entre 2011 y 2017 hay constancia de al menos catorce casos de malware (software malicioso) que han usado la esteganografía como herramienta infecciosa. En este caso, la esteganografía se utiliza en varios momentos del ataque: en primer lugar, cuando se está examinando al objetivo del ataque, para ocultar el escaneado; en segundo lugar, para obtener un acceso no autorizado, ocultando el proceso de infección o disfrazando aplicaciones maliciosas como inocentes; finalmente, también se

está usando para mantener en el tiempo un acceso no autorizado, ocultando el tráfico de datos y extrayendo, de manera encubierta, información del dispositivo afectado. A diferencia de la criptografía, que no oculta las comunicaciones, la esteganografía sí que lo hace, por lo que puede ser muy difícil, sino imposible, detectar este tipo de intrusiones con las herramientas de seguridad estándar.

En la actualidad, el malware que usa esteganografía a menudo se vale de contenidos digitales como portadores de la información. La técnica más habitual consiste en usar imágenes digitales para ocultar las configuraciones del software malicioso, para proporcionar una dirección de Internet desde la cual descargar componentes adicionales o, incluso, para ocultar directamente el código malicioso. Tampoco el secuestro de datos o ransomware ha quedado fuera de esta oleada y ya se han registrado varias infecciones que han utilizado imágenes o canales encubiertos en los protocolos de Internet para transmitir componentes del software de secuestro de datos, facilitando así la infección y dificultando la acción de las aplicaciones antimalware que detectan o bloquean este tipo de ataques.

En resumidas cuentas, parece que la tendencia al uso de la esteganografía, tanto para comunicaciones de grupos de delincuentes y terroristas como en el caso de la ciberdelincuencia para la propagación de infecciones de malware y ransomware, es algo que sí que debe de preocuparnos e instarlos a tomar las contramedidas oportunas. ¿Estamos, pues, indefensos? Por fortuna, no. La comunidad científica lleva décadas investigando las tecnologías de ocultación de la información y, entre ellas, la esteganografía. Quiere decir que ya se han desarrollado herramientas forenses que permiten, tanto a las autoridades como a los expertos en ciberseguridad de organizaciones y empresas, proteger sus sistemas y comunicaciones frente a este tipo de amenazas. Investigadores de todo el mundo centran su actividad en el desarrollo de nuevas técnicas de esteganografía y de sistemas de detección de anomalías que pueden usarse para discriminar si un objeto digital es solo lo que parece o, por el contrario, debe examinarse a fondo para determinar si contiene información oculta. Estas últimas técnicas se denominan estegoanálisis y constituyen la otra cara de la moneda de la esteganografía.

Los expertos en ciberseguridad y en técnicas forenses trabajamos para combatir las amenazas, para hacer de la Red un lugar algo menos salvaje Igual que para la criptografía existe el criptoanálisis, que trata de romper los sistemas criptográficos para descifrar la información secreta, en el ámbito de la esteganografía, y estegoanálisis consiste en analizar computacionalmente unos contenidos sospechosos para determinar si presentan algún tipo de desviación estadística respecto a sus análogos inocuos. En caso de hallar alguna anomalía, se analizará si ésta concuerda con alguna técnica esteganográfica concreta. Como es de suponer, el aprendizaje automático es, en estos momentos, uno de los mejores aliados para los analistas.

Nos hallamos, pues, inmersos en una suerte de “carrera armamentística” entre la esteganografía y el estegoanálisis, a la que podemos aplicar una sencilla analogía con los sistemas vivos. Igual que la naturaleza ha dotado a los organismos de un sistema inmunológico para combatir las infecciones biológicas, los expertos en ciberseguridad y en técnicas forenses trabajamos para combatir las amenazas que se valen de la esteganografía, creando nuevas soluciones de estegoanálisis y poniéndolas a disposición de las autoridades y de la sociedad para hacer de la Red un lugar algo menos salvaje. A causa de la emergencia de estas amenazas un conjunto de expertos de la comunidad científica, de las fuerzas de seguridad y de diferentes empresas y organizaciones de toda Europa, hemos impulsado la creación del grupo Criminal Use of Information Hiding. (CUING, Uso Criminal de la Ocultación de la Información) en colaboración con el European Cybercrime Centre (EC3) de la Europol.

Las actividades del grupo CUING se centran en crear conciencia sobre los usos maliciosos de estas técnicas, realizar un seguimiento del progreso del uso de estas tecnologías con fines delictivos, compartir inteligencia estratégica sobre las nuevas amenazas, trabajar conjuntamente para combatir estas amenazas, y educar y capacitar a los nuevos profesionales que se van a tener que enfrentar a este tipo de retos, tanto desde las autoridades como desde las organizaciones y empresas. Nuestra labor en este grupo, por lo tanto, sería comparable a la acción del sistema inmunológico de los seres vivos, realizando tareas constantes de vigilancia y supervisión, y ayudando a las autoridades y a los expertos en ciberseguridad a combatir los ataques en que los ciberdelincuentes intentan aprovechar las vulnerabilidades de los sistemas y las nuevas oportunidades que les brindan las técnicas de ocultación de la información.

**Semagrama:** Los semagramas ocultan información mediante el uso de símbolos o signos. Un semagrama visual utiliza objetos físicos cotidianos o de apariencia inocente, como garabatos o la ubicación de elementos en un escritorio o sitio web. Un semagrama de texto oculta mensajes modificando la apariencia del texto portador, como cambios sutiles en tamaño o tipo de fuente, adicionando espacios o florituras en letras o texto escrito a mano.

**Códigos abiertos:** Los códigos abiertos ocultan mensajes en portadores legítimos de manera no obvia para un observador desprevenido. El mensaje portador se denomina comunicación abierta, mientras que el mensaje oculto es comunicación encubierta. Esta categoría se subdivide en códigos de jerga y cifrados encubiertos. El código de jerga utiliza un lenguaje entendido por un grupo pero no por otros. Incluye warchalking (símbolos para indicar señales de red inalámbrica), terminología clandestina o conversaciones inocentes que transmiten significados especiales. Los códigos de referencia son un subconjunto, donde frases preestablecidas transmiten significado.

**Los cifrados encubiertos**: ocultan mensajes abiertamente en el medio portador para que puedan ser recuperados por quienes conozcan el secreto. Un cifrado de rejilla utiliza una plantilla para cubrir el mensaje, siendo las palabras en las aperturas el mensaje oculto. Un cifrado nulo oculta el mensaje según reglas preestablecidas, como "leer cada quinta palabra" o "mirar el tercer carácter de cada palabra". Con el aumento del almacenamiento y transmisión de datos en computadoras y redes, la esteganografía ha ingresado a la era digital. En este entorno, las aplicaciones permiten ocultar cualquier archivo binario en otro, siendo las imágenes y audio los soportes más comunes.

**10.1. Imagen y Audio Digital** Muchas técnicas de esteganografía digital comunes emplean imágenes gráficas o archivos de audio como medio portador. Es instructivo, entonces, revisar la codificación de imágenes y audio antes de discutir cómo funcionan la esteganografía y el estegoanálisis con estos portadores. La Figura 2 muestra el cubo de color RGB, un medio común para representar un color determinado mediante la intensidad relativa de sus tres colores componentes (rojo, verde y azul), cada uno con su propio eje (moreCrayons 2003). La ausencia de todos los colores produce el negro, que se muestra como la intersección del punto cero de los ejes tricolores. La mezcla de 100 por ciento rojo, 100 por ciento azul y la ausencia de verde forma magenta; el cian es 100 por ciento verde y 100 por ciento azul sin rojo; y 100 por ciento verde y 100 por ciento rojo sin azul se combinan para formar amarillo. El blanco es la presencia de los tres colores.

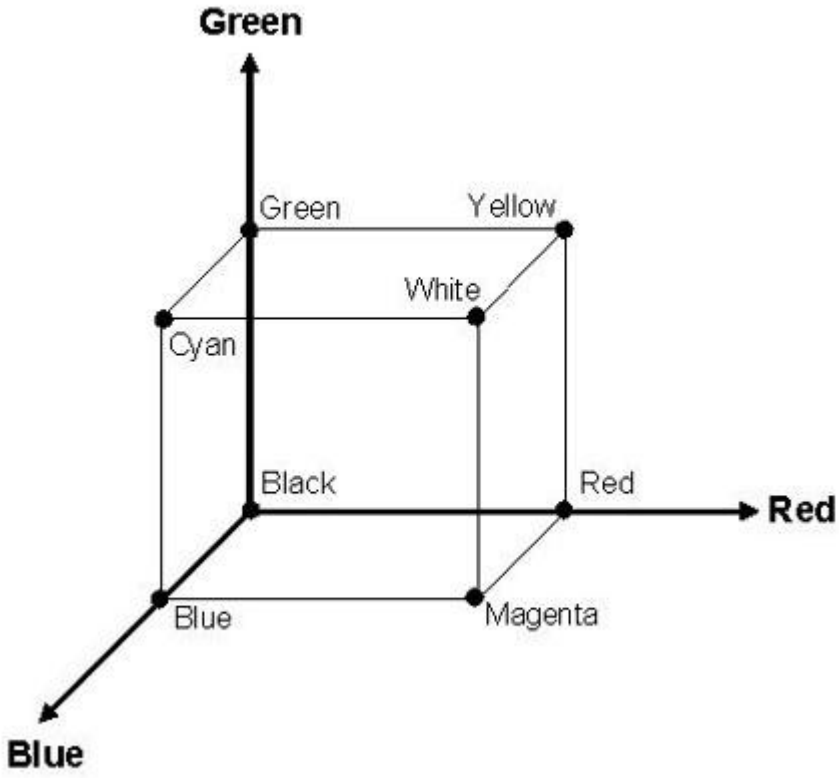



**Figura 2. El cubo de color RGB**

La Figura 3 muestra los niveles de intensidad RGB de algún color aleatorio. Cada componente RGB se especifica mediante un solo byte, de modo que los valores para cada intensidad de color pueden variar de 0 a 255. Este tono en particular se indica con un nivel de rojo de 191 (hex BF), un nivel de verde de 29 (hex 1D) y un nivel de azul de 152 (hex 98). Entonces, una imagen de magenta se modificaría con 24 bits, como 0xBF1D98. Este esquema de codificación de 24 bits admite 16.777.216 (2^24) colores únicos (Curran y Bailey 2003; Johnson y Jajodia 1998A).

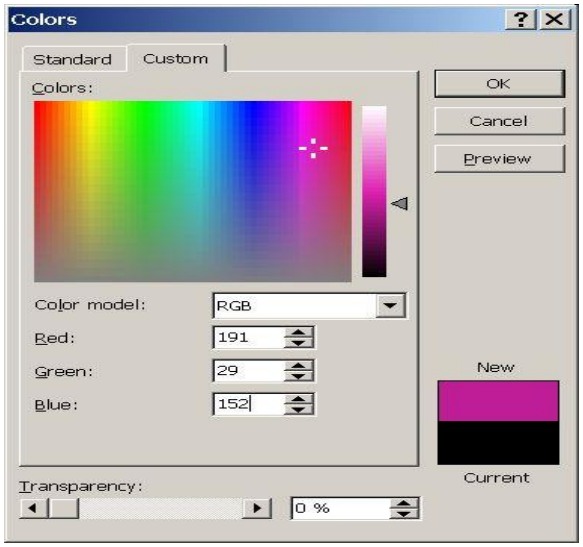



**Figura 3. Este cuadro de diálogo de selección de color muestra los niveles de rojo, verde y azul (RGB) de este color seleccionado.**

La mayoría de las aplicaciones de imágenes digitales actuales admiten color verdadero de 24 bits, donde cada elemento de la imagen (píxel) está codificado en 24 bits, que comprende los tres bytes RGB como se describió anteriormente. Otras aplicaciones codifican el color utilizando ocho bits/píxel. Estos esquemas también utilizan colores verdaderos de 24 bits, pero emplean una paleta que especifica qué colores se utilizan en la imagen. Cada imagen está codificada en ocho bits, donde el valor apunta a una entrada de color de 24 bits en la paleta. Este método limita el número único de colores en una imagen determinada a 256 (2^8). La elección de la codificación de colores obviamente afecta el tamaño de la imagen. Una imagen de 640 X 480 píxeles con color de ocho bits ocuparía aproximadamente 307 KB (640 X 480 = 307.200 bytes), mientras que una imagen de 1400 X 1050 píxeles con color verdadero de 24 bits requeriría 4,4 MB (1400 X 1050 X 3 = 4.410.000 bytes).

Las paletas de colores y el color de ocho bits se utilizan comúnmente con los formatos de imagen de formato de intercambio de gráficos (GIF) y mapa de bits (BMP). Generalmente se considera que GIF y BMP ofrecen compresión sin pérdidas porque la imagen recuperada después de la codificación y compresión es idéntica bit a bit a la imagen original (Johnson y Jajodia 1998A). El formato de imagen del Joint Photographic Experts Group (JPEG) utiliza transformaciones de coseno discretas en lugar de una codificación píxel por píxel. En JPEG, la imagen se divide en bloques de 8 X 8 para cada componente de color por separado. El objetivo es encontrar bloques donde la cantidad de cambio en los valores de los píxeles (la energía) sea baja. Si el nivel de energía es demasiado alto, el bloque se subdivide en 8 X 8 subbloques hasta que el nivel de energía sea lo suficientemente bajo. Cada bloque (o subbloque) de 8 X 8 se transforma en 64 coeficientes de transformada de coseno discretos que se aproximan a la luminancia (brillo, oscuridad y contraste) y crominancia (color) de esa parte de la imagen. Generalmente se considera que JPEG es una compresión con pérdida porque la imagen recuperada del archivo JPEG comprimido es una aproximación cercana, pero no idéntica, al original (Johnson y Jajodia 1998A; Monash University 2004; Provos y Honeyman 2003).

La codificación de audio implica convertir una señal analógica en un flujo de bits. El sonido analógico (voz y música) está representado por ondas sinusoidales de diferentes frecuencias. El oído humano puede oír frecuencias nominalmente en el rango de 20 a 20.000 ciclos/segundo (Hercios o Hz). El sonido es analógico, es decir, una señal continua. Almacenar el sonido digitalmente requiere que la onda sonora continua se convierta en un conjunto de muestras que puedan representarse mediante una secuencia de ceros y unos. La conversión de analógico a digital se logra muestreando la señal analógica (con un micrófono u otro detector de audio) y convirtiendo esas muestras a niveles de voltaje. Luego, el nivel de voltaje o señal se convierte en un valor numérico utilizando un esquema llamado modulación de código de pulso. El dispositivo que realiza esta conversión se denomina codificador-decodificador o códec.

La modulación de código de pulso proporciona solo una aproximación de la señal analógica original, como se muestra en la Figura 4. Si el nivel de sonido analógico se mide a un nivel de 4,86, por ejemplo, se convertiría a cinco en modulación de código de pulso. Esto se llama error de cuantificación. Diferentes aplicaciones de audio definen un número diferente de niveles de modulación de código de pulso para que este "error" sea casi indetectable para el oído humano. La red telefónica convierte cada muestra de voz a un valor de ocho bits (0-255), mientras que las aplicaciones de música generalmente utilizan valores de 16 bits (0-65.535) (Fries y Fries 2000; Rey 1983).

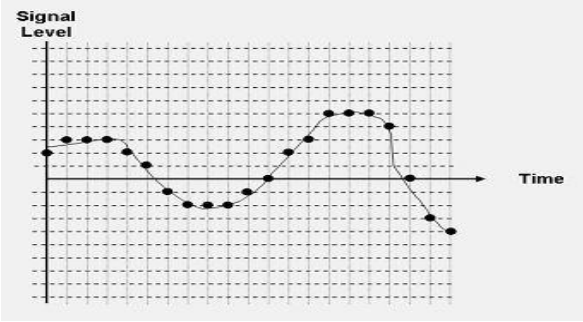



**Figura 4. Modulación de código de pulso simple**

Las señales analógicas deben muestrearse a una velocidad del doble del componente de frecuencia más alta de la señal para que el original pueda reproducirse correctamente sólo a partir de las muestras. En la red telefónica, la voz humana se transmite en una banda de frecuencia de 0 a 4.000 Hz (aunque en realidad sólo se utiliza entre 400 y 3.400 Hz para transmitir voz); por lo tanto, la voz se muestrea 8000 veces por segundo (una frecuencia de muestreo de 8 kHz). Las aplicaciones de audio musical asumen el espectro completo del oído humano y generalmente utilizan una frecuencia de muestreo de 44,1 kHz (Fries y Fries 2000; Rey 1983).

La velocidad de bits de la música sin comprimir se puede calcular fácilmente a partir de la frecuencia de muestreo (44,1 kHz), la resolución de modulación de código de pulso (16 bits) y el número de canales de sonido (dos) en 1.411.200 bits por segundo. Esto sugeriría que un archivo de audio de un minuto (sin comprimir) ocuparía 10,6 MB (1.411.200*60/8 = 10.584.000). De hecho, los archivos de audio se reducen mediante el uso de una variedad de técnicas de compresión. Un método obvio es reducir el número de canales a uno o reducir la frecuencia de muestreo, en algunos casos tan baja como 11 kHz. Otros códecs utilizan esquemas de compresión propietarios. Todas estas soluciones reducen la calidad del sonido.

**Tabla 1: Algunos formatos de audio digital comunes (Fries and Fries 2000)**

| Tipo de audio | Extensión de archivo | Códec |
|---|---|---|
| AIFF (Mac) | .aif, .aiff | Modulación de código de pulso (u otra) |
| AU (domingo/siguiente) | .au | ley μ (u otra) |
| CD de audio (CDDA) | n / A | Modulación de código de pulso |
| MP3 | .mp3 | Capa de audio MPEG III |
| Audio de Windows Media | .wma | propiedad de Microsoft |
| Tiempo rapido | .qt | Propiedad de Apple Computer |
| RealAudio | .ra, .ram | Propiedad de Real Networks |
| WAV | .wav | Modulación de código de pulso (u otra) |

**10.2 Métodos de portadores digitales:** Hay muchas formas en las que se pueden ocultar mensajes en los medios digitales. Los examinadores forenses digitales están familiarizados con los datos que permanecen en archivos inactivos o en espacios no asignados como restos de archivos anteriores, y se pueden escribir programas para acceder directamente al espacio inactivo y no asignado. También se pueden ocultar pequeñas cantidades de datos en la parte no utilizada de los encabezados de los archivos (Curran y Bailey 2003).

La información también se puede ocultar en un disco duro en una partición secreta. Una partición oculta no se verá en circunstancias normales, aunque la configuración del disco y otras herramientas pueden permitir un acceso completo a la partición oculta (Johnson et al. 2001). Esta teoría se ha implementado en un sistema de archivos esteganográfico ext2fs para Linux. Un sistema de archivos oculto es particularmente interesante porque protege al usuario de estar indisolublemente ligado a cierta información en su disco duro. Esta forma de negación plausible permite a un usuario afirmar que no está en posesión de cierta información o afirmar que ciertos eventos nunca ocurrieron. Bajo este sistema, los usuarios pueden ocultar la cantidad de archivos en la unidad, garantizar el secreto del contenido de los archivos y no alterar los archivos no ocultos mediante la eliminación del controlador de archivos de esteganografía (Anderson et al. 1998; Artz 2001; McDonald y Kuhn 2000). ).

Otro portador digital pueden ser los protocolos de red. El Protocolo de control de transmisión encubierto de Craig Rowland, por ejemplo, forma canales de comunicación encubiertos utilizando el campo de identificación en los paquetes del Protocolo de Internet o el campo de número de secuencia en los segmentos del Protocolo de control de transmisión (Johnson et al. 2001; Rowland 1996). Hay varias características del sonido que pueden alterarse de maneras que son imperceptibles para los sentidos humanos, y estas ligeras alteraciones, como pequeños cambios en el ángulo de fase, la cadencia del habla y la frecuencia, pueden transportar información oculta (Curran y Bailey 2003).

Sin embargo, los archivos de imagen y audio siguen siendo los medios portadores más fáciles y comunes en Internet debido a la gran cantidad de archivos portadores potenciales que ya existen, la capacidad de crear un número infinito de nuevos archivos portadores y el fácil acceso al software de esteganografía que operar en estos transportistas. Por esa razón, el enfoque del manuscrito volverá a los archivos de imagen y audio. El método de esteganografía más común en archivos de audio e imagen emplea algún tipo de sustitución o sobrescritura de bits menos significativos. El término de bit menos significativo proviene del significado numérico de los bits de un byte. El bit de orden superior o más significativo es el que tiene el valor aritmético más alto (es decir, $2^7=128$), mientras que el bit de orden inferior o menos significativo es el que tiene el valor aritmético más bajo (es decir, $2^0=1$). . Como ejemplo sencillo de sustitución del bit menos significativo, imagine "ocultar" el carácter 'G' en los siguientes ocho bytes de un archivo portador (los bits menos significativos están subrayados):

**1001010 1** **0000110 1** **1100100 1** **1001011 0**

**0000111 1** **1100101 1** **1001111 1** **0001000 0**

Una 'G' se representa en el Código estándar americano para el intercambio de información (ASCII) como la cadena binaria 01000111. Estos ocho bits se pueden "escribir" en el bit menos significativo de cada uno de los ocho bytes portadores de la siguiente manera:

**1001010 *0*** **0000110 1** **1100100 *0*** **1001011 0**

**0000111 *0*** **1100101 1** **1001111 1** **0001000 *1***

En el ejemplo anterior, solo se cambiaron la mitad de los bits menos significativos (que se muestran arriba en cursiva). Esto tiene cierto sentido cuando un conjunto de ceros y unos se sustituye por otro conjunto de ceros y unos. La sustitución de bits menos significativos se puede utilizar para sobrescribir codificaciones de color RGB legítimas o punteros de paleta en archivos GIF y BMP, coeficientes en archivos JPEG y niveles de modulación de código de pulso en archivos de audio. Al sobrescribir el bit menos significativo, el valor numérico del byte cambia muy poco y es menos probable que sea detectado por el ojo o el oído humanos.

La sustitución de bits menos significativa es una técnica simple, aunque común, de esteganografía. Sin embargo, su uso no es necesariamente tan simplista como parece el método. Sólo el software de esteganografía más ingenuo se limitaría a sobrescribir hasta el último bit menos significativo con datos ocultos. Casi todos utilizan algún tipo de medio para aleatorizar los bits reales del archivo portador que se modifican. Este es uno de los factores que dificulta la detección de la esteganografía. Otra forma de ocultar información en una imagen con paleta es alterar el orden de los colores en la paleta o utilizar la codificación de bits menos significativa en los colores de la paleta en lugar de en los datos de la imagen. Sin embargo, estos métodos son potencialmente débiles. Muchas herramientas de software de gráficos ordenan los colores de la paleta por frecuencia, luminancia u otro parámetro, y una paleta ordenada aleatoriamente se destaca en el análisis estadístico (Fridrich y Du 2000).

