MODIFICACIONES SINTÁCTICAS EN LENGUA ESPAÑOLA CON UTILIDAD EN ESTEGANOGRAFÍA LINGÜÍSTICA

ALFONSO MUÑOZ, IRINA ARGÜELLES, JUSTO CARRACEDO
UNIVERSIDAD POLITÉCNICA DE MADRID

Resumen: En la actualidad la publicación de estudios sobre esteganografía lingüística en lengua española es muy escasa, a pesar de los avances significativos en el procesamiento de lenguaje natural en esta lengua que han sido notorios en su aplicación a la traducción automática, data mining, resumen automático de textos, etc. En general, la esteganografía lingüística permite ocultar información en un texto en lenguaje natural, lo cual tiene utilidad, principalmente, en la protección-anonimato de comunicaciones digitales y en el marcado digital de textos (*Natural Language Watermarking*). El presente artículo, circunscrito dentro de una línea de investigación más amplia iniciada por los autores, presenta la posibilidad de utilizar la sintaxis en lengua española para ocultar información, hecho que podría tener utilidad, entre otras, para burlar censuras.

Palabras clave: *esteganografía lingüística, modificaciones sintácticas, marcado digital de textos, estegotexto*

Abstract: At present, the publication of studies on linguistic steganography in Spanish language is very scarce, despite the significant advances in the processing of natural language in Spanish that have lead to significant applications in automatic translation, data mining, automatic text summarizing, etc. Linguistic steganography can hide information in natural language texts, which is useful mainly in the protection of digital communications and digital watermarking of texts. This article, circumscribed within a broader line of research initiated by the authors, presents the possibility of using different syntactical structures of Spanish language to hide information with the aim, to circumvent censorship and facilitating freedom of expression within dictatorial regimes.

Keywords: *linguistic steganography, syntactical modifications, natural digital watermarking, stegotext*

1. El lenguaje natural en la ocultación de información: Esteganografía

La última década es testigo de la utilidad de la lingüística computacional en tecnologías tan dispares como son los sistemas de traducción automática, los algoritmos de reconocimiento del habla, algoritmos de análisis ortográficos, los sistemas de *data mining*, resumen automático de textos, etc. En general, algoritmos que aprovechan todo el conocimiento disponible en áreas tan variadas como puede ser el análisis del discurso, la lingüística de corpus, la lexicografía, la estadística de palabras, etc. Además de las anteriores, en los últimos 5 años se ha profundizado, especialmente en su aplicación a lengua inglesa, en una nueva vertiente de utilización de todo el conocimiento lingüístico disponible sobre una lengua concreta. Estamos hablando de la posibilidad de ocultar información en lenguaje natural, es decir, de la ciencia de la esteganografía lingüística. La ciencia de la esteganografía puede definirse como la ciencia y el arte de ocultar una información dentro de otra, que haría la función de *tapadera o cubierta*, con la intención de que no se perciba ni siquiera la existencia

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 229-247
Recibido: 30/12/2009
Aceptación comunicada:14/01/2010

de dicha información (Carracedo 2004:123-131). Cuando la cubierta o tapadera es un texto en lenguaje natural se habla de un tipo concreto de esteganografía, la esteganografía lingüística (en nuestro caso esteganografía y lingüística computacional). La ciencia de la esteganografía es complementaria a la ciencia de la criptografía, esta última si bien no oculta la existencia de un mensaje sí lo hace ilegible para quien no esté al tanto de un determinado secreto, una clave. En la práctica ambas ciencias pueden combinarse para mejorar la autenticidad y privacidad de las comunicaciones.

La ocultación de mensajes usando procedimientos esteganográficos puede tener fines legítimos o ilegítimos, que pueden ser beneficiosos para la proteger la privacidad de las comunicaciones o burlar censuras, o, por el contrario, ser vehículos para perpetrar actos criminales. Por este motivo, diversas entidades invierten cantidades importantes de dinero en la detección de este tipo de comunicaciones, es la ciencia del estegoanálisis, especialmente desde que el periódico sensacionalista Usa Today publicara en 2001 que la red terrorista Al-Qaeda y Ben-Laden utilizó estos procedimientos para intercambiar información en Internet sin ser detectados. Por este motivo, en el pasado analizamos e implementamos algunos procedimientos para detectar este tipo de comunicaciones ocultas en contenidos multimedia. Véase, por ejemplo, la herramienta StegSecret (Muñoz y Carracedo 2007).