Siguen surgiendo métodos de esteganografía más nuevos y complejos. Los métodos de esteganografía de espectro ensanchado son análogos a las transmisiones de radio de espectro ensanchado (desarrolladas en la

Segunda Guerra Mundial y comúnmente utilizadas en los sistemas de comunicación de datos en la actualidad), donde la "energía" de la señal se distribuye a lo largo de un amplio espectro de frecuencias en lugar de centrarse en un solo espectro. frecuencia, en un esfuerzo por hacer más difícil la detección y la interferencia de la señal. La esteganografía de espectro extendido tiene la misma función: evitar la detección. Estos métodos aprovechan el hecho de que las pequeñas distorsiones en los archivos de imagen y sonido son menos detectables en las partes de alta energía del portador (es decir, alta intensidad en archivos de sonido o colores brillantes en archivos de imagen). Incluso cuando se ven uno al lado del otro, es más fácil engañar a los sentidos humanos cuando se realizan pequeños cambios en sonidos fuertes y/o colores brillantes (Wayner 2002).

**10.3. Ejemplos de esteganografía** Actualmente hay más de 100 programas de esteganografía disponibles, desde descargas gratuitas hasta productos comerciales. Esta sección mostrará algunos ejemplos simples de esteganografía ocultando un mapa GIF de 11,067 bytes del aeropuerto de Burlington, Vermont (Figura 5) en archivos GIF, JPEG y WAV.

**Figura 5. Este mapa está oculto en los distintos operadores de este artículo.**

El primer ejemplo emplea Gif-It-Up, un programa de Nelsonsoft que oculta información en archivos GIF utilizando la sustitución de bits menos significativa (e incluye una opción de cifrado). La Figura 6 muestra una imagen GIF del centro comercial de Washington, DC por la noche, donde se ha utilizado Gif-It-Up para insertar el mapa del aeropuerto que se muestra en la Figura 5. El soporte original tiene 632.778 bytes de longitud y utiliza 249 colores únicos, mientras que el El archivo de esteganografía tiene una longitud de 677,733 bytes y utiliza 256 colores únicos. El tamaño del archivo es mayor en el archivo de esteganografía debido a una opción de extensión de color utilizada para minimizar la distorsión en la imagen de esteganografía. Si no se utiliza la extensión de color, las diferencias en el tamaño del archivo son ligeramente menos perceptibles.

**Figura 6. Un archivo GIF de operador que contiene el mapa del aeropuerto**

**Figura 7. La paleta del archivo del transportista del centro comercial de Washington antes (izquierda) y después (derecha) de que se ocultara el archivo del mapa.**

La Figura 7 muestra las paletas del archivo portador antes y después de la inserción del mensaje. Como todos los programas de inserción de bits menos significativos que actúan sobre imágenes en color de ocho bits, Gif-It-Up modifica la paleta de colores y generalmente termina con muchos pares de colores duplicados.

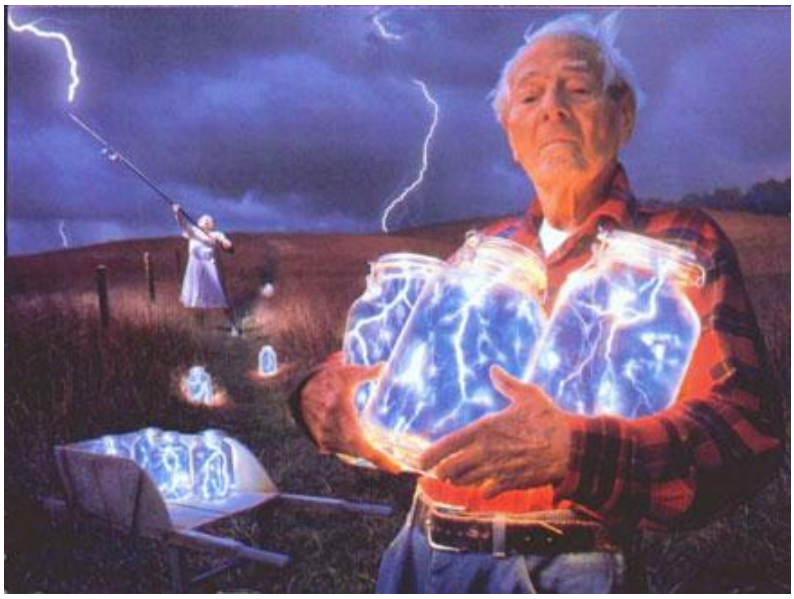


**Figura 8. Un archivo JPEG Carrier que contiene el mapa del aeropuerto**

JP Hide-&-Seek (JPHS) de Allan Latham está diseñado para usarse con archivos JPEG y compresión con pérdida. JPHS utiliza la sobrescritura de bits menos significativa de los coeficientes de transformación de coseno discretos utilizados por el algoritmo JPEG. El algoritmo criptográfico Blowfish se utiliza para la aleatorización y el cifrado de bits menos significativos (Johnson y Jajodia 1998B). La Figura 8 muestra un archivo JPEG de ejemplo con el mapa del aeropuerto incrustado. El archivo portador original tiene un tamaño de 207.244 bytes y contiene 224.274 colores únicos. El archivo de esteganografía tiene un tamaño de 207.275 bytes y contiene 227.870 colores únicos. No hay una paleta de colores para mirar porque JPEG utiliza codificación de colores de 24 bits y transformaciones de coseno discretas.

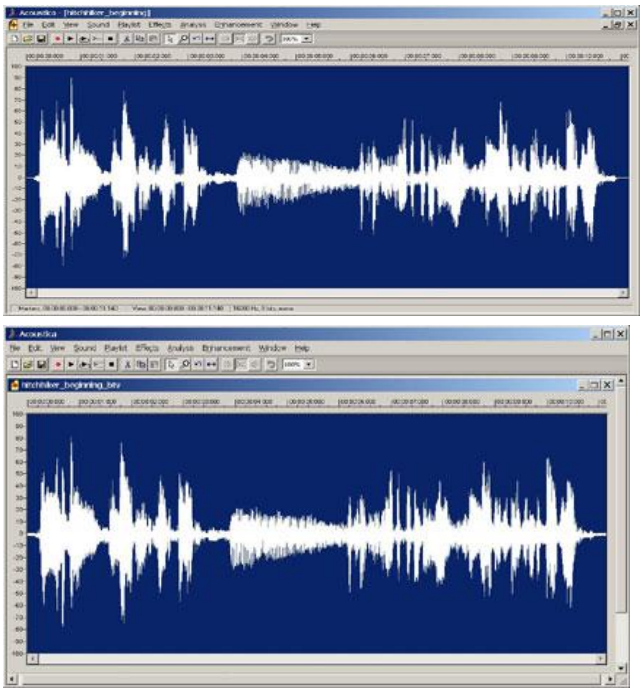


**Figura 9. Comparaciones del nivel de señal entre un archivo de operador WAV antes (arriba) y después (abajo) de insertar el mapa del aeropuerto.**

El último ejemplo emplea S-Tools, un programa de Andy Brown que puede ocultar información dentro de archivos GIF, BMP y WAV. S-Tools utiliza la sustitución de bits menos significativa en archivos que emplean compresión sin pérdidas, como modulación de código de impulsos y color de ocho o 24 bits. S-Tools emplea una contraseña para la aleatorización de bits menos significativa y puede cifrar datos utilizando el Estándar de cifrado de datos (DES), el Algoritmo internacional de cifrado de datos (IDEA), el Cifrado de resumen de mensajes (MDC) o Triple-DES (Johnson y Jajodia 1998A; Johnson y Jajodia 1998B; Wayner 2002). La Figura 9 muestra una comparación del nivel de señal entre un archivo de operador WAV antes y después de que se ocultara el mapa del aeropuerto. El archivo WAV original tiene una longitud de 178.544 bytes, mientras que el archivo WAV de esteganografía tiene una longitud de 178.298 bytes. Aunque el tamaño relativamente pequeño de la figura dificulta ver los detalles, se notan algunas diferencias al principio y al final de la muestra de audio (es decir, durante los períodos de silencio). (Algunas herramientas de esteganografía tienen inteligencia incorporada para evitar las partes de baja intensidad de la señal). Los archivos de audio son muy adecuados para ocultar información porque generalmente son relativamente grandes, lo que dificulta encontrar pequeños elementos ocultos.

Gif-It-Up, JPHS y S-Tools se utilizan arriba sólo con fines de ejemplo. Son gratuitos, fáciles de usar y realizan bien sus tareas. Hay muchos otros programas que se pueden utilizar para ocultar información en BMP, GIF, JPEG, MP3, Paintbrush (PCX), Portable Network Graphics (PNG), Tag Image File Format (TIFF), WAV y otros tipos de archivos portadores. El sitio web StegoArchive.Com tiene una muy buena lista de software gratuito, shareware y software de esteganografía comercial para DOS, Linux/Unix, MacOS, Windows y otros sistemas operativos (StegoArchive.com 2003).

Aunque la discusión anterior se ha centrado únicamente en archivos de imagen y audio, los medios de esteganografía no se limitan a este tipo de archivos. Otros tipos de archivos también tienen características que pueden aprovecharse para ocultar información. Hydan, por ejemplo, puede ocultar mensajes de texto en archivos ejecutables de OpenBSD, FreeBSD, NetBSD, Red Hat Linux y Windows XP. Desarrollado por Rakan El-Khalil, Hydan aprovecha la redundancia en el conjunto de instrucciones i386 e inserta información oculta definiendo conjuntos de instrucciones funcionalmente equivalentes, conceptualmente como una

imitación basada en gramática (por ejemplo, donde las instrucciones ADD son un bit cero y las instrucciones SUB son un poco). El programa puede ocultar aproximadamente un byte de mensaje por cada 110 bytes de instrucciones y mantiene el tamaño original del archivo de la aplicación. También se puede emplear el cifrado Blowfish (El-Khalil 2003).

**10.4. Detección de esteganografía**

El problema del prisionero (Simmons 1983) se utiliza a menudo para describir la esteganografía, aunque originalmente se introdujo para describir un escenario de criptografía. El problema involucra a dos prisioneros, Alice y Bob, que están encerrados en celdas separadas y desean comunicarse entre sí algún plan secreto. Alice y Bob pueden intercambiar mensajes entre ellos, pero William, el alcaide, puede leer todos los mensajes. Alice y Bob saben que William terminará las comunicaciones si descubre el canal secreto (Chandramouli 2002; Fridrich et al. 2003B).

William puede actuar de forma pasiva o activa. En el modelo de guardián pasivo, William examina cada mensaje y determina si reenviarlo o no en función de su capacidad para detectar un mensaje oculto. En el modelo de guardián activo, William puede modificar los mensajes si lo desea. Un guardián conservador o malicioso podría en realidad modificar todos los mensajes en un intento de interrumpir cualquier canal encubierto, de modo que Alice y Bob necesitarían utilizar un método de esteganografía muy robusto (Chandramouli 2002; Fridrich et al. 2003B).

La dificultad de la tarea del director dependerá en gran medida de la complejidad del algoritmo de esteganografía y de la cantidad de conocimientos previos de William (Chandramouli 2002; Fridrich et al. 2003B; Provos y Honeyman 2003). En un modelo de esteganografía pura, William no sabe nada sobre el método de esteganografía empleado por Alice y Bob. Ésta es una mala suposición por parte de Alice y Bob, ya que la seguridad mediante oscuridad rara vez funciona y es particularmente desastrosa cuando se aplica a la criptografía. Sin embargo, este es a menudo el modelo del analista forense digital que busca en un sitio web o en un disco duro el posible uso de la esteganografía.

La esteganografía de clave secreta supone que William conoce el algoritmo de esteganografía pero no conoce la esteganografía/clave criptográfica secreta empleada por Alice y Bob. Esto es consistente con la suposición que debe hacer un usuario de criptografía, según el Principio de Kerckhoff (es decir, "la seguridad del esquema criptográfico está en la gestión de claves, no en el secreto del algoritmo") (Kahn 1996). Sin embargo, esto también puede ser una suposición demasiado fuerte para la práctica, porque la información completa incluiría el acceso a la fuente del archivo portador.

El esteganálisis, la detección de la esteganografía por parte de un tercero, es una disciplina de investigación relativamente joven y pocos artículos aparecieron antes de finales de los años noventa. El arte y la ciencia del esteganálisis tienen como objetivo detectar o estimar información oculta basándose en la observación de algunas transferencias de datos y sin hacer suposiciones sobre el algoritmo de esteganografía (Chandramouli 2002). La detección de datos ocultos puede no ser suficiente. El esteganalista también puede querer extraer el mensaje oculto, desactivarlo para que el destinatario no pueda extraerlo y/o alterar el mensaje oculto para enviar información errónea al destinatario (Jackson et al. 2003). La detección y extracción de esteganografía generalmente es suficiente si el propósito es recopilar evidencia relacionada con un delito pasado, aunque la destrucción y/o alteración de la información oculta también pueden ser objetivos legítimos de aplicación de la ley durante una investigación en curso de grupos criminales o terroristas.

Las técnicas de esteganálisis se pueden clasificar de manera similar a los métodos de criptoanálisis, basándose en gran medida en la cantidad de información previa que se conoce (Curran y Bailey 2003; Johnson y Jajodia 1998B).

- Ataque exclusivo de esteganografía: el medio de esteganografía es el único elemento disponible para el análisis.
- Ataque de portador conocido: tanto el portador como el medio de esteganografía están disponibles para su análisis.
- Ataque de mensaje conocido: el mensaje oculto es conocido.

- Ataque de esteganografía elegido: tanto el medio de esteganografía como el algoritmo son conocidos.
- Ataque de mensaje elegido: se utiliza un mensaje conocido y un algoritmo de esteganografía para crear medios de esteganografía para futuros análisis y comparaciones.
- Ataque de esteganografía conocido: Se conocen el portador y el medio de esteganografía, así como el algoritmo de esteganografía.

Los métodos de esteganografía para medios digitales se pueden clasificar en términos generales como que operan en el dominio de la imagen o en el dominio de la transformación. Las herramientas de dominio de imagen ocultan el mensaje en el soporte mediante algún tipo de manipulación bit a bit, como la inserción del bit menos significativo. Las herramientas de dominio de transformación manipulan el algoritmo de esteganografía y las transformaciones reales empleadas para ocultar la información, como los coeficientes de transformada de coseno discreto en imágenes JPEG (Johnson y Jajodia 1998B).

De ello se deduce, entonces, que el estegoanálisis sigue en términos generales la forma en que funciona el algoritmo de esteganografía. Un método sencillo consiste en inspeccionar visualmente el soporte y los medios de esteganografía. Muchas herramientas de esteganografía simples funcionan en el dominio de la imagen y eligen bits de mensaje en el soporte independientemente del contenido del soporte. Aunque es más fácil ocultar el mensaje en el área de color más brillante o sonido más fuerte, es posible que el programa no busque esas áreas. Por lo tanto, la inspección visual puede ser suficiente para generar sospechas sobre un medio de esteganografía (Wayner 2002).

Un segundo enfoque es buscar rarezas estructurales que sugieran manipulación. La inserción de bits menos significativos en una imagen basada en una paleta a menudo provoca una gran cantidad de colores duplicados, dónde colores idénticos (o casi idénticos) aparecen dos veces en la paleta y difieren solo en el bit menos significativo. Los programas de esteganografía que ocultan información simplemente manipulando el orden de los colores en la paleta también provocan cambios estructurales. Los cambios estructurales a menudo crean una firma del algoritmo de esteganografía que se empleó (Jackson et al. 2003; Wayner 2002).

Las técnicas esteganográficas generalmente alteran las estadísticas del portador y, obviamente, los mensajes ocultos más largos alterarán al portador más que los más cortos (Farid 2001; Fridrich y Du 2000; Fridrich y Goljan 2002; Ozer et al. 2003). El análisis estadístico se emplea comúnmente para detectar mensajes ocultos, particularmente cuando el analista trabaja a ciegas (Jackson et al. 2003). Existe una gran cantidad de trabajo en el área del estegoanálisis estadístico. El análisis estadístico de archivos de imagen y audio puede mostrar si las propiedades estadísticas de los archivos se desvían de la norma esperada (Farid 2001; Ozer et al. 2003; Provos y Honeyman 2001). Estas llamadas estadísticas de primer orden (medias, varianzas, pruebas de chi-cuadrado ($X^2$)) pueden medir la cantidad de información redundante y/o distorsión en el medio. Aunque estas medidas pueden arrojar una predicción sobre si los contenidos han sido modificados o parecen sospechosos, no son definitivas (Wayner 2002).

El estegoanálisis estadístico se vuelve más difícil porque algunos algoritmos de esteganografía se esfuerzan por preservar las estadísticas de primer orden del archivo portador para evitar precisamente este tipo de detección. Cifrar el mensaje oculto también dificulta la detección porque los datos cifrados generalmente tienen un alto grado de aleatoriedad y los unos y los ceros aparecen con la misma probabilidad (Farid 2001; Provos y Honeyman 2001). La recuperación del mensaje oculto añade otra capa de complejidad en comparación con la simple detección de la presencia de un mensaje oculto. Recuperar el mensaje requiere conocimiento o una estimación de la longitud del mensaje y, posiblemente, una clave de cifrado y conocimiento del algoritmo criptográfico (Fridrich et al. 2003B).

Los algoritmos específicos del tipo de archivo portador pueden hacer que el análisis sea más sencillo. JPEG, en particular, ha recibido mucha atención en la investigación debido a la forma en que operan los diferentes algoritmos en este tipo de archivo. JPEG es un medio portador deficiente cuando se utiliza la inserción simple de bits menos significativos porque la modificación del archivo causada por la compresión JPEG facilita la tarea de detectar la información oculta (Fridrich y Du 2000). Existen varios algoritmos que ocultan

información en archivos JPEG y todos funcionan de manera diferente. JSteg incorpora secuencialmente los datos ocultos en los bits menos significativos, JP Hide&Seek; utiliza un proceso aleatorio para seleccionar los bits menos significativos, F5 utiliza una codificación matricial basada en un código Hamming y OutGuess conserva estadísticas de primer orden (Fridich et al. 2001; Fridich et al. 2002A; Fridrich et al. 2002B; Fridich et al. . 2003A; Provos y Honeyman 2001; Provos y Honeyman 2003).

Se han descrito pruebas estadísticas más avanzadas que utilizan estadísticas de orden superior, análisis lineal, campos aleatorios de Markov, estadísticas wavelet y más en archivos de imagen y audio (Farid 2001; Farid y Lyu 2003; Fridrich y Goljan 2002; Ozer et al. 2003). . La discusión detallada está más allá del alcance de este artículo, pero los resultados de esta investigación se pueden ver en algunas herramientas de detección de esteganografía. La mayoría de los estegoanálisis actuales se basan en firmas, de forma similar a los antivirus y los sistemas de detección de intrusiones. Los sistemas de estegoanálisis basados en anomalías apenas están comenzando a surgir. Aunque los primeros sistemas son precisos y robustos, los segundos serán más flexibles y podrán responder mejor rápidamente a las nuevas técnicas de esteganografía. Una forma de la llamada "detección de esteganografía ciega" distingue entre imágenes limpias y esteganográficas utilizando estadísticas basadas en la descomposición de ondas o el examen del espacio, la orientación y la escala en subconjuntos de la imagen más grande (Farid 2001; Jackson et al. 2003). .

Este tipo de estegoanálisis estadístico no se limita a archivos de imagen y audio. El programa Hydan conserva el tamaño del soporte original pero, al utilizar conjuntos de instrucciones "funcionalmente equivalentes", emplea algunas instrucciones que no se utilizan habitualmente. Esto abre a Hydan a la detección al examinar la distribución estadística de las instrucciones de un programa. Las versiones futuras de Hydan mantendrán la integridad del perfil estadístico de la aplicación original para defenderse de este análisis (El-Khalil 2003). La comunidad policial no siempre puede darse el lujo de saber cuándo y dónde se ha utilizado la esteganografía o el algoritmo que se ha empleado. Las herramientas genéricas que pueden detectar y clasificar la esteganografía son áreas donde la investigación aún está en sus inicios, pero ya están disponibles en herramientas de software, algunas de las cuales se describen en la siguiente sección (McCullagh 2001). Y el mismo ciclo se repite, como se ve en el mundo de las criptomonedas: el estegoanálisis ayuda a encontrar la esteganografía integrada pero también muestra a los escritores de nuevos algoritmos de esteganografía cómo evitar la detección.

**10.5 Herramientas para la detección de esteganografía:** Este artículo se centra claramente en el examinador informático forense en ejercicio más que en el investigador. Esta sección, entonces, mostrará algunos ejemplos de software actualmente disponible que puede detectar la presencia de programas de esteganografía, detectar archivos portadores sospechosos e interrumpir mensajes esteganográficamente ocultos. [15]

La detección de software de esteganografía en una computadora sospechosa es importante para el análisis forense posterior. Como muestra la investigación, muchos programas de detección de esteganografía funcionan mejor cuando hay pistas sobre el tipo de esteganografía que se empleó en primer lugar. Encontrar software de esteganografía en una computadora daría lugar a la sospecha de que en realidad existen archivos de esteganografía con mensajes ocultos en la computadora sospechosa. Además, el tipo de software de esteganografía encontrado afectará directamente cualquier estegoanálisis posterior (por ejemplo, S-Tools podría dirigir la atención a archivos GIF, BMP y WAV, mientras que JP Hide-&-Seek podría indicarle al analista que mire más de cerca los archivos JPEG). ). El software Gargoyle (anteriormente StegoDetect) de WetStone Technologies (WetStone Technologies 2004) se puede utilizar para detectar la presencia de software de esteganografía. Gargoyle emplea un conjunto de datos patentado (o conjunto de hash) de todos los archivos en las distribuciones de software de esteganografía conocidas, comparándolos con los hashes de los archivos sujetos a búsqueda. La Figura 10 muestra el resultado cuando Gargoyle apuntaba a un directorio donde se almacenan los programas de esteganografía. Los conjuntos de datos de Gargoyle también se pueden utilizar para detectar la presencia de criptografía, mensajería instantánea, registro de claves, caballos de Troya, descifrado de contraseñas y otro software nefasto.

[15] Esto no es de ninguna manera una encuesta de todas las herramientas disponibles, sino un ejemplo de las capacidades disponibles. StegoArchive.com enumera muchos programas de esteganálisis (StegoArchive.com 2003).

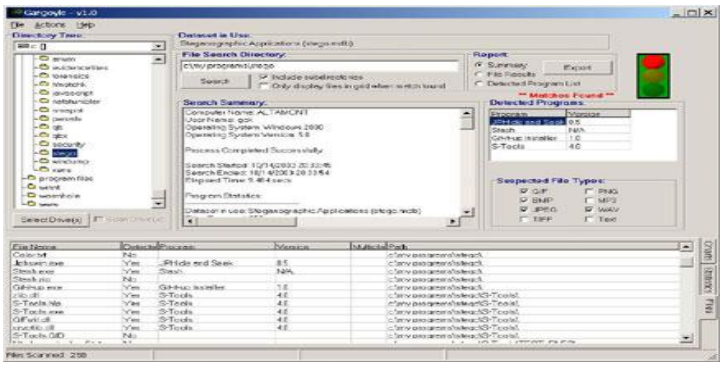


**Figura 10. Salida de Gargoyle cuando se dirige a uno de los directorios de un disco duro**

Forensic Toolkit de AccessData (AccessData 2003) y EnCase de Guidance Software (Guidance Software 2003) pueden utilizar los conjuntos de hash HashKeeper (Hashkeeper 2003), Maresware (Maresware 2003) y National Software Reference Library (National Software Reference Library 2003) para buscar una gran cantidad de datos. variedad de software. En general, estos conjuntos de datos están diseñados para excluir hashes de archivos "buenos" conocidos de los índices de búsqueda durante el análisis forense informático. Gargoyle también puede importar estos conjuntos de hash.

La detección de software de esteganografía sigue siendo más difícil por otra razón: el pequeño tamaño del software junto con la creciente capacidad de almacenamiento de los medios extraíbles. S-Tools, por ejemplo, requiere menos de 600 KB de espacio en disco y se puede ejecutar directamente, sin instalación adicional, desde un disquete o una memoria USB. En estas circunstancias, no se encontrarían restos del programa en el disco duro.

La segunda función importante del software de detección de esteganografía es encontrar posibles archivos portadores. Idealmente, el software de detección también proporcionaría algunas pistas sobre el algoritmo de esteganografía utilizado para ocultar información en el archivo sospechoso, de modo que el analista pueda intentar recuperar la información oculta.

Un programa de detección comúnmente utilizado es stegdetect de Niels Provos. Stegdetect puede encontrar información oculta en imágenes JPEG utilizando esquemas de esteganografía como F5, Invisible Secrets, JPHide y JSteg (OutGuess 2003). La Figura 11 muestra el resultado de xsteg, una interfaz gráfica para stegdetect, cuando se utiliza para examinar dos archivos en un disco duro: el soporte original y la imagen de esteganografía para la imagen JPEG que se muestra en la Figura 8. Tenga en cuenta que el archivo de esteganografía no solo está marcado como contiene información oculta, pero el programa también sugiere (correctamente) el uso del esquema de esteganografía JPHide.

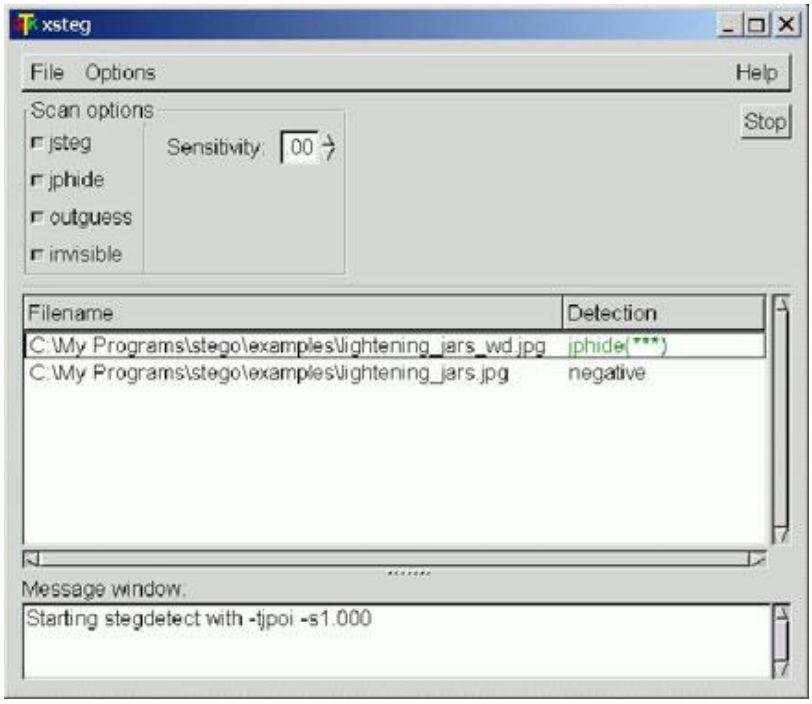



**Figura 11. El resultado de Xsteg al examinar dos archivos JPEG sospechosos**

Stego Watch de WetStone Technologies (WetStone Technologies 2004) analiza un conjunto de archivos y proporciona una probabilidad sobre cuáles son medios de esteganografía y el algoritmo probable utilizado para ocultarlos (lo que, a su vez, proporciona pistas sobre el software más probable empleado). El análisis utiliza una variedad de pruebas estadísticas seleccionables por el usuario basadas en las características del archivo portador que podrían verse alteradas por los diferentes métodos de esteganografía. Conocer el software de esteganografía disponible en la computadora sospechosa ayudará al analista a seleccionar las pruebas estadísticas más probables.

La Figura 12 muestra el resultado de Stego Watch cuando se apunta al archivo portador JPEG que se muestra en la Figura 8. La sección Algoritmos de detección de esteganografía en la pantalla muestra los algoritmos estadísticos empleados para el análisis y los que dieron frutos para esta imagen. Como se indicó anteriormente, Stego Watch identifica correctamente el software de esteganografía JPEG que se empleó.

Aunque aún no está disponible, el Instituto de Estudios de Tecnología de Seguridad de Dartmouth College ha desarrollado un software capaz de detectar datos ocultos en archivos de imágenes utilizando modelos estadísticos independientes del formato de la imagen o de la técnica de esteganografía. Este programa ha sido probado en 1.800 imágenes y cuatro algoritmos de esteganografía diferentes y fue capaz de detectar la presencia de mensajes ocultos con una precisión del 65 por ciento con una tasa de falsos positivos inferior al 0,001 por ciento (Dartmouth College 2003).

**Figura 12. Información de Stego Watch sobre un archivo JPEG que se sospecha que es un portador de esteganografía**

Encontrar esteganografía en un archivo que se sospecha que la contiene es relativamente fácil en comparación con extraer datos ocultos. La mayoría del software de esteganografía utiliza contraseñas para mantener el secreto, la aleatorización y/o el cifrado. Stegbreak, un programa complementario de stegdetect, utiliza un ataque de diccionario contra JSteg-Shell, JPHide y OutGuess para encontrar la contraseña de los datos ocultos pero, nuevamente, esto sólo se aplica a archivos JPEG (OutGuess 2003). De manera similar, Stego Break es un programa complementario de Stego Watch de WetStone que utiliza un ataque de diccionario en archivos sospechosos (WetStone Technologies 2004). Los esquemas de detección de esteganografía no ayudan directamente a recuperar la contraseña. Encontrar pistas apropiadas es donde entra en juego el resto de la investigación y la informática forense.

Un examinador forense informático que analiza pruebas en un caso penal probablemente no tenga motivos para alterar ningún archivo de pruebas. Sin embargo, un examen que forma parte de una vigilancia terrorista en curso bien podría querer alterar la información oculta incluso si no se puede recuperar. El contenido oculto, como la esteganografía y las marcas de agua digitales, puede atacarse de varias maneras para eliminarlo o modificarlo (Hernández Martin y Kutter 2001; Voloshynovskiy et al. 2001), y existe software diseñado específicamente para atacar las marcas de agua digitales. Dichos ataques tienen uno de dos efectos posibles: reducen la capacidad de carga de esteganografía del portaaviones (necesaria para evitar el ataque) o desactivan por completo la capacidad del portaaviones como medio de esteganografía.

Aunque este tema también está fuera del alcance de este artículo, se puede utilizar un ejemplo interesante de software de disrupción de esteganografía para cerrar esta discusión. 2Mosaic de Fabien Petitcolas emplea el llamado "ataque de presentación" principalmente contra imágenes de un sitio web. 2Mosaic ataca un sistema de marca de agua digital cortando una imagen en subimágenes más pequeñas. En el sitio web, la serie de imágenes pequeñas se colocan una al lado de la otra y aparecen igual que la imagen grande original (Petitcolas 2003).

**Figura 13. Una parte de la imagen JPEG con el mapa del aeropuerto oculto, creada por 2Mosaic**

La Figura 13 muestra un ejemplo de 2Mosaic cuando se usa contra la imagen JPEG de la Figura 8. En este caso, el archivo portador se divide en 165 subimágenes como se muestra arriba (11 filas de 15 subimágenes). El enfoque de 2Mosaic es obvio cuando se utiliza. El espectador de la imagen alterada sabe inmediatamente que algo anda mal.

Considere el siguiente escenario hipotético. Por acuerdo previo con miembros de una organización terrorista, el líder de la célula terrorista pone a la venta un artículo en eBay todos los lunes y publica una fotografía del mismo. El artículo a la venta es legítimo. Se aceptan ofertas, se recauda el dinero y los artículos se entregan diligentemente. Pero en algún momento preestablecido durante la semana, se publica una versión de la fotografía que contiene un mensaje oculto. Los miembros de la célula saben cuándo es ese momento y descargan el mensaje semanal. A menos que las personas estén bajo investigación activa, no está claro que alguien se dé cuenta de esta actividad. Este escenario, o uno similar, es un método viable para que terroristas o criminales se comuniquen, pero ¿es real? Después del 11 de septiembre de 2001, aparecieron varios artículos que sugerían que los terroristas de Al Qaeda empleaban la esteganografía (Kelly 2001; Kolata 2001; Manoo 2002; McCullagh 2001). En respuesta parcial a estos informes, se han hecho varios intentos de determinar la presencia de imágenes de esteganografía en Internet. Un conocido estudio buscó más de tres millones de imágenes JPEG en archivos de eBay y USENET.

Usando stegdetect, se encontró que entre uno y dos por ciento de las imágenes eran sospechosas, pero no se recuperaron mensajes ocultos usando stegbreak (Provos y Honeyman 2001; Provos y Honeyman 2003). Otro estudio examinó varios cientos de miles de imágenes de un conjunto aleatorio de sitios web y, utilizando también stegdetect y stegbreak, obtuvo resultados similares (Callinan y Kemick 2003). Aunque estos proyectos proporcionan un marco para buscar imágenes de esteganografía en un sitio web, no se pueden sacar conclusiones sobre las imágenes de esteganografía en Internet. En primer lugar, stegdetect sólo analiza imágenes JPEG. Otros tipos de imágenes nunca fueron examinados. En segundo lugar, se examinó un número limitado de sitios web, demasiado pocos para hacer afirmaciones definitivas sobre Internet en su conjunto. También es interesante observar que varios investigadores de esteganografía no publican deliberadamente información sobre qué sitios de Internet están examinando o qué están encontrando (Kolata 2001; McCullagh 2001). Hay pocas estadísticas concretas sobre la frecuencia con la que los funcionarios encargados de hacer cumplir la ley descubren software o medios de esteganografía en el curso de análisis forenses informáticos. Sin embargo, la evidencia anecdótica sugiere que muchos examinadores forenses informáticos no buscan rutinariamente software de esteganografía, y muchos podrían no reconocer dichas herramientas si las encontraran.

Además, las herramientas que se emplean para detectar software de esteganografía a menudo son inadecuadas, y el examinador frecuentemente depende únicamente de conjuntos de hash o de las propias herramientas de esteganografía (Kruse y Heiser 2001; Nelson et al. 2003; Security Focus 2003). Una búsqueda exhaustiva de evidencia de esteganografía en un disco duro sospechoso que podría contener miles de imágenes, archivos de audio y videoclips podría llevar días (Hosmer y Hyde 2003). De hecho, muchos examinadores forenses digitales consideran que la búsqueda de herramientas de esteganografía y/o medios de esteganografía es una parte rutinaria de cada examen (Security Focus 2003). Pero lo que parece faltar es un conjunto de directrices que proporcionen un enfoque sistemático para la detección por esteganografía. Incluso las directrices de búsqueda e incautación de pruebas digitales del Departamento de Justicia de Estados Unidos apenas mencionan la esteganografía (Departamento de Justicia de Estados Unidos, 2001; Departamento de Justicia de Estados Unidos, 2002).

El estegoanálisis será sólo una parte de una investigación; sin embargo, un investigador podría necesitar pistas de otros aspectos del caso para orientarlos en la dirección correcta. Un examinador forense informático podría sospechar el uso de esteganografía debido a la naturaleza del delito, los libros en la biblioteca del sospechoso, el tipo de hardware o software descubierto, grandes conjuntos de imágenes aparentemente duplicadas, declaraciones hechas por el sospechoso o los testigos, u otros factores. . Un sitio web puede resultar sospechoso por la naturaleza de su contenido o la población a la que sirve. Estos mismos elementos también podrían darle al examinador pistas sobre las contraseñas. Y la búsqueda de esteganografía no sólo es necesaria en investigaciones criminales y operaciones de recopilación de inteligencia. Los investigadores de contabilidad forense se están dando cuenta de la necesidad de buscar esteganografía, ya que ésta se convierte en una forma viable de ocultar registros financieros (Hosmer y Hyde 2003; Seward 2003).

Es imposible saber qué tan extendido está el uso de la esteganografía por parte de criminales y terroristas (Hosmer y Hyde 2003). Sin embargo, la verdad actual puede que ni siquiera importe. Es seguro que el uso de la esteganografía aumentará y será un obstáculo cada vez mayor para las actividades policiales y antiterroristas. Ignorar la importancia de la esteganografía debido a la falta de estadísticas es "seguridad a través de la negación" y no es una buena estrategia. La esteganografía no se encontrará si no se busca. Hay algunos informes de que los terroristas de Al Qaeda utilizaron la pornografía como medio de esteganografía (Kelly 2001; Manoo 2002). La esteganografía y la pornografía pueden ser tecnológica y culturalmente inesperadas para ese adversario en particular, pero demuestran la capacidad de trabajar "fuera de lo común". En las investigaciones informáticas, nosotros también debemos pensar e investigar de forma creativa.

## FORMAS BÁSICAS DE ESTEGANOGRAFÍA DIGITAL

| Criterio/Método | Esteganografía | Marca de Agua | Cifrado |
|---|---|---|---|
| Portador | Cualquier medio digital | Principalmente archivos de imagen/audio | Principalmente basado en texto, con algunas extensiones a archivos de imagen |
| Datos Secretos | Carga útil | Marca de agua | Texto plano |
| Clave | Opcional | Necesaria | N/A |

| | | | |
|---|---|---|---|
| Archivos de Entrada | Al menos dos a menos que sea antiincrustante | Uno | N/A |
| Detección | Ciega | Generalmente informativa (es decir, se necesita la cubierta original o la marca de agua para la recuperación) | Ciega |
| Autenticación | Recuperación completa de datos | Generalmente lograda mediante correlación cruzada | Recuperación completa de datos |
| Objetivo | Comunicación secreta | Preservación de derechos de autor | Protección de datos |
| Resultado | Archivo estego | Archivo con marca de agua | Texto cifrado |
| Preocupación | Detectabilidad/capacidad | Robustez | Robustez |
| Tipo de Ataques | Estegoanálisis | Procesamiento de imagen | Criptoanálisis |
| Visibilidad | Nunca | A veces (ver Fig. 2) | Siempre |

| | | | |
|---|---|---|---|
| Falla cuando | Es detectado | Es eliminado/reemplazado | Es descifrado |
| Relación con la cubierta | No necesariamente relacionado con la cubierta. El mensaje es más importante que la cubierta | Suele convertirse en un atributo de la imagen de la cubierta. La cubierta es más importante que el mensaje | N/A |
| Flexibilidad | Libre para elegir cualquier cubierta adecuada | Elección de la cubierta está restringida | N/A |
| Historia | Muy antigua excepto su versión digital | Era moderna | Era moderna |

## 11. ESTEGANOGRAFÍA DIGITAL AVANZADA

**11.1 Técnica ADN** : Sin duda, uno de los avances más significativos en el área de la esteganografía se basa en el contexto de la tecnología de Ingeniería celular. En el estudio del 11 de junio del 2020 propuesto por Dokyun Na (나도균) del Departamento de Ingeniería Biomédica, en la Universidad Chung-Ang en Seúl propone un estudio sobre "Esteganografía de ADN: ocultar mensajes secretos indetectables dentro de los polimorfismos de un solo nucleótido de un genoma y detectar errores inducidos por mutaciones".

En el estudio llevado a cabo por Na (2020), se diseñó una metodología innovadora de esteganografía de ADN con el propósito de ocultar mensajes de manera eficaz en las regiones variables, específicamente en los polimorfismos de un solo nucleótido, del genoma. Este enfoque se presenta como una medida estratégica para prevenir la piratería y garantizar la seguridad de la información transmitida. Asimismo, se implementó un algoritmo de verificación de suma en bloque, análogo al utilizado en la transmisión de datos de redes, con el fin de detectar posibles errores o mutaciones dentro de los mensajes cifrados. Este método no solo busca la confidencialidad de los mensajes ocultos, sino también la integridad de los mismos, brindando así una solución integral en el ámbito de la esteganografía de ADN (Na, 2020).

En el contexto de los avances en biología sintética e ingeniería metabólica, la creación de células diseñadas con aplicaciones industriales ha ganado prominencia, considerándose activos biológicos que requieren protección (Na, 2020). Ante esta necesidad, los investigadores han explorado métodos para marcar estas células con "marcas de agua", buscando salvaguardar su integridad y seguridad (Na, 2020). Los enfoques convencionales de marca de agua de ADN implican el cifrado de mensajes en secuencias de ADN, ya sea como códigos de barras insertados en el genoma o mezclados con fragmentos no relacionados, con la posterior capacidad de descifrar mediante técnicas como la reacción en cadena de la polimerasa (PCR) o electroforesis (Na, 2020).

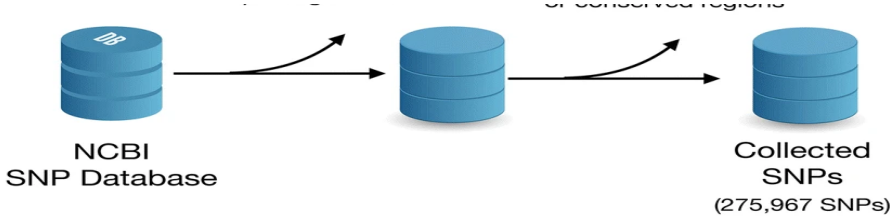
El ADN, con sus secuencias cuaternarias, ha captado un interés significativo como portador de información para almacenamiento, resolución de problemas y cifrado de mensajes (Na, 2020). La criptografía de ADN, en la cual los mensajes se cifran utilizando secuencias de ADN, ha sido empleada para ocultar información sensible (Na, 2020). No obstante, métodos anteriores, como el propuesto por Clelland et al., que inserta secuencias cifradas en el genoma humano, pueden resultar vulnerables con la tecnología actual de secuenciación de nueva generación (NGS), lo que impulsa la búsqueda de enfoques más robustos (Na, 2020).

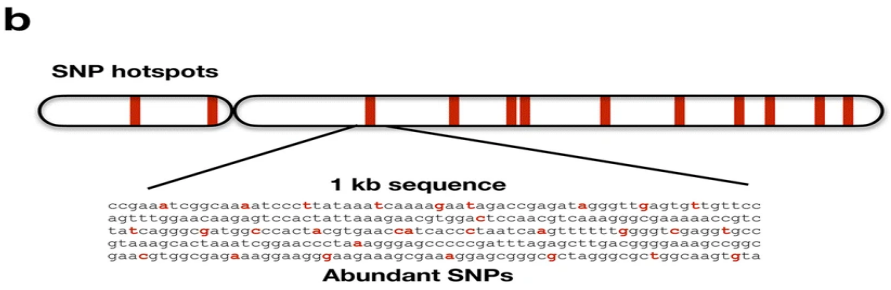
En respuesta a esta necesidad, el estudio presente propone una metodología innovadora de esteganografía de ADN, centrada en ocultar mensajes en regiones variables, específicamente polimorfismos de un solo nucleótido (SNP) en el genoma (Na, 2020). Este enfoque cifra el mensaje en secuencias de ADN e inserta estas secuencias en las regiones SNP, aprovechando la variabilidad natural de estas regiones para dificultar la detección de mensajes cifrados (Na, 2020). Para abordar las limitaciones asociadas con las mutaciones en el ADN, se emplea un algoritmo de verificación de suma en bloque, utilizado en la transmisión de datos, para detectar y corregir errores mutacionales, proporcionando así una solución más robusta y duradera para la esteganografía de ADN (Na, 2020). Este enfoque se posiciona como una herramienta valiosa para la marcación y protección de células diseñadas en aplicaciones industriales y de gestión (Na, 2020).

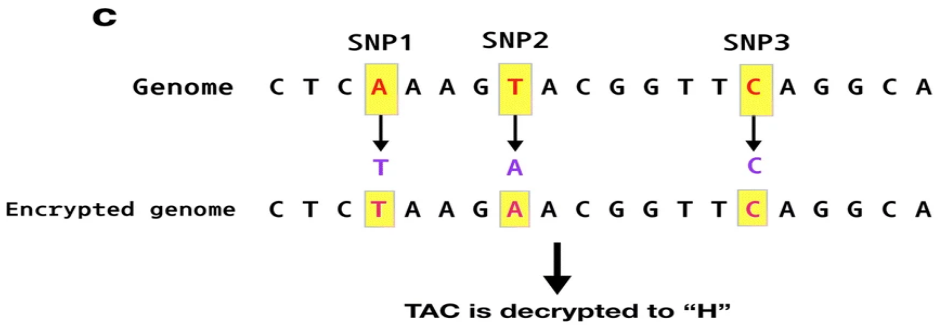
### 11.2 **Identificación de SNP y puntos de acceso de SNP**

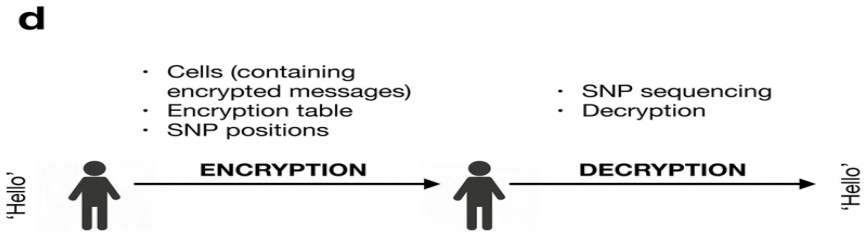
En la fase inicial de la investigación, se procedió a descargar la base de datos dbSNP (compilación 153) del NCBI con el objetivo de identificar regiones polimórficas dentro del genoma humano susceptibles de ser utilizadas para ocultar texto cifrado. La particularidad del genoma humano, en comparación con otros organismos, radica en la abundancia de SNP descubiertos, proporcionando así una rica fuente de información para la aplicación de esteganografía de ADN. Durante este proceso, se llevó a cabo una cuidadosa selección de SNP, permitiendo la presencia de uno de los cuatro nucleótidos (A/T/G/C) en la posición correspondiente, y descartando aquellos que se asociaban con patógenos (Tabla 1, Fig. 1a). Posteriormente, se eligieron las secuencias circundantes a los SNP, específicamente secuencias de 21 nucleótidos de longitud alrededor de los SNP, asegurándose de su singularidad en el genoma humano. Se excluyeron SNP ubicados dentro de elementos transponibles, islas CpG o regiones conservadas, utilizando bases de datos como Dfam para la identificación de elementos transponibles, Sequence Manipulation Suite para prever regiones de islas CpG, y las puntuaciones de conservación calculadas por PhastCons para identificar SNP en regiones conservadas con potencial impacto fenotípico. El resultado final de este proceso riguroso fue la identificación de 275,967 SNP seleccionados (Ver Tabla 1).

**Mensaje encriptado en lengua inglesa: HELLO**



En principio, todos los SNP[16] pueden servir para la ocultación de mensajes cifrados. Sin embargo, las tecnologías actuales de edición genética, como CRISPR/Cas, presentan limitaciones al editar múltiples genomas de manera simultánea. Para ilustrar, para cifrar una secuencia de ADN de 70 nucleótidos, que requeriría 70 SNP, se necesitaría llevar a cabo la edición de 70 nucleótidos. Por lo tanto, se emprendió la búsqueda de puntos críticos de SNP que facilitaran la modificación del genoma. Durante esta indagación, se identificaron regiones que albergaban más de 35 SNP dentro de un espacio de 1 kb. En situaciones donde existían dos puntos de acceso, la inserción de la secuencia cifrada podía lograrse con apenas dos iteraciones de recombinación genética. Se descubrieron cinco puntos de acceso de SNP que poseían al menos 35 SNP en 1 kb, siendo el más extenso con 120 SNP en una región de 1 kb (cromosoma 12, posiciones 88860531–88861530). En términos prácticos, empleando este punto de acceso, la ocultación de la secuencia de ADN cifrada (70 nt) se podría realizar mediante un único experimento de recombinación homóloga.