En la actualidad, el potencial de las redes sociales y los medios colaborativos para la ocultación de comunicaciones (Muñoz, Carracedo y Sánchez 2008) hizo ampliar nuestro enfoque de análisis a la posibilidad de intercambio de mensajes ocultos en lenguaje natural de manera no trivial. Por desgracia, los precedentes de esteganografía lingüística aplicada a lengua española son muy escasos (Calvo y Bolshakov 2004) por lo que ha sido necesario iniciar diversas investigaciones en este área (Muñoz y Carracedo 2009a y 2009c). En general, la utilidad de estas nuevas investigaciones puede resumirse en dos grandes temáticas:

a) Anonimato y privacidad. La posibilidad de ocultar información en textos en lenguaje natural permitiría intercambiar información dificultando su detección a personas y sistemas de monitorización automáticos (Echelon, Carnivore, Sitel…), mejorando, o eso se espera, la privacidad e incluso el anonimato de dichas comunicaciones. Sería de interés en términos de libertad de expresión.

b) Marcado digital de textos. En la actualidad, la integridad y la autenticidad de una información pueden ser garantizadas mediante la utilización de la denominada firma digital. En general, procedimientos que generan unos datos extras basados en la información que se quiere proteger y que se adjuntan a esa misma información. Estos procedimientos tienen el inconveniente que la información generada no está autocontenida en la información que se quiere proteger, luego pueden ser separadas con los problemas en términos de verificación que esto puede suponer. La posibilidad de ocultar información en un texto en lenguaje natural, si esta modificación no supusiera “alteraciones notorias” del texto utilizado como portador, facilitaría la inclusión autocontenida de firmas que podrían tener utilidad en autenticidad e integridad de escritos en lengua española. Una firma autocontenida permite garantizar que la firma de un autor presente en un artículo corresponde precisamente al texto que escribió y se puede, además, demostrar que él es el autor de dicho documento (*authorship proof*) así como “realizar un seguimiento” del mismo, por ejemplo, para medir la difusión de una obra.

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 229-247
Recibido: 30/12/2009
Aceptación comunicada:14/01/2010

2. Esteganografía Lingüística

La esteganografía lingüística puede definirse como aquel conjunto de algoritmos robustos que permiten ocultar una información, típicamente binaria, utilizando como tapadera información en lenguaje natural. A lo largo de los siglos se han documentado múltiples formas de ocultación de información utilizando tapaderas textuales (estegotextos) de lo más variadas (libros, telegramas, poesías, canciones, revistas, periódicos, etc.) con procedimientos variopintos. Por ejemplo, el sistema *newspaper code* en la época Victoriana o la verja de Cardano en el siglo XVI (Kahn 1996). En general, la mayoría de estos procedimientos son clasificables en términos de *open codes*[1] *y semagrams*[2]. En la actualidad, la esteganografía lingüística intenta mezclar principios de la ciencia de la esteganografía y la lingüística computacional (análisis automático del contenido textual, generación textual, análisis morfosintáctico, lexicografía computacional, descripciones ontológicas, etc.) para crear procedimientos no triviales aplicando el principio de Kerckhoffs (Kahn 1996). La aplicación de este principio a esteganografía indica que los algoritmos que faciliten la ocultación de la información serán públicos y la seguridad de todo el esquema dependerá exclusivamente de una información adicional conocida exclusivamente por emisor y receptor a modo de clave. Como puede suponerse, este tema no es nada sencillo, sobre todo si se considera que el estegotexto resultante debe resistir ataques estadísticos y lingüísticos por parte de analistas (humanos) y sistemas automáticos (máquinas). En este sentido, existen dos líneas principales de investigación:

a) Modificación de textos existentes para crear estegotextos. El mecanismo más tradicional de ocultación de información consiste en utilizar un texto existente para enmascarar información basándose en algunos elementos del texto o modificaciones del mismo. Procedimientos actuales, no exentos de problemas, consisten en: modificaciones léxicas -el mecanismo más utilizado es la ocultación de información basada en la sustitución de palabras por sus sinónimos-, modificaciones sintácticas y semánticas, ocultación mediante el "ruido" de las traducciones de un texto entre diferentes idiomas, ocultación basada en errores tipográficos y ortográficos, ocultación basada en símbolos de puntuación-abreviaturas y ocultación basada en modificaciones de la estructura o formato de un texto, por ejemplo, el uso de espacios de tamaño variable entre palabras, variación del estilo de fuente, etc (Bergmair 2007).

b) Generación automática de estegotextos. La modificación de textos existentes tiene el problema, al margen de otras cuestiones lingüísticas y estadísticas, de que un analista fuera capaz de conseguir el texto original al que se han aplicado las modificaciones, lo que le permitiría aplicar ataques basados en la comparación del texto original con el texto modificado. Para evitar estos ataques se han publicado diferentes propuestas para la

[1] Los *open codes* genéricamente se refieren a textos de apariencia inocente, que ocultan información recuperable utilizando ciertas letras, palabras, frases del texto o comunicación. Métodos basados en esto son: Cues, Null Ciphers, Jargon Code, Grilles, etc.