**Puntos críticos encontrados**

From: DNA steganography: hiding undetectable secret messages within the single nucleotide polymorphisms of a genome and detecting mutation-induced errors

| Chr | Positions | SNPs |
|---|---|---|
| **2** | **88860531–88861530** | **120** |
| **6** | **32663318–32664317** | **45** |
| **14** | **105863177–105864176** | **102** |
| **14** | **105864182–105865181** | **47** |
| **22** | **22880555–22881554** | **47** |

Este plantea interrogantes amplias sobre el exceso de sobreinformación actual. Así mismo el desarrollo de las nuevas formas de ocultamiento, caracterizan la fuerte influencia e inversión que hacen los gobiernos potencias en esteganografía, para establecer formas más avanzadas de ocultación lingüística y criptográfica de los mensajes..

[16] Un polimorfismo puntual, comúnmente conocido como SNP (Single Nucleotide Polymorphism), implica una variación en la secuencia de ADN que afecta a una única base, ya sea adenina (A), timina (T), citosina (C) o guanina (G), dentro de una región del genoma. Aunque se refiere específicamente a cambios en una sola base, la categorización de SNP suele extenderse para incluir variaciones que abarcan unos pocos nucleótidos, así como pequeñas inserciones y deleciones (indels). Es relevante destacar que, para ser oficialmente catalogado como un SNP, se requiere que dicha variación esté presente en al menos el 1% de la población. Si no cumple con este umbral, se clasifica más apropiadamente como una mutación puntual. En algunas instancias, estas variaciones de un solo nucleótido pueden asociarse con el término SNV (Single Nucleotide Variant), que, a diferencia de los SNPs, carece de restricciones en cuanto a su frecuencia en la población.

# 12. CARACTERÍSTICAS LINGÜÍSTICAS

## 12.1 LENGUA INGLESA

12.1.1. **El inglés como fenómeno social:** En el intrincado ámbito de la lingüística, el estudio de la gramática se erige como una exploración multifacética que trasciende los límites del mero conocimiento. Surge no sólo como un depósito de reglas lingüísticas, sino también como un fenómeno social, intrincadamente tejido en el entramado de la interacción humana. Este momento del trabajo se embarca en una odisea académica, adentrándose en las perspectivas matizadas que dan forma a nuestra comprensión de la gramática en la lengua inglesa.

Desde el principio, es imperativo considerar la gramática no solo como un corpus estático de conocimiento, sino como una fuerza dinámica, un fenómeno social que refleja la intrincada danza del lenguaje dentro de las comunidades. Noam Chomsky, una figura destacada en el campo de la lingüística, ha subrayado la creatividad inherente de la gramática en la lengua en su obra revolucionaria, postulando que nuestra competencia lingüística se extiende mucho más allá de las reglas explícitas que aprendemos. Según Chomsky, la gramática es un sistema innato y generativo que nos permite producir una variedad infinita de oraciones novedosas, trascendiendo así los confines de un simple conjunto estático de reglas (Chomsky, 1957). Además, la manifestación social de la gramática se desenvuelve como un aspecto crucial de la indagación lingüística. Dado que el lenguaje sirve inherentemente como un vehículo para la interacción social, la gramática se convierte en un medio a través del cual se modifican las normas y valores sociales. El sociolingüista William Labov, en sus estudios pioneros sobre la estratificación del inglés en entornos urbanos, demuestra cómo la variación lingüística se convierte en un marcador sociolingüístico, reflejando jerarquías e identidades sociales (Labov, 1972). El estudio de la gramática, por lo tanto, se extiende más allá de la sintaxis y la morfología, convirtiéndose en una lente a través de la cual desciframos las intrincadas capas de las estructuras sociales.

El análisis del discurso emerge como una puerta pivotal en la exploración de la gramática, presentando el lenguaje no solo como una estructura abstracta, sino como una interacción dinámica dentro de la comunicación. Erving Goffman, en su trabajo sobre la "posición" de los participantes en la conversación, arroja luz sobre cómo las personas navegan a través del discurso, gestionando activamente sus roles y relaciones dentro del marco lingüístico (Goffman, 1981). El discurso se convierte en una ventana hacia la gramática subyacente, donde la sintaxis y la semántica convergen para construir significado en contexto. La dimensión escrita de la gramática, encapsulada en tratados gramaticales, gramáticas y guías de estilo, ofrece una síntesis integral de las normas lingüísticas. Gramáticos destacados como Strunk y White, en su obra atemporal "The Elements of Style," proporcionan una guía destilada de la gramática inglesa, haciendo hincapié en la claridad y la precisión en la expresión escrita (Strunk & White, 1918). Estas gramáticas escritas no solo sirven como manuales prescriptivos, sino también como artefactos que encapsulan las normas en evolución del lenguaje.En síntesis, el intrincado tapiz de la gramática inglesa se despliega a través de una miríada de perspectivas, desde las ideas generativas de Chomsky hasta las exploraciones sociolingüísticas de Labov y los análisis de discurso de Goffman. Es una entidad dinámica, a la vez un depósito de conocimiento y un reflejo de las dinámicas sociales. Al atravesar las intersecciones de conocimiento, manifestación social, discurso y síntesis escrita, desentrañamos la riqueza incorporada en la gramática del idioma inglés. Este viaje nos impulsa no solo a descifrar las estructuras lingüísticas, sino a desentrañar los hilos intrincados que unen la gramática con la esencia misma de la interacción humana.

## 12.2. LENGUA RUSA

**12.2.1 Características lingüísticas y esteganográficas del Ruso:** La lengua rusa, rica en historia y evolución, ha sido objeto de profundo estudio por lingüistas destacados como Alan Timberlake y Charles Townsend. A lo largo de los siglos, el ruso ha experimentado transformaciones fonológicas, morfológicas y sintácticas que han dado forma a su estructura actual. Alan Timberlake, experto en lingüística eslava, ha contribuido significativamente con su obra "Reference Grammar of Russian", una referencia invaluable que abarca exhaustivamente las reglas gramaticales y las estructuras lingüísticas del ruso contemporáneo. Su trabajo ha sido fundamental para comprender la complejidad y la elegancia de la gramática rusa. Charles Townsend, cuyo libro "Russian W" se suma a la bibliografía de la monografía, puede ofrecer perspectivas adicionales sobre diversos aspectos de la lengua. Su enfoque podría abordar fenómenos lingüísticos específicos, así como el uso práctico del ruso en contextos cotidianos. En conjunto, estos autores y sus obras proporcionan una sólida base para explorar la evolución del ruso, desde sus raíces hasta los desafíos contemporáneos en la descripción y el análisis del idioma. Este curso promete ser una inmersión completa en la riqueza lingüística de la lengua rusa.

El ruso ha experimentado una larga evolución que se remonta al primer milenio de nuestra era. A partir del siglo V, los hablantes eslavos establecieron asentamientos en una vasta área de Europa Central y del Este, desde el Danubio en el sur hasta el Elba en el noroeste. Al este, se desplazaron hacia el norte desde el valle del Dniéper hasta el Golfo de Finlandia y el Alto Volga, desplazando gradualmente o asimilando a los habitantes bálticos y fineses anteriores. El ruso se desarrolló a partir de los dialectos eslavos hablados en el norte de este territorio eslavos del este. En el siglo IX, esta área eslava del este quedó bajo el control de mercaderes-guerreros escandinavos. La cristianización de esta tierra en 988 fue seguida por la sumisión al "yugo mongol" desde el siglo XIII hasta el siglo XV. Como agente favorecido de la Horda de Oro, el pequeño principado de Moscú trajo cada vez más tierras bajo su control. A finales del siglo XV, cuando el yugo mongol fue definitivamente eliminado, Moscú se había convertido en el centro político y eclesiástico de las tierras eslavas del este y el centro del área de la lengua rusa.

El ruso no es solo una lengua hablada, sino también una lengua escrita utilizada para todos los fines culturales. La forma moderna del ruso tomó forma a lo largo del siglo XVIII. La morfología y fonología se basan en el dialecto de Moscú. En su vocabulario, sintaxis y retórica, el ruso, aunque se basa en elementos eslavos nativos, tiene una larga historia de adaptación e internalización de modelos extranjeros, como los bizantinos, franceses y, más recientemente, ingleses. Curiosamente, la palabra moderna "héccrbq" ('ruso') es un adjetivo que deriva del sustantivo "Hécm" ('Rus'). Según una venerable etimología, Hécm era un nombre descriptivo para los escandinavos basado en el étimo germánico 'remar', ya que los escandinavos eran principalmente remeros. En las tierras eslavas del este, Hécm se usaba inicialmente para los señores escandinavos y su principado de Kiev. Con el tiempo, se extendió a todas las tierras eslavas del este. Moscovia se apropió del nombre para su identidad política, cultura y lengua a medida que consolidaba su poder.

El ruso es la primera lengua de aproximadamente 150 millones de personas. Según una estimación de 2002, la Federación Rusa tenía una población total de 145 millones de personas, entre las cuales el 81.5 por ciento, o 118 millones, eran rusos étnicos. A mediados de la década de 1990, había otros 25 millones de rusos en los países recién independizados que surgieron tras la desintegración de la Unión Soviética (Novaia Rossiia 1994). En conjunto, eso haría un total de 143 millones de rusos étnicos. A esa cifra se podría agregar un porcentaje sustancial, aunque indeterminado, de los 27 millones restantes de miembros de otras nacionalidades que residen en la Federación Rusa. Según estadísticas recientes, la tasa de crecimiento poblacional en la Federación Rusa es negativa

(-0.33%), lo que implicaría que el número de hablantes de ruso no aumentará en el futuro previsible.

**Aspectos sociolingüísticos:** La Revolución Rusa de 1917 marcó un punto de inflexión crucial en la historia de Rusia, llevando al derrocamiento del régimen zarista y al ascenso de los bolcheviques liderados por Vladimir Lenin. Esta revolución no solo transformó el orden político y social, sino que también influyó en diversos aspectos culturales, incluido la lengua. Con la creación de la Unión Soviética en 1922, se inició un período de cambios radicales en todos los aspectos de la sociedad, incluido el lenguaje. Las autoridades soviéticas, conscientes del papel fundamental de la lengua en la construcción de una identidad nacional y en la transmisión de la ideología comunista, emprendieron reformas significativas en la lengua rusa.

Uno de los objetivos clave de estas reformas era hacer que el ruso fuera más accesible a las masas, eliminando posibles barreras lingüísticas y facilitando la alfabetización generalizada. En 1918, se implementó una reforma ortográfica que simplificó la escritura, eliminando algunas letras y alterando ciertas reglas gramaticales. El objetivo era simplificar el proceso de aprendizaje y lectura, especialmente para aquellos que no tenían acceso a una educación formal.Además de las reformas ortográficas, se llevaron a cabo esfuerzos para estandarizar la lengua rusa y eliminar influencias extranjeras consideradas burguesas o contrarrevolucionarias. Se abogó por un ruso más "puro" y se eliminaron palabras y expresiones que se percibían como aristocráticas o vinculadas a la vieja élite.

La lengua, bajo el control del Estado, se convirtió en un instrumento poderoso para la propaganda y la difusión de la ideología comunista. Se fomentó la creación de un lenguaje específico para describir los principios del socialismo y para glorificar los logros del régimen. Esta manipulación del lenguaje no solo buscaba transmitir un mensaje político, sino también consolidar la identidad soviética y promover la lealtad al Estado.Sin embargo, estas reformas no estuvieron exentas de críticas. Algunos intelectuales y escritores argumentaron que las medidas para simplificar la lengua comprometían su riqueza y expresividad. La tensión entre la necesidad de un lenguaje accesible para las masas y la preservación de la riqueza cultural y lingüística fue un desafío constante.

**Evolución Ortográfica del Ruso: Desde sus Primeros Cambios Hasta las Reformas del Siglo XX:** El desarrollo de la lengua rusa a lo largo de la historia ha sido moldeado por una serie de cambios ortográficos significativos. Desde sus primeras transformaciones en el antiguo eslavo oriental hasta las reformas del siglo XX, la escritura rusa ha experimentado un fascinante proceso de evolución. La formación del estado ruso centralizado en los siglos XV y XVI llevó a la necesidad de estandarizar el lenguaje utilizado en asuntos administrativos y legales. La introducción del "script civil" por Pedro I en 1708 marcó el inicio de la modernización de la escritura rusa. Aunque esta reforma no fue específicamente ortográfica, la sustitución de Ѧ por Я y la eliminación efectiva de varias letras y diacríticos resultaron en una distinción visual entre la escritura rusa y la eslavonia eclesiástica.

En el siglo XVIII, la ortografía rusa se volvió inconsistente, tanto en la práctica como en la teoría. La influencia de figuras como Mikhail Lomonosov, defensor de la ortografía morfofonémica, y Vasily Trediakovsky, defensor de la ortografía fonémica, reflejó tensiones sobre la mejor manera de abordar la escritura. A lo largo de los siglos XIX y XX, se realizaron ajustes ad hoc a medida que la lengua literaria rusa asumió su forma moderna y altamente estandarizada. Estos incluyeron la introducción de la letra ё y la pérdida gradual de ѵ y ѳ en favor de и, і, ф y т.

**Explicación de los Cambios:**

- **A.** Сѣверо-Американскіе Соединенные Штаты a Северо-Американские Соединённые Штаты:
- **B.** La eliminación de la letra ѣ (yat) y su fusión con la letra e.
  Россія a Россия:
- **C.** La sustitución de la letra i (і) por la letra и.
  Петроградъ a Петроград:
- **D.** La eliminación del "signo duro" ъ al final de la palabra.
  театръ a театр:
- **E.** La sustitución de la letra s (ze) por la letra т.
  дѣти a дети:.

| Ortografía Antigua | Ortografía Moderna | Traducción |
|---|---|---|
| Сѣверо-Американскіе Соединенные Штаты | Северо-Американские Соединённые Штаты | Estados Unidos de América del Norte |
| Россія | Россия | Rusia |
| Петроградъ | Петроград | Petrogrado |
| театръ | театр | teatro |
| дѣти | дети | niños |
| Іисусъ Христосъ | Иисус Христос | Jesucristo |

## 12.3. LENGUA CHINA

El chino, una lengua tonal, presenta una característica distintiva donde el tono de una sílaba tiene el poder de alterar significativamente el significado de una palabra. Por ejemplo, la palabra "ma" 媽 o 妈 puede adquirir significados diversos como "madre", "cáñamo", "caballo" o "regañar" según el tono empleado. Esta complejidad tonal añade una capa adicional de sutileza y expresión al idioma. Además, el chino se clasifica como una lengua aislante, lo que implica que carece de conjugaciones y declinaciones. En lugar de ello, las palabras se combinan para formar frases y oraciones, proporcionando un enfoque único y eficiente en la construcción del significado.

Desde una perspectiva morfológica, el chino exhibe características de una lengua aglutinante. Este rasgo se manifiesta a través de la formación de palabras mediante la adición de sufijos y prefijos a las raíces. Estos elementos afijos no solo modifican el significado de la raíz, sino que también indican la función gramatical de la palabra dentro de la oración. En el ámbito académico de la lingüística china, se destacan varios autores relevantes. Yuen Ren Chao, creador del sistema de romanización Wade-Giles para el chino mandarín, ha dejado una huella significativa en este campo. Asimismo, sociólogos como John DeFrancis y Jerry Norman han contribuido de manera notable al estudio y comprensión profunda de la lengua china.

**Yuen Ren Chao**

***Romanización Wade-Giles:*** Chao desempeñó un papel fundamental en el desarrollo del sistema de romanización Wade-Giles, que facilitó la representación fonética del chino mandarín con caracteres latinos. Este sistema ha sido crucial para los hablantes no nativos al aprender y pronunciar palabras chinas.

***Fonología del Chino Mandarín***: Chao realizó contribuciones significativas al estudio de la fonología del chino mandarín, analizando los tonos y las características distintivas de los sonidos en esta lengua tonal.

**John DeFrancis:**

***Estructura y Gramática***: DeFrancis se destacó en desentrañar la estructura gramatical del chino, ofreciendo análisis detallados sobre la sintaxis y la morfología. Sus obras profundizan en cómo las palabras y frases se combinan para formar significado.

***Historia de la Escritura China***: Exploró la evolución de los caracteres chinos a lo largo de la historia, proporcionando una perspectiva enriquecedora sobre el desarrollo y la complejidad de la escritura china.

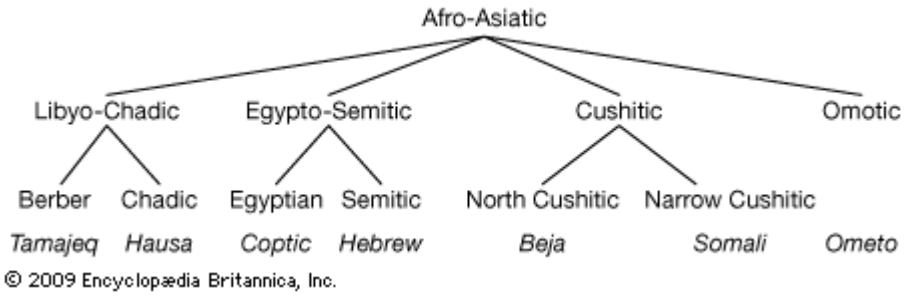
**Jerry Norman:**

***Variación Lingüística:*** Norman ha abordado la variación lingüística dentro del chino, examinando las diferencias entre dialectos y variantes regionales. Su trabajo ha contribuido a la comprensión de la diversidad lingüística en el contexto chino.

***Análisis Sociolingüístico:*** Norman ha explorado las dimensiones sociolingüísticas del chino, investigando cómo factores sociales influyen en el uso del idioma, desde dialectos regionales hasta variaciones en contextos específicos. [17]

## 12.4. LENGUA HEBREA

**S**egún la descripción lingüística propuesta por Britannica[18] el hebreo es una lengua semítica[19] del grupo central septentrional (también llamado noroccidental), está estrechamente relacionado con el fenicio y el moabita, con los cuales a menudo los eruditos lo ubican en un subgrupo cananeo. Hablado en la antigüedad en Palestina, el hebreo fue suplantado por el dialecto occidental del arameo a partir del siglo III a.C.; sin embargo, la lengua continuó utilizándose como lengua litúrgica y literaria. Fue revitalizada como lengua hablada en los siglos XIX y XX y es la lengua oficial de Israel.



Imagen

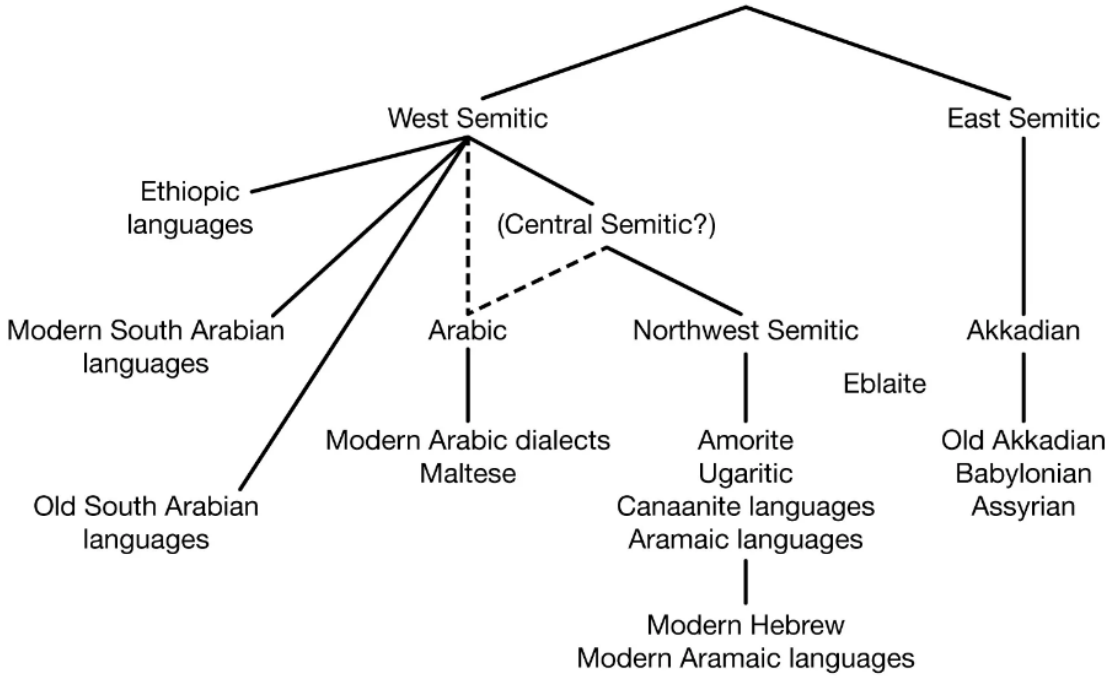
La historia del hebreo se divide generalmente en cuatro períodos principales: el hebreo bíblico o clásico, hasta aproximadamente el siglo III a.C., en el cual se escribió la mayor parte de la Biblia hebrea (Antiguo Testamento); el hebreo mishnaico o rabínico, la lengua de la Mishná (una colección de tradiciones judías), escrito alrededor del año 200 d.C. (este tipo de hebreo nunca se usó entre la gente como lengua hablada); el hebreo medieval, desde aproximadamente el siglo VI hasta el siglo XIII, cuando se tomaron prestadas muchas palabras del griego, español, árabe y otras lenguas; y el hebreo moderno, la lengua de Israel en tiempos modernos. Los eruditos están de acuerdo en que la forma más antigua del hebreo es la de algunos de los poemas en la Biblia, especialmente el "Cántico de Débora" en el capítulo 5 de los Jueces. Las fuentes de palabras prestadas que aparecieron por primera vez durante este período incluyen las otras lenguas cananeas, así como el acadio. El hebreo también contiene un pequeño número de palabras sumerias

[17] Yuen Ren Chao, "Language and Symbolic Systems," 1968. Jerry Norman, "Chinese," 1988. John DeFrancis, "The Chinese Language: Fact and Fantasy," 1984. Yuen Ren Chao, "A Grammar of Spoken Chinese," 1968.

[18] Britannica, T. Editors of Encyclopaedia (2023, October 11). Hebrew language. Encyclopedia Britannica. https://www.britannica.com/topic/Hebrew-language

[19] Que pertenece al grupo de lenguas camitosemíticas que se hablan principalmente en el sudoeste de Asia y en el norte de África.

tomadas de una fuente acadia. El hebreo de la Biblia muestra poca variedad dialectal, pero las huellas literarias dentro del texto indican cierto grado de variación según la geografía. Además de la Biblia, se conserva un pequeño número de inscripciones en hebreo del período bíblico; la más antigua de estas es una breve inscripción en caracteres fenicios que data del siglo IX a.C.

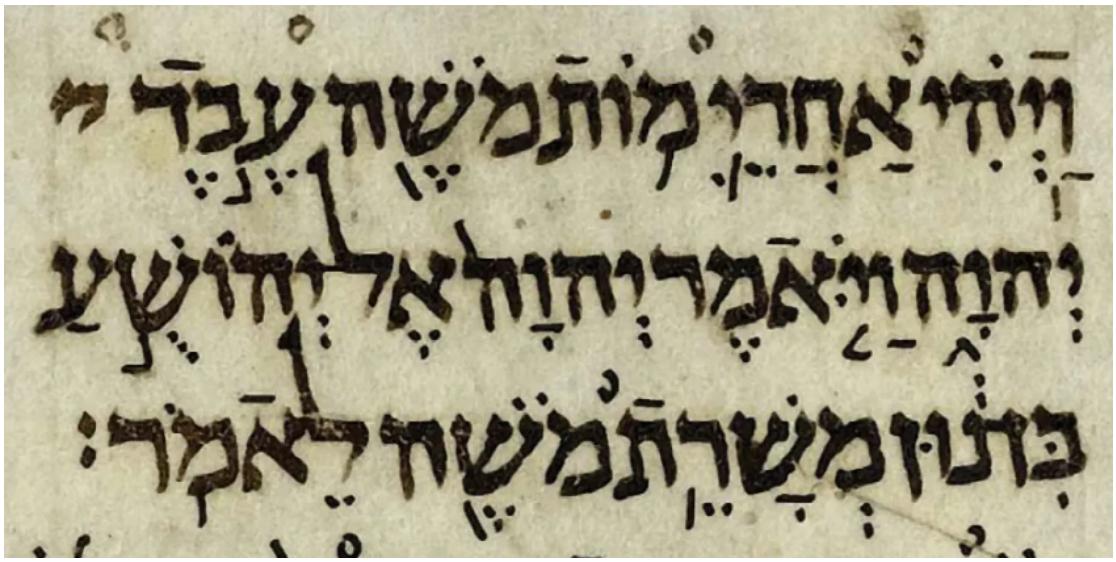



El uso del habla disminuyó desde el siglo IX hasta el siglo XVIII. Sin embargo, el lenguaje medieval experimentó un desarrollo, aunque espasmódico, en diversas direcciones. El culto del poema litúrgico llamado piyyûṭ (una palabra griega en sí misma) entre los siglos VI y IX enriqueció el vocabulario escrito al dar nuevos significados a palabras antiguas e inventar otras nuevas, especialmente en el llamado estilo Kaliriano, y los poetas hispano-hebreos del período 900-1250 siguieron la tendencia. En este período, también se añadieron alrededor de 2,000 o 3,000 términos científicos, filológicos y filosóficos; algunos de estos se formaron haciendo un nuevo uso de raíces antiguas, como en el caso de 'geder' ('cerca'), que también servía para 'definición'. Algunos se basaron en palabras hebreas existentes como 'kammût' ('cantidad'), de 'kammāh' ('¿Cuánto?'); y otros, como 'ʾaqlîm' ('clima') y 'ṭibʿî' ('natural'), se adaptaron de lenguajes extranjeros, principalmente griego y árabe.

El hebreo moderno, basado en el lenguaje bíblico, contiene muchas innovaciones diseñadas para satisfacer las necesidades modernas; es la única lengua coloquial basada en un lenguaje escrito. Las palabras cotidianas se derivan de frases hebreas existentes (por ejemplo, 'šaʿôn' ('reloj') de 'šaʿah' ('hora')) o se toman prestadas de idiomas contemporáneos (por ejemplo, 'mishmesh' ('albaricoque') del árabe 'mishmish'). Un puñado de palabras se reutilizan a partir de expresiones bíblicas (por ejemplo, 'ḥašmal' ('electricidad') de 'ámbar'). La pronunciación es una modificación de la utilizada por los judíos sefardíes (hispano-portugueses) en lugar de la de los judíos ashkenazíes (de Europa del Este). Las antiguas consonantes guturales no se distinguen claramente (excepto por los hablantes palestinos) o se pierden. La sintaxis se basa en la de la Mishná. Característico del

hebreo en todas sus etapas es el uso de raíces de palabras que constan generalmente de tres consonantes, a las que se añaden vocales y otras consonantes para derivar palabras de diferentes partes del discurso y significado.

La lengua se escribe de derecha a izquierda en un guión semítico del norte de 22 letras. Solo se escribían consonantes en el período más antiguo del lenguaje, y algunas de esas consonantes se emplearon más tarde para representar vocales largas. En el siglo VII d.C., los masoretas en Tiberíades introdujeron en el sistema de escritura marcas diacríticas, que representaban vocales cortas y otra información fonológica.



**ASPECTOS SOCIOLINGÜÍSTICOS DEL HEBREO MODERNO**

El análisis exhaustivo realizado por Ricardo Muñoz Solla de la Universidad de Salamanca en 1998 destaca la peculiar situación que experimenta el hebreo moderno como medio de comunicación cotidiana en el Estado de Israel, que ya tiene más de cincuenta años. A pesar de su actual uso, el

hebreo tiene una historia extensa que se remonta a tiempos bíblicos, y los estudios lingüísticos han buscado caracterizar la diferencia entre el hebreo que comenzó a utilizarse a finales del siglo XIX y el hebreo de épocas anteriores, que mantenía viva únicamente la tradición escrita y servía como vínculo para los judíos que vivían en la diáspora. Muñoz Solla aborda tres aspectos cruciales para comprender la evolución de la lengua hebrea desde el siglo XIX hasta su consolidación como idioma nacional de Israel. Primero, examina el resurgimiento del hebreo en tres etapas: el surgimiento como lengua oral entre 1890 y 1916, la estandarización entre 1916 y 1948, y la modernización léxica desde 1948 hasta la actualidad. Destaca el papel crucial de Eliezer Ben-Yehudah a partir de 1880 en la generalización del uso del hebreo como lengua franca en los territorios palestinos.

Luego, el autor aborda el normativismo e innovación lingüística en el desarrollo del hebreo moderno. Señala que el resurgimiento nacionalista llevó a la uniformidad de diversos registros bajo una koine, representando la forma estandarizada de la nueva lengua. Destaca la influencia directa del hebreo bíblico, postbíblico, misnaico y medieval, que creó una estrecha vinculación con el hebreo moderno, aunque este último desarrolló estructuras y vocabulario particulares, a menudo alejados de las normas gramaticales preexistentes. Por último, Muñoz Solla examina la influencia de la inmigración y el aprendizaje del hebreo en la constitución de la sociedad israelí. Describe cómo las sucesivas oleadas migratorias han afectado directamente diversos aspectos del hebreo, desde la homogeneización en el hebreo estándar hasta la selección lingüística realizada por diferentes generaciones de hablantes según su procedencia. Además, destaca el papel de la política lingüística del Ministerio de Educación israelí en la aculturación de los inmigrantes y la estandarización del hebreo moderno, aunque señala los desafíos actuales relacionados con la integración de grupos de inmigrantes rusos reticentes al aprendizaje del hebreo.

## NORMATIVISMO E INNOVACIÓN LINGÜÍSTICA

La tradición purista en la concepción de la nueva lengua condicionó el desarrollo del hebreo en su esfuerzo por definir los criterios que prevalecerían en el uso diario de la lengua, mientras que los usos coloquiales fueron los generadores de las principales innovaciones lingüísticas. La normalización fonética concilió las dos tradiciones más importantes de pronunciación, la askenazí y la sefardí, favoreciendo el predominio de esta última. Los rasgos más característicos de cada una fueron eliminados, dando lugar a un sistema fonético de compromiso entre ambas. De esta manera, el hebreo moderno desarrolló dos variantes de pronunciación. La más generalizada, fruto de esta reestructuración de la "israelí general", se enfrenta a la "israelí oriental", utilizada mayoritariamente por las comunidades sefardíes. En el estilo formal, se adhiere a la modalidad estándar israelí. La expansión de la modalidad israelí se apoya fuertemente en el papel decisivo de los medios de comunicación y en la valoración positiva otorgada por los hablantes al dominio de esta pronunciación, que goza de mayor prestigio social.

Dentro del campo léxico, las innovaciones son tan numerosas que resulta difícil ofrecer un panorama sencillo de sistematización. La necesidad de designar nuevos conceptos, inexistentes en las tradiciones anteriores, ha llevado a una constante incorporación de palabras de otras lenguas. Aunque en el ámbito científico el inglés sirve como base para la creación de nuevos términos, en el ámbito coloquial la situación es más variada, considerando el origen de los hablantes, que conservan como señal de identidad cultural muchas palabras en sus usos expresivos y familiares. El influjo de las lenguas occidentales es tan intenso que ha llevado a una occidentalización significativa de las estructuras del hebreo, otorgándole una fuerte personalidad europea al hebreo

estándar. Entre la actual tendencia hacia una nueva "re-hebraización" de los usos y la aceptación del hebreo como lengua occidental, el mayor desafío actual, que refleja hasta qué punto persiste la lucha entre el normativismo y la innovación, sigue siendo la influencia de sustratos europeos. Si en un primer momento fue el eslavo y el germánico, el hebreo moderno se enfrenta ahora al difícil objetivo de equilibrar las presiones del inglés en una sociedad en gran parte americanizada, y la presencia de numerosos inmigrantes que son reticentes a perder sus formas lingüísticas maternas. Mientras tanto, el yidis o el judeoespañol van disminuyendo en número de hablantes y se han convertido en lenguas de identificación comunitaria (el yidis para los judíos ultraortodoxos y el judeoespañol para los descendientes sefardíes).

**BEN-YEHUDA אליעזר בן-יהודה**

Eliezer Ben-Yehuda nació en Luzhki cerca de Vilna, Lituania, en 1858, y su nombre era Eliezer Yitzhak Perelman. El trabajo de Damian Alejandro Dzienciarsky en 2020 resalta el trabajo en torno a la conmemoración de los 130 años desde la creación del Comité de la lengua hebrea. Yehuda fue sin duda el precursor de la Academia de la lengua hebrea. A menudo, se atribuyen a Eliezer Ben-Yehuda títulos como el resucitador del hebreo y el padre de la lengua sagrada en relación con este evento. Sin embargo, se destaca que el resurgimiento del hebreo no fue obra exclusiva de un individuo, sino de un grupo de intelectuales y estudiosos. La conmemoración oficial, en 1989, de los cien años del comité como "los cien años de la lengua hebrea" subraya la influencia del comité en la posterior creación de la Academia.

El Vaad halashón, liderado por Ben-Yehuda y David Yellin, junto con figuras como el Rabino Chaim Hirschensohn y Avraham Moshe Luntz, desempeñó un papel crucial en la contribución al hebreo. Sin embargo, entre el cierre del comité y la apertura de la academia, pasó un período de inactividad considerable debido a diferencias de ideas y falta de coherencia entre sus miembros.Al analizar la vida de Ben-Yehuda, se revela su ferviente deseo de emigrar a la Tierra de Israel, motivado por la emancipación de su pueblo. Ben-Yehuda, un miembro activo del movimiento "Chovevei Tzion", llegó a Palestina en 1881 como parte de un grupo comprometido con el renacimiento nacional judío en la tierra de los patriarcas. Se destaca su contribución al nuevo yshuv, una forma incipiente de organización que allanó el camino para la creación del Estado de Israel en 1948, marcando logros significativos en la hebraización de la educación y la cultura.

En este contexto, Ben-Yehuda, tras llegar a Jerusalén, trabajó en el periódico Havazelet y se desempeñó como maestro de hebreo en la escuela Alliace, utilizando un método innovador de enseñar el idioma solo en hebreo. Aunque su salud se debilitó y dejó la enseñanza, se dedicó a escribir y editar varios periódicos. La figura de Ben-Yehuda, por tanto, se presenta como parte integral de un esfuerzo colectivo para el resurgimiento del hebreo en Palestina. El hebreo carecía de muchas palabras para referenciar las nuevas creaciones,  situaciones, adjetivos y verbos de finales del siglo XIX. Podríamos citar, por ejemplo, dos listas de palabras renovadas por Ben Yehuda (Sivan, 1961, p. 67). La primera es una lista de palabras que el público hablante aceptó e incorporó al hebreo. Algunos ejemplos:

אומנות (omanut) - arte
אפנה (ofná) - moda
אופניים (ofanaim) - bicicleta
אקדח (ekdach) - revolver
בובה (bubá) - muñeca
בורג (boreg) - tornillo
גלידה (glidá) - un helado

הפתעה (aftaá) - sorpresa
חביתה (chavitá) - omelette
חיידק (chaidak) - virus
חייל (chaial) - soldado
מברק (mivrak) - telegrama
מגבת (maguevet) - toalla
מגהץ (maguetz) - plancha de ropa
מטרייה (mitriá) - paraguas
מילון (milón) - diccionario
מעטפה (maatafá) - sobre de carta
מסעדה (misadá) - restaurante
מספרה (mispará) - peluquería
מקהלה (makhelá) - coro
מקלחת (miklachat) - bañadera
משטרה (mishtará) - policía
משרד (misrad) - oficina
עיתון (itón) - diario, periódico
פצצה (ptzatzá) - bomba
רכבת (rakevet) - tren
שעון (shaón) - reloj

**Palabras no aceptadas por los hablantes:**
אבכמץ (avchemetz) - oxígeno
כוחית (kochit) - dinamita
מקושית (makoshit) - piano

En este artículo, se destaca la monumental obra de Eliezer Ben-Yehuda, el Gran Diccionario del Antiguo y Nuevo Hebreo, que consta de 16 volúmenes más una introducción, abarcando desde 1908 hasta 1959. Este diccionario, apreciado por lingüistas y lectores, refleja la innovación lingüística y la creación de nuevas palabras durante la actividad de Ben-Yehuda. El enfoque clave aquí es cómo Ben-Yehuda logró difundir sus creaciones en una época sin medios de comunicación modernos. Se revela que la difusión y aceptación de sus propuestas, incluidas las nuevas palabras, se debieron en gran medida a la publicidad en los periódicos de la época. Ben-Yehuda aprovechó la morfología fija del hebreo para crear neologismos, y su habilidad para publicitar estas innovaciones en los periódicos fue crucial para su reconocimiento y aceptación por el público. Se destacan ejemplos específicos, como la creación de la palabra "kayas" para un ladrón de poca monta, nacida por necesidad de espacio en un titular. La influencia de Ben-Yehuda en la difusión de nuevas palabras mediante la publicidad en los periódicos se presenta como un factor decisivo en el resurgimiento del hebreo y su impacto duradero en el lenguaje.

## 13 . COMPARACIÓN SOCIOPOLÍTICA

Colombia ostenta el título del conflicto sin resolver más antiguo del mundo. La persistencia y los cambios en este conflicto están vinculados a una extensa lista de factores estructurales, entre los cuales se incluyen el persistente problema agrario, el acceso limitado a la participación política, el auge del narcotráfico, las presiones del contexto internacional y la fragmentación institucional y territorial del Estado (Grupo de Memoria Histórica, 2016, p. 111). En el informe "Basta Ya" del Centro Nacional de Memoria Histórica se identifican cuatro periodos clave: (a) la transformación de la violencia bipartidista en violencia subversiva (1958-1982); (b) la expansión guerrillera, las políticas de paz y el surgimiento paramilitar (1982-1996); (c) el recrudecimiento del conflicto armado (1996-2005); y (d) la desmovilización de las Autodefensas Unidas de Colombia (AUC) y la ofensiva contrainsurgente (2005-2012).

El primer periodo (1958-1982) se caracterizó por el surgimiento de guerrillas, en contraste con el auge de la movilización social y la marginalidad del conflicto armado. El segundo periodo (1982-1996) abarcó la crisis y el colapso parcial del Estado, la irrupción del narcotráfico a nivel global, la promulgación de la Constitución Política de 1991 y la expansión territorial y militar de las guerrillas. El tercer periodo (1982-1996) se destacó por las expansiones simultáneas de guerrillas y grupos paramilitares, la crisis y recomposición del Estado en medio del conflicto armado, y la radicalización política de la opinión pública hacia una solución militar del conflicto (GMH, 2016, p. 112). También fueron centrales los cambios organizacionales en el narcotráfico y su tratamiento internacional a través de la lucha contra el terrorismo. La cuarta categoría (2005-2012) abarcó la ofensiva militar del Estado, que llevó al reacomodo militar de las guerrillas. Además, las negociaciones con los paramilitares derivaron en un rearme y asentamiento violento de estos grupos en el territorio colombiano (2005-2012).

Por otro lado, abordar el conflicto colombiano implica remontarse a la fundación de la república. Sin embargo, para los fines de este trabajo y para establecer un contexto que permita comprender ambos fenómenos, se eligió como punto de partida la década de los 90. Previo a la elección de César Gaviria como presidente de Colombia, tres candidatos presidenciales fueron asesinados, marcando el camino hacia una Asamblea Nacional Constituyente y la decisión del Estado y la sociedad de enfrentar al narcotráfico y las guerrillas.
Gaviria extendió la mano a las guerrillas para que participaran en la Asamblea Constituyente, siempre y cuando abandonaran las armas. Prometió una lucha implacable contra los paramilitares, calificándolos como la peor amenaza para Colombia (Comisión para el Esclarecimiento de la Verdad, la Convivencia y la No Repetición, 2022, p. 241). La nueva Constitución de Colombia logró el consenso de diversos sectores políticos, sociales y étnicos, proporcionando un instrumento que garantizaba el equilibrio de poderes y el disfrute de los derechos de los colombianos. A pesar de esto, "la constitución no realizó reformas en el sector de seguridad, a pesar de las graves masacres y asesinatos ocurridos meses antes en Urabá y Magdalena Medio" (p. 251), lo que desencadenó procesos posteriores que legitimaron la unión de las fuerzas legítimas del Estado con grupos de narcotráfico y paramilitares. En este periodo, también se formalizaron las redes de inteligencia militar en cada fuerza y se asimilaron las redes de informantes (CEV, 2022, p. 271), elementos cruciales para comprender el auge del paramilitarismo.

Durante el gobierno de César Gaviria, se realizaron intentos fallidos de paz con la Coordinadora Guerrillera Simón Bolívar (SGSB). Además, se implementó la Estrategia Nacional contra la Violencia, que se centró en tres objetivos: garantizar el monopolio del uso de la fuerza por parte del Estado, recuperar la capacidad de la justicia y ampliar el alcance institucional del Estado (p. 268). A pesar de estos esfuerzos, a mediados de los años 90, las FARC se fortalecieron, estableciendo que cada bloque debía ser autónomo, lo que llevó al aumento de secuestros y extorsiones en las regiones. "Convirtieron a las clases medias y sectores más pobres en víctimas recurrentes de las guerrillas" (CEV, 2022, p. 278).

En 1994, tras la muerte de Fidel Castaño, sus hermanos crearon las Autodefensas Campesinas de Córdoba y Urabá (ACCU), precursoras del fenómeno paramilitar. "No eran simplemente un ejército privado de narcotraficantes, sino que tenían una vocación territorial" (CEV, 2022, p. 299) y lograron aliados en la política, las fuerzas militares y el narcotráfico. Aunque las autodefensas habían formado parte de proyectos en gobiernos anteriores a la década de los 90, fueron prohibidas por Virgilio Barco. Posteriormente, en 1995, Ernesto Samper creó los Servicios Comunitarios de Vigilancia y Seguridad Privada (Convivir). Según el informe final de la Comisión para el Esclarecimiento de la Verdad, muchos grupos armados se registraron bajo la figura de las Convivir, y en 1996, se crearon más de 60 de estas.

**Narcotráfico y Conflictos Armados: Colombia y Afganistán como Estudios de Caso:** La relación entre el narcotráfico y los conflictos armados ha sido una preocupación global, y dos países, Colombia y Afganistán, han emergido como casos de estudio particulares debido a su papel destacado en la producción de cocaína y heroína, respectivamente. Este ensayo explorará la conexión entre la financiación de la guerra y las potencias de siembra de drogas en estos dos países, utilizando datos de la ONU para comprender la complejidad de esta interrelación.
Colombia ha sido históricamente reconocida como el mayor cultivador de cocaína a nivel mundial. La producción y distribución de esta droga han alimentado los conflictos armados internos, proporcionando financiamiento a grupos rebeldes y paramilitares. Según datos de la ONU, la relación entre el narcotráfico y el conflicto en Colombia es innegable. Los cárteles de drogas y las guerrillas han encontrado en el cultivo de coca una fuente sostenible de ingresos para financiar sus actividades. La década de 1990 fue testigo de la brutalidad de los cárteles y grupos rebeldes que luchaban por el control de las rutas de drogas. Los esfuerzos gubernamentales y la asistencia internacional han intentado abordar este problema, pero la conexión persistente entre el narcotráfico y la violencia armada plantea desafíos multifacéticos.

En Afganistán, el escenario es diferente pero igualmente complejo. La nación se ha convertido en el principal productor mundial de heroína, un título que se ha mantenido a lo largo de los años. La producción de opio, materia prima para la heroína, ha crecido en medio de la inestabilidad política y los conflictos armados. Según datos de la ONU, la conexión entre el narcotráfico y la guerra en Afganistán es intrínseca. Las plantaciones de adormidera, aprovechadas por los talibanes y otros actores armados, han proporcionado una fuente crucial de financiamiento para los grupos insurgentes. La lucha contra el narcotráfico se entrelaza con los esfuerzos por estabilizar el país, y los programas de erradicación se enfrentan a desafíos significativos en un entorno donde la producción de opio a menudo se percibe como la única opción económica viable.

La intersección entre el narcotráfico y los conflictos armados en Colombia y Afganistán plantea cuestiones críticas sobre las estrategias globales de lucha contra las drogas y los esfuerzos para alcanzar la paz. La dependencia económica de estas actividades ilícitas crea un ciclo de violencia difícil de romper. La experiencia colombiana destaca la necesidad de enfoques integrales que aborden tanto la oferta como la demanda de drogas, combinando estrategias de desarrollo económico con medidas de seguridad. En Afganistán, la complejidad política y social exige una comprensión profunda de las dinámicas locales para abordar eficazmente la producción de opio. En ambos casos, la comunidad internacional desempeña un papel vital en el apoyo a soluciones sostenibles. La lucha contra el narcotráfico debe ir de la mano con iniciativas de construcción de paz y desarrollo, reconociendo la conexión intrínseca entre la economía del narcotráfico y los conflictos armados.

**Desafíos Emergentes y Esteganografía Criminal:** la conexión entre el narcotráfico y la financiación de la guerra ha llevado a la evolución de métodos cada vez más sofisticados para el intercambio clandestino de información. En un mundo digital, la esteganografía, la técnica de ocultar datos dentro de otros datos, se ha vuelto una herramienta clave para aquellos involucrados en actividades ilícitas.

La esteganografía criminal, en este contexto, agrega una capa adicional de complejidad a la lucha contra el narcotráfico y los conflictos armados. Los actores criminales utilizan técnicas avanzadas para ocultar mensajes, transacciones y coordinación, desafiando los esfuerzos de las fuerzas del orden y agencias de inteligencia. Datos de la ONU pueden proporcionar una visión parcial de estos fenómenos, pero la clandestinidad inherente a la esteganografía criminal hace que la detección y prevención sean extremadamente desafiantes.

La evolución de la esteganografía criminal plantea preguntas cruciales sobre la adaptabilidad de las estrategias de seguridad y la necesidad de capacidades avanzadas en el ámbito digital. La comunidad internacional enfrenta el reto de mantenerse al día con las innovaciones tecnológicas utilizadas por aquellos que buscan eludir la ley. La cooperación global y el intercambio de información se vuelven esenciales para abordar esta nueva dimensión de la delincuencia organizada.

**LA DIVISA NARANJA:** El inicio de las comunicaciones militares en Colombia se remonta a 1944, cuando mediante Decreto No 2787 de noviembre 16, se reconoce como un Servicio Técnico de apoyo para el combate. La importancia de esta especialidad, evidenciada durante los años siguientes, motivó, 52 años después, su promoción a la categoría de arma.Soportados en tecnología de punta, los integrantes de la Divisa Naranja tienen la delicada responsabilidad de garantizar el contacto permanente y seguro entre las diferentes unidades de la Fuerza.

**Evolución y Consolidación de las Comunicaciones Militares en el Ejército Nacional:** Las comunicaciones militares han desempeñado un papel crucial a lo largo de los conflictos bélicos, no solo como componente esencial del mando, sino también como garantes del alcance y la coordinación de operaciones militares a lo largo de la historia. En el contexto colombiano, este elemento vital tiene una trayectoria que se remonta a 1932, durante el conflicto contra el Perú, cuando el Ministerio de Guerra estableció una Sección de Comunicaciones para enlazar puestos fronterizos y mantener la conexión de las tropas con el país. La década de 1940 marcó la consolidación de la especialidad, con la formación de hombres especializados en comunicaciones provenientes de la Compañía de Comunicaciones del Socorro. En 1944, la creación de la Escuela de Capacitación Técnica en Comunicaciones impulsó aún más el desarrollo de esta área como especialidad técnica, fortaleciendo la formación de oficiales y suboficiales del Ejército.

La experiencia internacional, especialmente la participación del Batallón Colombia en la Guerra de Corea, contribuyó al proceso de consolidación de las comunicaciones militares. En 1956, fueron oficialmente reconocidas como arma y servicio técnico del Ejército Nacional. La década del setenta presenció los primeros pasos hacia la implementación de la Red Integrada de Comunicaciones (RIC), un proyecto crucial para la integración y eficiencia de las comunicaciones en las Fuerzas Militares.Sin embargo, en 1997, ante el recrudecimiento de la violencia en Colombia, las comunicaciones militares adquirieron un nuevo estatus mediante la Ley 492, siendo reconocidas como especialidad orgánica del Ejército Nacional, contribuyendo así al desarrollo del conflicto armado en el país. A medida que avanzó el nuevo milenio, se intensificaron los esfuerzos para modernizar las comunicaciones militares. La creación del Comando Conjunto Cibernético en 2012 y el Batallón de Comunicaciones N° 3 en 2015 reflejan la adaptación a las nuevas amenazas y la importancia de la ciberdefensa.

En la actualidad, las comunicaciones militares se estructuran en especialidades como Comando, Control y Comunicaciones, Computación, Ciberdefensa, Guerra Electrónica en Defensa y Protección, y Logística de Comunicaciones. Estas especialidades, delineadas por la Escuela de Comunicaciones, reconocida como institución de educación superior, reflejan la evolución continua y la proyección hacia el futuro de las comunicaciones militares en el Ejército Nacional de Colombia.

## 14. LITERATURA EN LENGUA ESPAÑOLA

Varios estudios propuestos por la Universidad Complutense de Madrid exploran la aplicación de la esteganografía lingüística en lengua española, basada en modelos N-Gram y la Ley de Zipf. Este análisis, realizado por Alfonso Muñoz Muñoz e Irina Argüelles Álvarez de la Universidad Politécnica de Madrid, se enfoca en la generación automática de textos en español que oculten información específica. El artículo plantea hipótesis y lleva a cabo experimentos mediante la implementación de un algoritmo. Los resultados de las pruebas indican la viabilidad de utilizar modelos N-Gram y características derivadas de la Ley de Zipf para generar estegotextos con una calidad lingüística tal que un lector humano podría no diferenciarlo de otro texto auténtico. Estos estegotextos permitirían la ocultación de al menos 0,5 bits por palabra generada (Muñoz y Argüelles, Universidad Politécnica de Madrid).El interés en la última década en esta ciencia se debe a que puede dar solución a dos problemas comunes: la privacidad y el anonimato y el marcado digital de textos. Para ello se utilizan dos grandes familias de algoritmos clasificados en técnicas de modificación de textos existentes y en técnicas de generación automática de estegotextos.

**Técnicas en adverbios:** Las técnicas más tradicionales de ocultación de información consisten en modificar elementos de un texto existente para esconder la información. En la última década, se han propuesto diversas técnicas de este tipo (Bergmair, 2007), abordando problemas específicos para diversas lenguas como inglés, español, japonés, chino, árabe, ruso, entre otras. Aunque estas técnicas pueden utilizarse para la protección y anonimato de comunicaciones, la comunidad científica muestra mayor interés en su aplicación para el marcado digital de textos, conocido como Natural Language Watermarking (NLW). Entre las variantes de estas técnicas se encuentran algoritmos basados en modificaciones léxicas (Chapman y Davida, 1997), modificaciones sintáctico-semánticas (Muñoz, Argüelles y Carracedo, 2009; Muñoz y Argüelles, 2012), ocultación basada en la traducción entre lenguas (Grothoff et al., 2005; Meng et al., 2010), ocultación basada en errores ortográficos y tipográficos (Topkara, Topkara y Atalla, 2007), y ocultación basada en la estructura y formato de un texto (Bergmair, 2007).

El problema de mantener en secreto o destruir el texto original o portador, en el cual se realizarán las modificaciones para crear el esteganotexto resultante con la información enmascarada, es crucial en estas técnicas basadas en la modificación de textos existentes. La comparación entre el texto original y el esteganotexto facilitaría la tarea de detección para un estego-analista. Por estos motivos, se destaca la importancia del método alternativo basado en la generación automática de esteganotextos. Este enfoque no solo permite mejorar el esteganotexto generado considerando diversas consideraciones lingüísticas y estadísticas, sino que también complica la tarea del esteganoanalista al generar un esteganotexto único para cada comunicación enmascarada. Sin embargo, la generación automática de esteganotextos presenta enormes desafíos en términos de semántica y coherencia global, especialmente para textos de longitud media y temática variada. A pesar de la viabilidad en la generación de textos con validez léxica y sintáctica, la solución para la semántica y coherencia global aún no está clara.

Considerando las ventajas y limitaciones de ambos métodos, este trabajo se centra en la posibilidad de generar estegotextos automáticamente en lengua española utilizando modelos N-Gram[20]. Además, se introduce un concepto innovador de edición manual de estegotextos a posteriori, facilitando la producción de estegotextos que se asemejen a aquellos escritos por humanos.El trabajo de Arguelles propone que Los algoritmos de generación automática de textos deben considerar la calidad léxica, sintáctica y semántica, así como la cohesión y la coherencia del esteganotexto resultante. Para aproximarse a este problema, desde finales del siglo XX se han propuesto dos grandes líneas de generación, que se pueden aplicar conjuntamente: unas basadas en imitación gramatical y otras basadas en imitación estadística de un texto "típico" en una lengua concreta.

a) Modificaciones Léxicas: Estos procedimientos consisten en la ocultación de información mediante la sustitución/modificación de palabras. El método más utilizado es la sustitución basada en el uso de sinónimos. Desde que esta idea fuera trabajada por Chapman y Davida en 1997 es considerada como una excelente opción y estudiada en diversas lenguas (Bergmair 2007). El mayor problema con esta técnica es que, o no existen, o son muy pocos los sinónimos puros en una lengua, es decir, dos palabras que signifiquen exactamente lo mismo en cualquier contexto. Por este motivo, conseguir herramientas prácticas con estos principios, ya sea para ocultación de información en general o para el marcado digital de textos, requiere de sofisticados mecanismos para determinar cuál es la ambigüedad de una palabra en un contexto determinado y para saber si puede ser reemplazada o no por otra palabra. Para ello, se requieren procedimientos WSD (Word Sense Disambiguation) y estudios estadísticos que indiquen cuáles son los más aconsejados de entre los sinónimos disponibles para una palabra.

b) Modificaciones Sintácticas-Semánticas: Los algoritmos de ocultación más robustos basados en esteganografía lingüística deberían ser capaces de aplicar modificaciones 33 sintáctico-semánticas a un texto para ocultar información sin perder la coherencia y la semántica del texto. Esta investigación está completamente abierta en su aplicación a diferentes lenguas. Algunos recursos sintácticos documentados en inglés, chino, coreano, turco, ruso, persa, etc., para ocultar información son (Bergmair 2007): el cambio de voz activa/pasiva, el movimiento de los adverbios dentro de la oración, el cambio de orden de los términos unidos por conjunciones (por ejemplo, listo y guapo o guapo y listo), etc. Por otro lado, se sigue trabajando en mecanismos que aprovechándose de descripciones semánticas (ontologías aplicadas a la esteganografía) faciliten la ocultación de información considerando la semántica y la coherencia de un texto. Un ejemplo sencillo consiste en la inserción de sintagmas o términos semánticamente "vacíos", es decir, que no afectan al contexto sintáctico-semántico, por ejemplo, en inglés, la inserción de adverbios de evaluación delante de una frase como: Basically, it seems that, etc.

c) Modificaciones basadas en el ruido de las traducciones de un texto entre diferentes lenguas: La idea de estos procedimientos consiste en ocultar información basándose en la posibilidad de traducir una oración de una lengua concreta por varias oraciones "equivalentes" en la lengua de destino, entre las cuales se puede elegir estableciendo un sistema binario de ocultación de información (Bergmair 2007).

d) Ocultación basada en errores tipográficos y ortográficos: Abreviaturas y símbolos de puntuación. Estos mecanismos pueden parecer triviales desde un punto de vista lingüístico, sin

[20] Un n-grama es un conjunto de n elementos consecutivos en un documento de texto, que puede incluir palabras, números, símbolos y puntuación.

embargo, pueden presentar utilidad esteganográfica si los textos creados con estas modificaciones se insertan en "canales" donde este tipo de errores sean frecuentes. Por ejemplo, incluir textos esteganográficos basados en faltas ortográficas en foros en Internet donde los textos escritos presentan muchos errores de este tipo. De la misma forma, se han documentado diferentes procedimientos para la ocultación de información utilizando abreviaturas de palabras, por ejemplo, de manera ingeniosa en mensajes sms/mms (ShiraliShahreza 2007). En esta línea, los símbolos de puntuación (punto, coma, punto y coma, etc.), más exactamente su colocación o no en zonas de un texto, también pueden ser utilizados para establecer sistemas binarios de ocultación. La creación de reglas generales para aplicar esta idea a los textos de una lengua no es nada sencilla. Por ejemplo, en turco puede usarse una coma después de un sujeto si ésta está relativamente distante del verbo. La vaguedad de esta norma hace que su automatización sea difícil (Bergmair 2007).

e) Ocultación basada en formato: Este es el mecanismo más tradicional para ocultar información y se basa en el formato-estructura de un texto. Los recursos más utilizados son: el uso de caracteres invisibles, separación entre líneas o palabras (por ejemplo, uso de espacios de tamaño variable entre palabras) y codificación de información basada en sucesivos cambios del formato del texto (estilo de fuente, color, tamaño de letra, subrayado, negrita, cursiva, mayúsculas, etc.). Estos mecanismos (Bergmair 2007) pueden favorecer la ocultación de una cantidad razonable de información pero no están exentos de problemas.

A menudo, simples ataques activos que anulan el formato, eliminan la información oculta. De los procedimientos comentados, las modificaciones basadas en sustituciones léxicas o modificaciones sintáctico-semánticas presentan, a priori, mayor interés por su mayor robustez y porque la información modificada (el texto) no dependería necesariamente del canal donde se emite, así, de esta forma, el esteganotexto resultante podría ser impreso en papel, escaneado, intercambiado oralmente, etc. y esto no afectaría a la seguridad del sistema planteado. Por desgracia, existen pocos estudios públicos que analicen el potencial de esta tecnología en lengua española. Por eso, en lo que sigue, centramos el interés en el potencial de las modificaciones sintácticas en lengua española con utilidad esteganográfica y en el análisis de su productividad desde el punto de vista esteganográfico. Entendemos que una estructura lingüística o alteración es productiva esteganográficamente cuando facilita la posibilidad de ocultar la información necesaria para el objetivo perseguido. Por ejemplo, si se desea autocontener una información a modo de firma digital de unos 40 bits en artículos periodísticos (las firmas digitales tradicionales son de un mínimo de 128 bits), artículos en torno a 500 palabras, estas condiciones podrían limitar el uso de ciertas estructuras lingüísticas que no aparecen muy a menudo en textos breves. 2. Modificaciones sintácticas con fines esteganográficos en español. Antecedentes No se conocen estudios públicos sobre la utilidad de las modificaciones sintácticas en frases en lengua española con utilidad en esteganografía lingüística y en marcado digital de textos (natural language watermarking). En teoría, la manipulación sintáctica de una frase con fines esteganográficos se basa en el hecho de que las frases son combinaciones de sintaxis y semántica, y de que la semántica de una frase podría ser expresada por más de una estructura sintáctica. Es importante recordar que el fin es poder alterar una frase sintácticamente sin que la semántica de la misma o la coherencia global del texto se vea afectada. Si existe más de una posibilidad de expresar "lo mismo" puede elegirse entre las opciones disponibles y la decisión por una u otra opción es lo que permitirá ocultar información.

Con estos principios como marco se inició en (Muñoz et al. 2010) un primer acercamiento al análisis de la conveniencia o no de utilizar transformaciones ya usadas con éxito en otras lenguas e indagar sobre otras nuevas. En este marco, se publicó en (Muñoz et al. 2010) un análisis del potencial de algunas modificaciones sintácticas en lengua española, entendiendo por modificación sintáctica aquélla que no consiste exclusivamente en modificar, añadir o suprimir algún término de una frase, cambios que, desde nuestra perspectiva, suponen modificaciones léxicas a diferencia de cómo los entienden otros autores (Murphy y Vogel 2007). En este estudio se profundizó en una transformación sintáctica basada en el cambio activa/pasiva y se destacó el potencial de dos líneas de transformación cuyos resultados parciales necesitan un análisis más exhaustivo (movimiento de adjetivos y movimiento de adverbios). Es en esta investigación en curso y en algunos de sus resultados en los que se centra este trabajo que profundiza en la posibilidad de usar modificaciones basadas en la reordenación de complementos del verbo con utilidad esteganográfica. A continuación se describen brevemente los resultados obtenidos en el trabajo previo (Muñoz et al. 2010) y se enlaza con los nuevos resultados.

**Transformación sintáctica basada en el cambio activa-pasiva:** En lengua inglesa se ha documentado en los últimos años la posibilidad de utilizar la transformación de una frase de activa a voz pasiva y viceversa con utilidad esteganográfica. Por ejemplo, la oración Peter builds a house/A house is built by Peter. La ocultación de información basada en esta transformación consiste en asignar un bit 0 o 1 a cada frase en voz activa y el bit contrario 1 o 0 a cada frase en voz pasiva. La información que se quiera ocultar se convertirá a binario y cada frase seleccionada del texto tapadera se modificará, si es necesario, para que refleje el código binario que se desea ocultar. En (Muñoz et al. 2010) utilizando el corpus LEXESP se cuantificó la utilidad de esta transformación en lengua española. Como indica Lázaro Carreter (1980), el uso de la pasiva en español no es común y éste es el mayor problema en su aplicación a esteganografía.

***Ejemplificación en el español colombiano***

Para hacer este piloto vamos a hacer “esteganografía” ocultando las coordenadas de una “supuesta” operación militar en la región del Cauca. El nombre clave de nuestra operación será llamada “ Rigatoni”. La orden será dada por un comando NN por medio de una cuenta de Twitter específica que por seguridad solo será dada al miembro a cargo de la comunicación dentro de la operación. Todo el mensaje se ocultará en la receta de un rigatoni italiano.

El tweet en x tendrá la siguiente información en cuestión:

**Ingredientes:**

500 g de rigatoni
Agua (para cocinar la pasta)
2 cucharadas de sal (para cocinar la pasta)
2 cucharadas de aceite de oliva
3 tazas de salsa de tomate
3 dientes de ajo picados
500 g de albóndigas o carne molida (opcional)
Queso rallado al gusto (parmesano o mozzarella)
Hierbas frescas como albahaca o perejil para decorar (opcional)

**Instrucciones**

1. Cocina los rigatoni siguiendo las instrucciones del paquete en agua con sal. Escúrrelos cuando estén al dente.

2. Mientras se cocina la pasta, calienta el aceite de oliva en una sartén grande a fuego medio. Agrega el ajo picado y saltea hasta que esté fragante.

3. Si estás usando albóndigas o carne molida, agrégala a la sartén y cocina hasta que esté bien dorada y cocida.

4. Vierte la salsa de tomate en la sartén y deja que hierva a fuego lento. Puedes sazonar con sal y pimienta al gusto.

5. Mezcla la pasta cocida con la salsa en la sartén, asegurándote de que esté bien cubierta.

6. Sirve el rigatoni en platos individuales y espolvorea queso rallado por encima. Decora con hierbas frescas si lo deseas.

**Información que será ocultada**

**Coordenadas operación militar: Aures (Cauca) 03°06′00″N - 76°45′00″W UT-5:00**
**Información:** *Despliegue Delta-Niner en curso. Todos los escuadrones preparados para la operación. Confirmo coordenadas y autorizo el despegue. Mantengan la formación y sigan el plan de vuelo designado. Buena suerte y vuelvan con seguridad a la base. ¡Despeguen!*

Para hacer esto, podemos utilizar el siguiente algoritmo:
a) Convertir el mensaje a binario.
b) Convertir la receta a binario.
c) Combinar los dos mensajes binarios.
d) Convertir el mensaje combinado a texto.
e) Codificar el mensaje combinado en Java.

Para hacer esto, necesitamos un algoritmo que nos permita convertir texto a binario y viceversa. Podemos utilizar el siguiente algoritmo para hacer esto:

a) Convertir cada letra del texto a su valor ASCII.
b) Convertir cada valor ASCII a su representación binaria.
c) Concatenar todas las representaciones binarias.
d) Una vez que tengamos el mensaje y la receta en binario, podemos combinarlos simplemente concatenados. Luego, podemos convertir el mensaje combinado de nuevo a texto y modificarlo en Java.

**Código en Java**

```
import java.util.Base64;


public class Steganography {

    public static void main(String[] args) {

        String message = "Despliegue Delta-Niner en curso. Todos los escuadrones preparados para la operación. Confirmo coordenadas y autorizo el despegue. Mantengan la formación y sigan el plan de vuelo designado. Buena suerte y vuelvan con seguridad a la base. ¡Despeguen!";

        String recipe = "500 g de rigatoni\nAgua (para cocinar la pasta)\n2 cucharadas de sal (para cocinar la pasta)\n2 cucharadas de aceite de oliva\n3 tazas de salsa de tomate\n3 dientes de ajo picados\n500 g de albóndigas o carne molida (opcional)\nQueso rallado al gusto (parmesano o mozzarella)\nHierbas frescas como albahaca o perejil para decorar (opcional)";

        String combined = combineMessages(message, recipe);

        String encoded = encodeMessage(combined);

        System.out.println(encoded);

    }


    public static String combineMessages(String message, String recipe) {

        String binaryMessage = textToBinary(message);

        String binaryRecipe = textToBinary(recipe);

        return binaryMessage + binaryRecipe;

    }


    public static String textToBinary(String text) {

        byte[] bytes = text.getBytes();
```

```
        StringBuilder binary = new StringBuilder();
        for (byte b : bytes) {
            int value = b;
            for (int i = 0; i < 8; i++) {
                binary.append((value & 128) == 0 ? 0 : 1);
                value <<= 1;
            }
        }
        return binary.toString();
    }

    public static String encodeMessage(String message) {
        byte[] bytes = message.getBytes();
        String encoded =
Base64.getEncoder().encodeToString(bytes);
        return encoded;
    }
}
```
[21]

[21] Este código toma el mensaje y la receta, los combina y los codifica en Base64. Para decodificar el mensaje, simplemente se puede decodificar el mensaje en Base64 y luego separar el mensaje y la receta. Así mismo se genera un canal seguro desde el decodificador partiendo de la premisa de seguir su rastro digital, el cual es invisible para los posibles vunerabuladores de la información.

**14.5.Herramientas y recursos tecnológicos (software específico, bibliotecas de lingüística computacional)**

- Steghide: Esta herramienta de esteganografía destaca por su versatilidad al permitir esconder datos en varios tipos de archivos. Su aplicación se extiende a través de diferentes formatos, ofreciendo una solución integral para ocultar información de manera eficiente.

- OpenPuff: Este software combina capacidades de esteganografía y cifrado para brindar una doble capa de protección a los datos sensibles. Su utilidad radica en ocultar y resguardar información de manera segura, proporcionando una opción integral para la seguridad de los archivos.

- OutGuess: Especializado en archivos de imagen, OutGuess utiliza algoritmos de compresión para ocultar información de forma efectiva. Su enfoque en imágenes lo convierte en una herramienta específica para aquellos que buscan ocultar datos en este tipo de archivos.

- StegFS: Este programa va más allá al ser un sistema de archivos cifrado que incorpora características de esteganografía. Ofrece una solución integral para aquellos que buscan proteger sus datos mediante un cifrado avanzado junto con la capacidad de ocultar información.

- Invisible Secrets: Diseñado para la protección de datos sensibles, Invisible Secrets combina esteganografía y cifrado. Su enfoque integral proporciona una capa adicional de seguridad, permitiendo ocultar información de manera efectiva mientras se mantiene la confidencialidad.

- QuickStego: Con un diseño fácil de usar, QuickStego es una herramienta que permite ocultar datos en imágenes de manera rápida y sencilla. Su simplicidad lo hace accesible para usuarios que buscan una solución práctica y rápida para la esteganografía.

- LSB-Steganography: Esta implementación de esteganografía se centra en el método de bits menos significativos (Least Significant Bit). Proporciona una opción específica para aquellos que desean utilizar este enfoque particular en la ocultación de datos.

- Camouflage: Este software de esteganografía destaca por su capacidad para ocultar archivos en imágenes. Proporciona una opción discreta y efectiva para aquellos que buscan integrar datos en archivos visuales de manera encubierta.

- Hide and Seek: Utilizando técnicas de cifrado, Hide and Seek es una herramienta de esteganografía que se especializa en ocultar datos. Su enfoque en la seguridad de la información lo convierte en una opción valiosa para aquellos que buscan proteger sus archivos de manera efectiva.

- ImageHide: Centrado en archivos de imagen, ImageHide se especializa en ocultar datos de manera específica en este tipo de archivos. Su enfoque específico lo hace ideal para aquellos que buscan una solución especializada en la esteganografía visual.

# 15. CONCLUSIONES

La esteganografía, especialmente combinada con la criptografía, es una herramienta poderosa que permite a las personas comunicarse sin posibles escuchas, incluso sabiendo que existe una forma de comunicación en primer lugar. Los métodos utilizados en la ciencia de la esteganografía han avanzado mucho en los últimos siglos, especialmente con el auge de la era informática. La mayoría de las técnicas que se pueden usar en imágenes también se pueden aplicar en archivos de audio. La mejor forma de eliminar mensajes ocultos de un texto plano puede ser reescribir y reformular los contenidos. Reescribir usando diferentes palabras y construcciones de oraciones seguramente eliminará todas las formas de reproducir un mensaje oculto, ya que se encargará de casi todas las formas posibles en que los datos pueden almacenarse dentro de un texto sin formato.

El esquema de posición de los caracteres ya no funcionará porque se han cambiado las palabras, y lo mismo vale para las diferenciaciones en espacios en blanco, ya que el texto tendrá un nuevo diseño. En un futuro próximo, el uso más importante de las técnicas esteganográficas probablemente residirá en el campo de la marca de agua digital. Los proveedores de contenido están ansiosos por proteger sus obras con derechos de autor contra la distribución ilegal y las marcas de agua digitales brindan una forma de rastrear a los propietarios de estos materiales. Aunque no impedirá la distribución en sí, permitirá que el proveedor de contenidos inicie acciones legales contra los infractores de los derechos de autor, ya que ahora pueden ser rastreados. Existen diferentes técnicas y continúan desarrollándose, mientras que las formas de detectar mensajes ocultos también avanzan rápidamente. Dado que la detección nunca puede garantizar la búsqueda de toda la información oculta, se puede utilizar junto con métodos para derrotar a la esteganografía, a fin de minimizar las posibilidades de que se produzca una comunicación oculta.

Incluso entonces, la esteganografía perfecta, en la que la clave secreta simplemente señalará partes de una fuente encubierta que forman el mensaje, pasará desapercibida, porque la fuente encubierta no contiene información alguna sobre el mensaje secreto. La esteganografía también podría verse limitada por las leyes, ya que los gobiernos ya afirmaron que los delincuentes usan estas técnicas para comunicarse. Más restricciones en el uso de tecnologías de protección de la privacidad no son muy improbables, especialmente en este período de tiempo con gran ansiedad de ataques terroristas y de otro tipo. Es aquí donde la lingüística puede aportar herramientas para el entendimiento y detección temprana de peligros potenciales. La seguridad cibernética no debe basarse meramente en matemática y criptografía, sino también del arte y los procesos del lenguaje que podemos generar en las computadoras. Este segundo avance nos puede ayudar a intentar generar una aplicación y profundización lingüística en el futuro proceso, pensado desde las redes sociales hacia una aplicación en lengua española.

**15.1. Recomendaciones:** Claramente, la esteganografía es un campo que demanda dos aspectos fundamentales desde la perspectiva lingüística. En primer lugar, se requiere la especialización de estudiantes en diversas áreas que abarquen desde la fonética hasta la pragmática, semiótica, entre otros; en conjunto con disciplinas como la lingüística computacional. Dada la riqueza del español como lengua, la anticipación se presenta como un elemento crucial. El segundo aspecto a desarrollar debe establecerse en la intersección entre la lingüística y la seguridad, considerando todos los factores relacionados con las redes y aspectos militares. Esta colaboración integral es esencial para abordar los desafíos específicos de la esteganografía en el contexto lingüístico, garantizando una respuesta efectiva y adaptada a las complejidades de la lengua y los factores de seguridad involucrados.

**15.2. Resumen de los Hallazgos Clave**: En el contexto colombiano, la necesidad de implementar esteganografía lingüística como parte de políticas independientes de proveedores internacionales de ciberseguridad se destaca como una medida crucial. La lingüística ofrece canales esteganográficos que pueden ser fundamentales para establecer estudios y estrategias de seguridad adaptadas a la realidad colombiana. La investigación revela que la implementación de técnicas esteganográficas en el ámbito lingüístico puede contribuir significativamente a las luchas geopolíticas y sociales, así como a los desarrollos militares en ciencia y tecnología. En el contexto colombiano, marcado por desafíos únicos y una compleja dinámica sociopolítica, la esteganografía lingüística emerge como una herramienta esencial para fortalecer la seguridad y la criptografía independiente a proveedores.Un punto de partida, a futuro, con más información podría ser " El ataque IFX" un proveedor importante de ciberseguridad en Colombia, el cual ocurrió en 2023. [22]

**15.3. Contribuciones al Campo:** Esta línea de investigación ha aportado insights valiosos al campo de la ciberseguridad en Colombia. La integración de esteganografía lingüística como parte de las políticas independientes de seguridad ofrece una perspectiva innovadora para abordar los desafíos específicos del entorno colombiano. Se destaca la importancia de adaptar las estrategias de seguridad a las características lingüísticas y sociopolíticas únicas de la región. Además, se resalta la contribución al ámbito militar y tecnológico, donde la esteganografía lingüística puede desempeñar un papel crucial en operaciones encubiertas y protección de información sensible. Estos hallazgos sugieren que el desarrollo de capacidades esteganográficas específicas puede ser fundamental para fortalecer la posición de Colombia en materia de seguridad nacional.

**15.4. Recomendaciones para Futuras Investigaciones:** Con base en los resultados obtenidos, se recomienda llevar a cabo investigaciones adicionales que profundicen en las aplicaciones prácticas de la esteganografía lingüística en entornos específicos, como el gubernamental, militar y empresarial. Además, se sugiere explorar nuevas formas de optimización para identificar y prevenir actividades criminales desde las entidades gubernamentales, aprovechando las capacidades esteganográficas.

La investigación futura podría también abordar la implementación de políticas específicas que fomenten el desarrollo y la adopción de tecnologías esteganográficas en el ámbito nacional. Esto incluiría la colaboración con expertos en lingüística, criptografía y ciberseguridad para crear estrategias integrales y adaptadas a la realidad colombiana. En conclusión, la esteganografía lingüística se presenta como una herramienta estratégica clave en el ámbito de la seguridad colombiana, ofreciendo soluciones innovadoras y adaptadas a los desafíos únicos de la región.[23]

---

[22] Ver "El ataque cibernético que sacude a Colombia" en: https://www.javeriana.edu.co/pesquisa/ciberataque-ifx-networks-colombia/

[23] La principal entidad en el campo de la lingüística computacional es la Association for Computational Linguistics (ACL). Además, existen ramas regionales como la European Association for Computational Linguistics en Europa y la American Association for Computational Linguistics en América. Cada dos años, se celebra la International Conference on Computational Linguistics (COLING), donde se presentan documentos de interés para aquellos que desean seguir de cerca el desarrollo de la lingüística computacional.Otros congresos relevantes para los lingüistas computacionales incluyen la bienal de Applied Natural Language Processing (ANLP), el congreso de Empirical Methods in Natural Language Processing (EMNLP) y la International Conference on Spoken Language Processing (ICSLP).

**16, REFERENCIAS**

Aho, Alfred; Hopcroft, John; Ullman, Jeffrey. (1983). Data Structures and Algorithms. Reading, MA: Addison Wesley Publishing Company.

Bar Adon, A. (1990). Las madres y su aporte en el resurgimiento y consolidación del hebreo (1882-1914), en Lashon veIvrit 3, Jerusalén, editorial Nimá, pp. 5-25. [En hebreo].

Blanc, H. (1956). "A note on Israeli Hebrew 'psycho-phonetics'". Word, 12, pp. 106-114.

Cano, J. (2021). Modos de operación de la ciberseguridad empresarial. Capacidades básicas para navegar en el contexto digital. Global Strategy. Global Strategy Report No. 44. https://global-strategy.org/modos-de-operacion-de-la-ciberseguridad-empresarial-capacidades-basicas-para-navegar-en-el-contexto-digital/

Cole, Ronald A. (1997). Survey of the State of the Art in Human Language Technology. Cambridge: Cambridge University Press.

Colomina et al. (2022). El mundo en 2023: diez temas que marcarán la agenda internacional. CIDOB Notes Internationals. No. 238. https://bit.ly/3YHt7uK

DAPF (s.f.) Decreto 1010 de 2000. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=64913

Dzienciarsky, D. (2019). A discriminação sociolinguística dos judeos magrebinos, en Arquivo Maaravi v. 13 n. 25, Belo Horizonte, editorial UFMG, pp. 199-212. https://doi.org/10.17851/1982-3053.13.25.199-212.

EPM (2022). Resultados financieros. Primer trimestre 2022. https://www.epm.com.co/site/Portals/6/documentos/informe-1t2022.pdf

EPM (s.f.). Grupo EPM. https://www.grupo-epm.com/site/ticsa/institucional/grupo-epm

Fellman, J. (1990). Eliezer Ben-Yehuda y el renacimiento de la lengua hebrea, en Leshonenu 40-volumen del compendio del año del hebreo, editorial de la Academia de lengua hebrea, Jerusalén, pp. 215-221. [En hebreo].

Frank, S. L. (2009). "Surprisal-based comparison between a symbolic and a connectionist model of sentence processing". In Proceedings of the 31st Annual Conference of the Cognitive Science Society, pp. 1–8.

Grune, Dick; Jacobs, Ceriel J. H. (1990). Parsing Technique: A Practical Guide. Chichester: Ellis Horwood Limited. Also available at: http://www.cs.vu.nl/~dick/PTAPG.html.

Guarnieri, F., Fliss, M., & Bancroft, C. (1996). Making DNA Add. Science, 273, 220–223.

Guinovart, J. G. (1998). Fundamentos de Lingüística Computacional: bases teóricas, líneas de investigación y aplicaciones.

Halliday, Michael A. K. (1985). An Introduction to Functional Grammar. London: Edward Arnold.

Halvorsen, K., & Wong, W. P. (2012). Binary DNA nanostructures for data encryption. PLoS ONE, 7, e44212.

Hale, J. T. (2001). "A probabilistic Earley parser as a psycholinguistic model". In Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies, pp. 1–8.

Hutchins, W. John. (1986). Machine Translation: Past, Present, Future. Chichester: Ellis Horwood.

Infobae (2022). Invima: sitio web fue blanco de ciberataque. https://www.infobae.com/america/colombia/2022/02/09/invima-sitio-web-fue-blanco-de-ciberataque/

Jelinek, Frederik. (1999). Statistical Methods for Speech Recognition. Cambridge: MIT Press.

Jonoska, G. Păun, & G. Rozenberg (Eds.), Aspects of molecular computing: essays dedicated to Tom Head, on the occasion of his 70th birthday (pp. 167–188). Berlin: Springer.

Kahn, D. (1967). The code breakers: The comprehensive History of Secret Communication from Ancient Times to the Internet. New York: Scribner.

Karmi, Sh. (1997). ONE People ONE Language: The revival of the Hebrew language en una perspectiva interdisciplinaria, Tel Aviv, editorial del Ministerio de Seguridad. [En hebreo].

Kay, Martin. (1984). Functional Unification Grammar: A Formalism for Machine Translation. In Coling-84. 10th International Conference on Computational Linguistics. 22nd Annual Meeting of the Association for Computational Linguistics. 1984. New Jersey, Morristown. 75-78.

Kutscher, E. Y. (1982). A History of the Hebrew Language. Jerusalem: The Magnes Press.

Lázaro Carreter, F. (1980). "Sobre la pasiva en español". Estudios de lingüística, 61-70. Barcelona: Crítica.

Lee, J. W., Chan, C. T. Y., Slomovic, S., & Collins, J. J. (2018). Next-generation biocontainment systems for engineered organisms. Nature Chemical Biology, 14, 530–537.

Leier, A., Richter, C., Banzhaf, W., & Rauhe, H. (2000). Cryptography with DNA binary strands. Biosystems, 57, 13–22.

Levy, R. (2008). "Expectation-based syntactic comprehension". Cognition 106, 1126–1177.

Lipton, R. (1995). DNA solution of hard computational problems. Science, 268, 542–545.

Manning, Christopher D. & Schütze, Hinrich. (1999). Foundations of Statistical Natural Language Processing. MA: MIT Press, Cambridge.

Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MA: MIT Press, Cambridge.

McKeown, Kathy. (1984). Text Generation. Cambridge, MA: Cambridge University Press.

Mirkin, R. (1989). Capítulos sobre la historia de los diccionarios hebreos de las personas privadas, en Leshonenu 53, volumen del compendio del año del hebreo, editorial de la Academia de lengua hebrea, Jerusalén, pp. 7-56. [En hebreo].

Molino, A., Lewin, I., Jofen, A., & Marantz, A. (2019). A multifactorial approach to the study of language disorders: A case study of an English-Hebrew speaker. Frontiers in Psychology, 10, 1393.

Morimoto, T., & Thorpe, M. F. (2001). Unsupervised construction of large phonetic lexica. In Proceedings of the 7th International Conference on Spoken Language Processing, pp. 665–668.

Nunberg, G. (1979). "The non-uniqueness of semantic solutions: polysemy". Linguistics and Philosophy, 3, 143–184.

Pantel, P., & Lin, D. (2002). "Discovering word senses from text". In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 613–619.

Peleg, Y. (1991). Estudios y ensayos lingüísticos, en Leshonenu 55, volumen del compendio del año del hebreo, editorial de la Academia de lengua hebrea, Jerusalén, pp. 277-287. [En hebreo].

Pinkas, A. (2019). Punctuation annotation in the Hebrew Bible. Literary and Linguistic Computing, 34(4), 763–770.

Rau, J., & Mezulis, S. (2019). Chasing the triple bottom line: Strengthening public-private partnerships for sustainable development. OECD Development Co-operation Working Papers, No. 58, OECD Publishing, Paris.

Rodríguez, J. E. (2020). La contribución de las mujeres a la lingüística computacional, en Revista Española de Lingüística Computacional, 36, pp. 189-204.

Sag, Ivan A. (2010). English Grammar in English and Russian. Stanford: CSLI Publications.

Sag, I. A., Wasow, T., & Bender, E. M. (2003). Syntactic Theory: A Formal Introduction (2nd ed.). Stanford, CA: CSLI Publications.

Sánchez, E. (2022). Ciberseguridad: retos y tendencias en 2022. Global Strategy. Global Strategy Report No. 45. https://global-strategy.org/ciberseguridad-retos-y-tendencias-en-2022/

Shieber, Stuart M. (1986). An Introduction to Unification-Based Approaches to Grammar. Stanford: CSLI Publications.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), 1929–1958.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., ... & Van Ess-Dykema, C. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics, 26, 339–373.

Stoyanovich, J., Howe, B., Abiteboul, S., & Miklau, G. (2017). "Fides: Towards a platform for responsible data science". Proceedings of the VLDB Endowment, 10(12), 1854–1857.

Teahan, W. J. (1998). A probabilistic model of lexical and syntactic access and disambiguation. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1, pp. 898–902.

Thomas, M., Pang, B., & Lee, L. (2006). "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts". In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 327–335.

Turing, A. M. (1936). "On Computable Numbers, with an Application to the Entscheidungsproblem". Proceedings of the London Mathematical Society, 2(42), 230–265.

Varshney, L. R., Chen, Y., & Li, Y. (2011). Binary code analysis and optimization. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 41, 1597–1609.

Volk, Martin; Edelman, Shimon. (1985). A Computational Model of Natural Language Communication: Interpretation, Inference, and Production. In Artificial Intelligence and Natural Language, eds. M. Brady & R. C. Berwick, 319-358. Academic Press, Orlando.

Vossen, P., Aitenbichler, E., Dengel, A., Hügle, A., Immenhauser, B., Kett, H., ... & Polthier, K. (2001). Towards a multimedia global information system. Computers & Graphics,

**CODE LIBRARY**

FROM:https://github.com/maniksingh92-xx/simple-steganography/blob/master/src/SteganoImgProcess.java

```
/* IMAGE STEGANOGRAPHY : SteganoImgProcess.java
 *
 * This source code form is protected under GPL v2.0.
 * To read more about the license, follow the link:
 * http://www.gnu.org/licenses/gpl-2.0.html
 *
 * Created by Manik Singh <maniksingh92@live.com> as a project
 * during the summer training.
 *
 * Initiating date : 7-JUL-2014
 *
 * >> SteganoImgProcess.java contains two methods encode(), for
 * encoding text into image; and decode(), for decoding text from
 * an image.
 */

import java.awt.image.BufferedImage;
import java.io.File;
import java.io.IOException;
import java.util.ArrayDeque;
import java.util.Deque;

import javax.imageio.ImageIO;

public class SteganoImgProcess {
```

```
String ext;
int encodedMsgOffset;

//Method for encoding text into source image
boolean encode(BufferedImage input, BufferedImage output, int width, int height, String msg, String outputName) {


int msgLength = msg.length(); //original message length
//add overhead "!encoded!" to identify start of encoded message,
//along with original message length which will instruct when to stop
//the decoding process, once the encoded message has been extracted
String message = "!encoded!" + msgLength + "!" + msg;
msgLength = message.length(); //message length included the overhead

//Creating an array of integers which will hold the entire message
//in the form of bit format. A message is made up of characters,
//each character is of the size of one byte. 1 byte = 8 bits.
//Each byte here is divided into 4 parts of 2 bits, and stored in
//the twoBitMessage array. Thus, the size of twoBitMessage is
//4 times the length of the actual message.
int[] twoBitMessage = new int[4 * msgLength];

char currentChar;
for(int i =0; i < msgLength ; i++) {
        currentChar = message.charAt(i); // extracting character at position i from string
        twoBitMessage[4*i + 0] = (currentChar >> 6) & 0x3; //storing 1st and 2nd bit from the left
```

```
                twoBitMessage[4*i + 1] = (currentChar >> 4) & 0x3; //storing 3rd
and 4th bit from the left
                twoBitMessage[4*i + 2] = (currentChar >> 2) & 0x3; //storing 5th
and 6th bit from the left
                twoBitMessage[4*i + 3] = (currentChar)      & 0x3; //storing 7th and
8th bit from the left
            }

            int pixel, pixOut, count = 0;;
            loop: for(int i = 0; i < width; i++) {
                for(int j = 0; j < height; j++) {
                    if(count < 4*msgLength) { //ensuring that loop only iterates
till the entire message has been encoded
                        pixel = input.getRGB(i, j); //Grab the RGB value
from the pixel of the source image at position (i,j)

                        //Bit operator AND is used to convert the two LSB to
zero.
                        //Once that is established, bit operator OR is used to
copy
                        //the two bit message in place of these two LSB.
                        //Thus, we encode two bits of our message in one
pixel.
                        //Effectively, 4 pixels, carrying 8 bits of encoded bits
                        //in total, will carry the information of one character.
                        pixOut = (pixel & 0xFFFFFFFC) |
twoBitMessage[count++]; //modified RGB value with encoded data

                        output.setRGB(i, j, pixOut); //Set the modified RGB
value at given pixel.
```

```
                        } else {
                                break loop;
                        }
                }

        }

        try {
                ImageIO.write(output, "png", new File(outputName)); //create .PNG
file with encoded image data
                return true;
        } catch (IOException e) {
                e.printStackTrace();
                return false;
        }

}// end of encode()


String decode(BufferedImage input, int width, int height) {

        if(!isEncoded(input, width, height)) {
                return null;
        }

        int msgLength = getEncodedLength(input, width, height);

        StringBuffer decodedMsg = new StringBuffer();
        Deque<Integer> listChar = new ArrayDeque<Integer>();
```

```
                int pixel, temp, charOut, ignore = 0, count = 0;
                loop: for(int i = 0; i < width; i++) {
                        for(int j = 0; j < height; j++) {
                                if(ignore < 36 + 4*(String.valueOf(msgLength).length()+1))
{
                                        ignore++;
                                        continue;
                                }

                                if(count++ == 4*msgLength) {
                                        break loop;
                                }
                                pixel = input.getRGB(i, j); //grab RGB value at specified
pixel
                                temp = pixel & 0x03; //extract 2 LSB from encoded data

                                listChar.add(temp); //add the extracted data to a queue for
later processing

                                if(listChar.size() >=4) { //once we have 8 bits of data
extracted
                                        //combine them to create a byte, and store this byte as
a character
                                        charOut = (listChar.pop() << 6) | (listChar.pop() <<
4) | (listChar.pop() << 2) | listChar.pop() ;
                                        decodedMsg.append((char)charOut);
                                }
                        }

                }
```

```
            String outputMsg = new String(decodedMsg); //generate extracted message

            return outputMsg;
        } //end of decode()

        boolean isEncoded(BufferedImage input, int width, int height) { //Check for "!encoded!" at starting

            StringBuffer decodedMsg = new StringBuffer();
            Deque<Integer> listChar = new ArrayDeque<Integer>();

            int pixel, temp, charOut, count = 0;
            loop: for(int i = 0; i < width; i++) {
                for(int j = 0; j < height; j++, count++) {

                    if(count == 45) { //remain in loop till first 9 characters are extracted
                        break loop;
                    }
                    pixel = input.getRGB(i, j); //grab RGB value at specified pixel
                    temp = pixel & 0x03; //extract 2 LSB from encoded data

                    listChar.add(temp); //add the extracted data to a queue for later processing

                    if(listChar.size() >=4) { //once we have 8 bits of data extracted
                        //combine them to create a byte, and store this byte as a character
```

```
                                        charOut = (listChar.pop() << 6) | (listChar.pop() << 4) | (listChar.pop() << 2) | listChar.pop() ;
                                        decodedMsg.append((char)charOut);     //else     add character to a StringBuffer
                                        count++;
                                    }
                                }
                            }

                            String check = new String(decodedMsg);
                            System.out.println(check + " " + check.length());
                            if (check.compareTo("!encoded!") == 0) {
                                System.out.println("true");
                                return true;
                            } else {
                                return false;
                            }

        } //end of isEncoded() method

        int getEncodedLength(BufferedImage input, int width, int height) { //method to get actual length of message encoded

                            StringBuffer decodedMsg = new StringBuffer();
                            Deque<Integer> listChar = new ArrayDeque<Integer>();

                            int pixel, temp, charOut, count = 0;
                            loop: for(int i = 0; i < width; i++) {
                                for(int j = 0; j < height; j++) {
                                    if(count < 36) { //ignore the 36 bits or 9 bytes, equal to "!encoded!"
```

```
                        count++;
                        continue;
                    }

                    pixel = input.getRGB(i, j); //grab RGB value at specified pixel
                    temp = pixel & 0x03; //extract 2 LSB from encoded data

                    listChar.add(temp); //add the extracted data to a queue for later processing

                    if(listChar.size() >=4) { //once we have 8 bits of data extracted
                        //combine them to create a byte, and store this byte as a character
                        charOut = (listChar.pop() << 6) | (listChar.pop() << 4) | (listChar.pop() << 2) | listChar.pop() ;
                        if((char)charOut == '!') { //terminate process if character extracted is '!'
                            break loop;
                        } else {
                            decodedMsg.append((char)charOut);      //else add character to a StringBuffer
                        }
                    }
                }

            }

            String length = new String(decodedMsg);
```

```
            System.out.println("length is " + Integer.parseInt(length));

            return Integer.parseInt(length);
        } //end of getEncodedLength() method

        String getExt() {
            return ext;
        }

        int getOffset() {
            return encodedMsgOffset;
        }
}
```

**Illinois Institute Of Technology**

**ITM 311 LAB 08 SP 19 STEGANOGRAPHY AND ARRAYS.docx**

**PROJECT:** Steganography and Secret Messages - Using Arrays in Java
***Objective: To type a simple Java program, execute ( run ) the program for some particular values, observe the output and then modify the program.***

**PROJECT DESCRIPTION:** Type, compile and run the basic Java program that is shown inFigure 1, which follows. Then compile and run your program, observe the output then modify the program. Information About This Project Steganography is related to cryptography and attempts to disguise a message in a plaintext statement block. Consider receiving the text message that is shown below. the plaintext message reads as: Hello Stan and Marilyn! Let us meet for lunch at noon on Tuesday. We will be waiting at the East building of train station C. I will ask Hank, Quincy, Denise, Eddy and Roni to join us there! Best Regards, KZ Using the grid below, write the above message into the cells below - use one letter per cell.Use lower case letters. 012345678910111213141516171819202122 2 3 line 1HelloStanandMarilyn! line 2Letusmeetforlunch line 3atnoononTuesday. line 4Wewillbewaiting line 5attheEastbuildingof line 6trainstationC. line 7IwillaskHank,Quincy, line 8Denise,EddyandRoni line 9tojoinusthere! line10 BestRegards, line11 KZ line 12 Then attempt to extract a complete message from the plaintext. A partialexample of a practical covert, spy - like message that can be taken from theplaintext would be: time to launch drone BETA

[1] Akyildiz I. F., Su W., Sankarasubramaniam Y. and Cayircl E., “A survey on sensor networks”, in: Proceedings of the IEEE Communication Magazine, Vol. 40, pp. 102-114, August 2002.

[2] Ilyas M. and Mahgoub I., “Handbook of Sensor Networks: Compact Wireless and Wired Sensing Systems”, in: Proceedings of the CRC Press, London, Washington, D.C., 2005.

[3] Kahn J.M., Katz R.H. and Pister K.S.J., “Next century challenges: mobile networking for smart dust”, in: Proceedings of the ACM MobiCom 99, Washington, USA, pp. 271–278,1999.

[4] D. Chen and K. Varshney, “QoS Support in Wireless Sensor Networks: A Survey” Department of EECS, Syracuse University Syracuse, NY, U.S.A 13244, 2004

[5] A. Ganz, Z. Ganz, and K. Wongthavarawat, Multimedia Wireless Networks:Technologies, 'Standards, and QoS , Prentice Hall, Upper SaddleRiver, NJ, 2004.

[6] E. Crawley et al., “A Framework for QoS-Based Routing in the Internet, ” RFC 2386, http://www.ietf.org/rfc/rfc.2386.txt, Aug. 1998.

[7] Z. Demetrios, “A Glance at Quality of Services in Mobile Ad-Hoc Networks, ” http://www.cs.ucr.edu/ csyiazti/cs260.html, November 2001.

[8] D. Zeinalipour, S. Aristeidou, S. Kazeli, “IP Quality of Services (in Greek), ” http://www.cs.ucr.edu/ csyiazti/downloads/papers/ipqos/ papers/ip-qos.pdf, 1999

[9] K. Wui, J. Harms, “QoS Support in Mobile Ad Hoc Networks, ” Crossing Boundaries – an interdisciplinary Journal, Vol 1, No 1, Fall 2001.

[10] S. Chakrabarti and A. Mishra, “QoS Issues in Ad Hoc Wireless Networks, ” IEEE Communications Magazine, pp. 142-148, February 2001.

[11] S. Meguerdichian, F. Koushanfar, M. Potkonjak, and M. B. Srivastava, “Coverage Problems in Wireless Ad-hoc Sensor Networks, ” in proceedings of IEEE Infocom, 2001, pp. 1380-1387

[12] S. Meguerdichian, F. Koushanfar, G. Qu, and M. Potkonjak, “Exposure in Wireless Ad-hoc Sensor Networks, ” in Mobile Computing and Networking, 2001, pp. 139-150.

[13] R. Iyer and L. Kleinrock, “QoS Control for Sensor Networks, ” in ICC 2003, May 2003.

[14] S. Tilak, N. Abu-Ghazaleh and W. Heinzelman, “A taxonomy of wireless micro-sensor network communication models, ” ACM Mobile Computing and Communication Review(MC2R), June 2002.

[15] Narendra K. S., Thathachar M. A. L.; “Learning automata: An introduction”; Prentice Hall, 1989.

[16] Narendra K.S., Thathachar M. A. L. “Learning automata a survey”; IEEE Transactions on Systems, Man and Cybernetics, vol. 4, no. 4, July 1974.

[17] Mance, E., and Stephanie, S., H., “Reinforcement learning: A tutorial,” Wright Laboratory, 1996.

[18] Sutton, R. S., and Barto, A.G.; “Reinforcement learning: Introduction”; MIT Press, 1998.

[19] Lakshmivarahan S., Thathachar M. A. L.; “Absolutely expedient learning algorithms for stochastic automata”; IEEE Transactions on Systems, Man and Cybernetics, vol. 6, pp. 281-286, 1973.

[20] Mars, P., Chen, J. R., and Nambiar, R., Learning algorithms theory and applications in signal processing, control and communications, CRC Press, 1996.

[21] Lakshmivarahan S.; “Learning algorithms: theory and applications”; New York: Springer-Verlag, 1981.

[22] Lakshmivarahan S., Thathachar M. A. L.; “Optimal non-linear reinforcement schemes for stochastic automata”; Information Science, vol. 4, pp. 121-128, 1982.

[23] Lakshmivarahan S., Thathachar M. A. L.; “Absolute expediency of Q and S-model learning algorithm”; IEEE Transactions on Systems, Man and Cybernetics, vol. 6, pp. 222-226, 1976.

[24] Viswanathan R., Narendra K. S.; “Expedient and optimal variable structure stochastic automata”; Technical report CT-31, Dunham Lab., Yale University, New Haven, Connecticut, U.S.A., April 1970.

[25] Mason L.G.; “An optimal learning algorithm for s-model environments”; IEEE Transactions on Automatic Control, pp. 493-496, October 1973.

[26] Thathachar M. A. L., Bhaskar R. H.; “Learning automata with changing number of actions”; IEEE Transactions on Systems, Man and Cybernetics, vol. 17, no. 6, November 1987.

[27] Meybodi M. R., Beigy H., Taherkhani M.; “Cellular Learning Automata and Its Applications”; Journal of Science and Technology, University of Sharif, No. 25, pp.54-77, Autumn/Winter 2003-2004.

[28] Meybodi M. R., Kharazmi M. R.; “Cellular Learning Automata and Its Application to Image Processing”; Journal of Amirkabir, Vol. 14, No. 56A, pp. 1101-1126, 2004.

[29] Meybodi M. R, Kharazmi M. R., “Image Restoration Using Cellular Learning Automata”, in Proceedings of the Second Iranian Conference on Machine Vision, Image Processing and Applications, Tehran, Iran, pp. 261-270, 2003.

[30] Meybodi M. R., Khojaste M. R., “Application of Cellular Learning Automata in Modeling of Commerce Networks”, in Proceedings of 6th Annual International Computer Society of Iran Computer Conference CSICC2001, Isfahan, Iran, pp. 284-295, 2001.

[31] Meybodi M. R., Mehdipour F., “VLSI Placement Using Cellular Learning Automata”, in Proceedings of 8th Annual International Computer Society of Iran Computer Conference CSICC2001, Mashhad, Iran, pp. 195-203, 2003.

[32] Meybodi, M. R., and Taherkhani, M., “Application of Cellular Learning Automata to Modeling of Rumor Diffusion”, in Proceedings of 9th Conference on Electrical Engineering, Power and Water institute of Technology, Tehran, Iran, pp. 102-110, May 2001.

[33] Beigy H., Meybodi M. R.; “A Mathematical Framework for Cellular Learning Automata”; Advances on Complex Systems, Vol. 7, No. 3, pp. 1-25, 2004.

[34] V. Raghunathan, C. Schurgers, S. Park, and M. B. Srivastava, “Energy-Aware Wireless MicrosensorNetworks,” IEEE Signal Processing Magazine, Vol 19, pp 40-50, March 2002.

[35] M. Cardei and J. Wu “Coverage inWireless Sensor Networks” Department of Computer Science and Engineering Florida Atlantic University Boca Raton, FL 33431, 2004.

[36] S. Slijepcevic and M. Potkonjak, “Power Efficient Organization of Wireless Sensor Networks,” Proc. of IEEE International Conference on Communications, Vol 2, pp 472-476,Helsinki, Finland, June 2001.

[37] M. Cardei, D. MacCallum, X. Cheng, M. Min, X. Jia, D. Li, and D.-Z. Du, “Wireless Sensor Networks with Energy Efficient Organization”, Journal of Interconnection Networks, Vol 3, No 3-4, pp 213-229, Dec 2002.

[38] D. Tian and N. D. Georganas, “A Coverage-Preserving Node Scheduling Scheme for Large Wireless Sensor Networks,” Proc. of the 1st ACM Workshop on Wireless Sensor Networks and Applications, 2002.

[39] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, “Energy-Efficient Communication Protocols for Wireless Microsensor Networks,” Proc. of HICSS, Jan 2000.

[40] F. Ye, G. Zhong, S. Lu, and L. Zhang, “Energy Efficient Robust Sensing Coverage in Large Sensor Networks,” Technical Report UCLA, 2002.

[41] C.-F. Huang and Y.-C. Tseng. The coverage problem in a wireless sensor network. In ACM International Workshop on Wireless Sensor Networks and Applications (WSNA), 2003, pp. 115–121.

[42] M. Cardei and D.-Z. Du, “ImprovingWireless Sensor Network Lifetime through Power Aware Organization”, accepted to appear in ACM Wireless Networks.

[43] S. Meguerdichian, F. Koushanfar, M. Potkonjak, and M. Srivastava, “Coverage Problems in Wireless Ad-Hoc Sensor Networks,” IEEE Infocom 2001, Vol 3, pp 1380-1387, April 2001.

[44] G. Xing, X. Wang, Y. Zhang, C. Lu, R. Pless, and C. Gill, “Integrated coverage and connectivity configuration for energy conservation insensor networks,” ACM Transactions on Sensor Networks, vol. 1, no. 1,pp. 36–72, August 2005.

[45] Crossbow Technology, Inc., http://www.xbow.com/Products/Product_pdf_files/Wireless_pdf/6020-0042-01_A_MICA2.pdf.

[46] A. R. Shahani, D. K. Schaeffer, T. H. Lee, ”A 12 mW Wide Dynamic Range CMOS Front-End for a portable GPS Receiver”, Digest of Technical Papers, IEEE International Solid State Circuits Conference, Vol. 40, pp. 368-369, February 1997.

[47] S. Capkun, M. Hamdi, J. P. Hubaux, ”GPS-free Positioning in Mobile Ad-hoc Networks”, Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34), January 3-6, 2001.

[48] D. Niculescu, B. Nath, ”Ad Hoc Positioning System (APS)”, IEEE Global Telecommunications Conference 2001, Vol. 5 , pp. 2926-2931, November 2001.

[49] C. Savarese , K. Langendoen, J. Rabaey, ”Robust Positioning Algorithms for Distributed Ad-hoc Wireless Sensor Networks”, Proceedings of the 2002 USENIX Annual Technical Conference, Monterey, pp. 317-27, June 2002.

[50] D. Perkins, R. Tumati, ”Reducing Localization Errors in Sensor AdhocNetworks”, 23rd IEEE International Conference on Performance, Computing and Communications, Phoenix, pp. 723 – 729, April 2004.

[51] ] T. Yan, T. He, J. A. Stankovic, ” Differentiated Surveillance for Sensor Networks” SenSys’03, November 5–7, 2003, Los Angeles, California, USA.

[52] D. Tian and N. Georganas, “A coverage-preserving node scheduling scheme for large wireless sensor networks,” in Proc. of First ACM International Workshop on Wireless Sensor Networks and Applications, Atlanta, GA, September 2002, pp. 32–41.

[53] F. Ye, G. Zhong, J. Cheng, S. Lu, and L. Zhang, “Peas: A robust energy conserving protocol for long-lived sensor networks,” in Proc. of International Conference on Distributed Computing Systems (ICDCS’03),Providence, RI, May 2003, pp. 28–37.

[54] O. Younis and S. Fahmy, “Distributed Clustering in Ad-hoc Sensor Networks: A Hybrid, Energy-Efficient Approach”, In Proc. of IEEE INFOCOM, volume 1, pp. 629-640, March 2004.

[55] W.Heinzelman, A. Chandrakasan, and H. Balakrishnan, Energy-Efficient Communication Protocol for wireless microsensor networks, Proc of Hawaii conference on System Science, Jan 2000.

[56] Seema Bandopadhyay, Edward J.Coyle, An Energy efficient hierarchical clustering algorithm for wireless sensor networks, Proc. of IEEE INFOCOM 2003, San Francisco, CA.

[57] Srejan Raghuwanshi and Amitabh Mishra, A Self adpative clustering based algorithm for increased energy efficiency and scalability in WSN,Proc. IEEE VTC 2000.

[58] Gaurav Gupta, Mohammed Younis, Load balanced clustering in wireless sensor networks, proc of IEEE ICC 2003.

[59] Li Chunwang, Chung Wei Wang, A cross layer design of clustering architecture for WSN, Proc. of IEEE ICNSC, 2004 Taipei, Taiwan.

[60] Ewa Hansen, Jonas Neander, Mikael Nolin, Mats Bjorkman; Efficient Cluster Formation for Sensor Networks, MRTC report ISSN 1404-3041 ISRN MDH-MRTC-199/2006-1-SE, M?lardalen Real-Time ResearchCentre, M?lardalen University.

[61] S. Basagni, “Distributed Clustering Algorithm for Ad-hoc Networks,” in International Symposium on Parallel Architectures, Algorithms, and Networks (I-SPAN), 1999.

[62] S. Banerjee and S. Khuller, “A Clustering Scheme for Hierarchical Control in Multi-hop Wireless Networks,” in Proceedings of IEEE INFOCOM, April 2001.

[63] M. Gerla, T. J. Kwon, and G. Pei, “On Demand Routing in Large Ad Hoc Wireless Networks with Passive Clustering,” in Proceeding of WCNC,2000.

[64] C. R. Lin and M. Gerla, “Adaptive Clustering for Mobile Wireless Networks,” in IEEE J. Select. Areas Commun., September 1997.

[65] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, “An Application-Specific Protocol Architecture for Wireless MicrosensorNetworks,” IEEE Transactions on Wireless Communications, vol. 1,no. 4, pp. 660–670, October 2002.

[66] S. Bandyopadhyay and E. Coyle, “An Energy-Efficient HierarchicalClustering Algorithm for Wireless Sensor Networks,” in Proceedingsof IEEE INFOCOM, April 2003.

[67] D. Estrin, R. Govindan, J. Heidemann, and S. Kumar, “Next Century Challenges: Scalable Coordination in Sensor Networks,” in Proceedings of the ACM/IEEE International Conference on Mobile Computing and Networking (MOBICOM), August 1999.

[68] H. Chan and A. Perrig, “ACE: An Emergent Algorithm for Highly Uniform Cluster Formation,” in Proceedings of the First European Workshop on Sensor Networks (EWSN), January 2004.

[69] V. Mhatre and C. Rosenberg, “Design Guidelines for Wireless Sensor Networks Communication: Clustering and Aggregation,” Ad-hoc Networks ournal, 2004.

[70] J. Kulik, W. R. Heinzelman, and H. Balakrishnan, “Negotiation-Based Protocols for Disseminating Information in Wireless Sensor Networks,” ACM Wireless Networks, vol. 8, no. 2-3, pp. 169–185, 2002. [Online]. Available: citeseer.nj.nec.com/kulik99negotiationbased.html

[71] M. Esnaashari, M. R. Meybodi, “Irregular Cellular Learning Automata and Its Application to Clustering in Sensor Networks” ICEE, 2007

[72] Farajzadeh, N. and Meybodi, M. R., “Learning Automata-based Clustering Algorithm for Sensor Networks”, Proceedings of 12th Annual CSI Computer Conference of Iran, Shahid Beheshti University, Tehran, Iran, pp. 780-787, Feb. 20-22, 2007

[73] Esnaashari, M. and Meybodi, M. R., “A Cellular Learning Automata based Clustering Algorithm for Wireless Sensor Networks”, Sensor Letters, 2008, to appear

[74] M. Ilyas, I. Mahgoub, “Handbook of Sensor Networks : Compact Wireless and Wired Sensing Systems”, CRC Press, London, Washington, D.C., 2005.

[75] K. Akkaya, M. Younis, “A Survey on Routing Protocols for Wireless Sensor Networks”, Elsevier Ad Hoc Network Journal, pp. 325-349, 2005.

[76] R. Shah and J. Rabaey, “Energy Aware Routing for Low Energy Ad Hoc Sensor Networks”, in Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC), Orlando, Florida, March 2002.

[77] Y. Xu, W. C. Lee, J. Xu, and G. Mitchell, “Processing Window Queries in Wireless Sensor Networks”, IEEE International Conference on Data Engineering (ICDE’06), Atlanta, GA, April 2006.

[78] R. Rosemark and W. C. Lee, “Decentralizing Query Processing in Sensor Networks”, the Second International Conference on Mobile and Ubiquitous Systems: Networking and Services (Mobiquitous’05), San Diego, CA, July, 2005, pp. 270-280.

[79] J. Winter, Y. Xu, and W. C. Lee, “Energy Efficient Processing of K Nearest Neighbor Queries in Locationaware Sensor Networks”, the Second International Conference on Mobile and Ubiquitous Systems: Networking and Services (Mobiquitous’05), San Diego,CA, July, 2005, pp. 281-292.

[80] G. Bontempi and Y. Le Borgne, “An adaptive modular approach to the mining of sensor network data”, Workshop on Data Mining in Sensor Networks, SIAM SDM, Newport Beach, CA, USA, April 2005.

[81] C. Liu, K. Wu, and J. Pei, “A Dynamic Clustering and Scheduling Approach to Energy Saving in Data Collection from Wireless Sensor Networks”, In Proceedings of the Second Annual IEEE Communications Society Conference on Sensor and AdHoc Communications and Networks (SECON’05), Santa Clara, California, USA, September, 2005.

[82] O. Younis and S, Fahmy, “An Experimental Study of Routing and Data Aggregation in Sensor Networks”, In Proceedings of the International Workshop on Localized Communication and Topology Protocols for Ad hoc Networks (LOCAN), held in conjunction with The 2nd IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS-2005), November 2005.

[83] R. Virrankoski and A. Savvides, “TASC : Topology Adaptive Spatial Clustering for Sensor Networks”, Second IEEE Intl. Conf. on Mobile Ad Hoc and Sensor systems”, Washington, DC, November, 2005.

[84] S. Soro and W. Heinzelman, “Prolonging the Lifetime of Wireless Sensor Networks via Unequal Clustering,” Proceedings of the 5th International Workshop on Algorithms for Wireless, Mobile, Ad Hoc and Sensor Networks (IEEE WMAN ’05), April 2005.

[85] M. Lotfinezhad and B. Liang, “Effect of partially correlated data on clustering in wireless sensor networks,” in Proceedings of the IEEE International Conference on Sensor and Ad hoc Communications and Networks (SECON), Santa Clara, California, October 2004.

[86] C. Guestrin, P. Bodik, R. Thibaux, M. Paskin and S. Madden, “Distributed Regression : An Efficient Framework for Modeling Sensor Network Data”, Intel corporation, 2004.

[87] K. Dasgupta, K. Kalpakis and P. Namjoshi, “An Efficient Clustering-based Heuristic for Data Gathering and Aggregation in Sensor Networks”, IEEE Wireless Communications and Networking Conference, Vol.4, No. 1, March, 2003.

[88] J. Beaver, M. A. Sharaf, A. Labrinidis, and P. K. Chrysanthis, “Location-Aware Routing for Data Aggregation in Sensor Networks”, Proc. of the 2nd Hellenic Data Management Symposium, 2003.

[89] P. Radivojac, U. Korad, K. M. Sivalingam and Z. Obradovic, “Learning from Class-Imbalanced Data in Wireless Sensor Networks,” IEEE Semiannual Vehicular Technology Conference, VTC-Fall 2003, Vol. 5, pp. 3030-3034, Orlando, Florida, U.S.A., October 2003.

[90] B. Karp and H. Kung. Greedy perimeter stateless routing for wireless networks. In Proceedings of the Sixth Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom 2000), pages 243–254, Boston, MA, August 2000.

[91] P. Beyens, M. Peeters, K. Steenhaut and A. Nowe, “Routing with Compression in Wireless Sensor Networks: a Q-learning Approach”, In “Fifth European Workshop on Adaptive Agents and Multi-Agent Systems (AAMAS 05), Paris, France.”, 2005.

[92] M. Esnaashari, M. R. Meybod, “A Learning Automata Based Data Aggregation Method for Sensor Networks”, CSICC 2007

[93] Nishimura C. E. and Conlon D. M., “IUSS Dual Use: Monitoring Whales and Earthquakes Using SOSUS”, in: Proceedings of the Mar. Tech., Soc. Journal, Vol. 27, No. 4, pp. 13-21, 1994.

[94] http://www.pmel.noaa.gov/vents/acoustics/sosus.html

[95] Al-Karaki J. N. and Kamal A. E., ”Routing techniques in wireless sensor networks: a survey”. in: Proceedings of the IEEE Wireless Communications, Vol. 11, pp. 6-28, 2004.

[96] Janakiram D., Venkateswarlu R. and Nitin S., ”A survey on programming languages, middleware and applications in wireless sensor networks”, in: Proceedings of the IITM-CSE-DOS-2005-04, 2005.

[97] Janakiram D., Venkateswarlu R. and Nitin S., ”A survey on programming languages, middleware and applications in wireless sensor networks”, in: Proceedings of the IITM-CSE-DOS-2005-04, 2005.

[98] Tsypkin Ya. Z.; “Adaptation and learning in automatic systems”; New York: Academic Press, 1971.

[99] Tsetlin, M. L., Automaton theory and modeling of biological systems, New York: Academic Press, 1973.

[100] Narendra K.S., Viswanathan R.; “Learning models using stochastic automata”; Proceedings of International Conference of Cybernetics and Society, Washington DC, October 1972.

[101] Varshavski V. I., Vorontsova I.P.; “On the behavior of stochastic automata with variable structure”; Avtomatika & Telemechanika, vol. 24, pp. 353-360, 1963.

[102] Fu K. S., McMurtry G. J.; “An application of stochastic automata to the synthesis learning systems”; Technical Report TR-EE 65-17,Purdue University, 1965.

[103] Fu K. S., Li T. J.; “Formulation of learning automata and automata games”; Information Science, vol. 1, no. 3, pp. 237-256, 1969.

[104] Fu K. S., Li T. J.; “On stochastic automata and languages”; Information Science, vol. 1, no. 4, pp. 403-419, 1969.

[105] Fu K. S.; “Stochastic automata as models of learning systems”; in Computer and Information Sciences II, Ed. Lou J.T., New York: Academic Press, 1967.

[106] McLaren R. W.; “A stochastic automaton model for synthesis of learning systems”; IEEE Transactions on System Science and Cybernetics, vol. 2, pp. 109-114, 1966.

[107] Chandrasekharan B., Shen D. W. C.; “On expediency and convergence in variable structure stochastic automata”; IEEE Transactions on System Science and Cybernetics, vol. 5, pp. 145-149, 1968.

[108] Najim K., Poznyak A. S.; editors, “Learning automata: theory and application”, Tarrytown, New York: Elsevier Science Publishing Ltd., 1994.

[109] Oommen B. J., Ma D. Y.”Deterministic learning automata solutions to equipartioning problem”; IEEE Transactions on Computers, vol. 37, pp. 2-14, January 1988.

[110] Oommen, B. J., Croix St., de E. V.; “String taxonomy using learning automata”; IEEE Transactions on Systems, Man and Cybernetics, vol. 27, pp. 354-365, April 1997.