[2] Tipo de técnicas que consisten en la utilización (variación) de la estructura y formato de los elementos de un texto, aunque visibles, no por ello son fáciles de detectar. Por ejemplo, la codificación de una información con espacios de tamaño variable.

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 229-247
Recibido: 30/12/2009
Aceptación comunicada:14/01/2010

generación automática de textos “a medida” con fines esteganográficos, es decir, procedimientos para generar textos que lleven ya incluida la información que se quiera ocultar (estegotextos). Esta idea, que suena apasionante, en la práctica resulta de una enorme complejidad, ya que si bien es viable generar textos con validez léxica y sintáctica, la semántica y la coherencia global son a día de hoy temas sin una solución clara. Existen dos líneas generales de generación automática de estegotextos, en ocasiones combinadas, basadas en imitación gramatical y basadas en imitación estadística. A continuación se citan algunos ejemplos destacables.

En la década de los 90 Peter Wayner (1995) adelantó algún trabajo a este respecto en lengua inglesa relacionando los avances del lingüista A. Noam Chomsky y su gramática generativa con la esteganografía lingüística mediante el uso de CFGs (*Context-Free-Grammar*). Pocos avances cualitativos en generación automática se han documentado desde entonces. Su idea, más tarde implementada en herramientas automáticas como Nicetext (Chapman 1997), consiste en la posibilidad de imitar la gramática de uno o varios textos de entrenamiento generando reglas gramaticales que luego se utilizarán para generar las frases del estegotexto que se desea crear. La ocultación de la información se realiza mediante la selección de las palabras concretas que se utilizarán en cada frase. Entre los múltiples problemas documentados de esta propuesta destaca la dificultad de generar estegotextos con coherencia global y la aparición de patrones estadísticos que delatarían la presencia de información oculta, por ejemplo, patrones basados en la distribución no equitativa de palabras en las partes de un texto (Zhi-li et al. 2008).

Otra alternativa en la generación automática de estegotextos consiste en la utilización de algoritmos que imiten la estadística de textos de entrenamiento. Un procedimiento sencillo para conseguir esto es el siguiente: dado uno o más textos de entrenamiento se anotan las colocaciones de las palabras presentes en los textos y las frecuencias de repetición de cada una (modelo N-Gram), es decir, anotamos qué palabra va detrás de qué otra y con qué probabilidad. Lo único que debe hacer el algoritmo de generación de estegotextos es, después de seleccionar una palabra, elegir de una manera aleatoria la siguiente de entre las posibles que indican las colocaciones (Wayner 1992). Si la selección es aleatoria es más probable que salgan las palabras más probables que las que son menos, y por tanto se imite la estadística de la fuente de entrenamiento. Si se imita la estadística de la fuente es más probable que el estegotexto resultante que está basado en los textos de entrenamiento, tenga validez léxica y sintáctica (validez léxica y sintáctica que deben tener los textos de entrenamiento). Esta idea se comprobó en lengua española mediante la implementación de la herramienta Stelin (Muñoz y Carracedo 2009a). No obstante, el mayor problema de la imitación estadística reside en la generación de estegotextos con coherencia global, a día de hoy un tema realmente complejo. A pesar de este problema en (Muñoz 2009d) indicamos como puede mejorarse este aspecto mediante anotación manual. Este proceso consiste en la posibilidad de añadir manualmente palabras entre cada dos palabras del estegotexto generado con la intención que el emisor pueda mejorar manualmente la apariencia global del estegotexto, minimizando los errores producidos en la generación automática. Las palabras añadidas tienen que tener unas condiciones para que el receptor pueda descartarlas sin problemas para recuperar la información oculta. Esta idea permite crear pequeños estegotextos de gran calidad, eso sí, entre sus limitaciones existe la necesidad de edición manual a posteriori del estegotexto generado automáticamente. Esta edición hace que para una relación de esfuerzo-tiempo razonable (Muñoz 2009d) sólo sea viable la ocultación de una pequeña centena de caracteres,

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 229-247
Recibido: 30/12/2009
Aceptación comunicada:14/01/2010

en torno a 1000 bits, lo cual puede ser interesante para el intercambio de mensajes cortos de información.

3. Modificaciones sintácticas con fines esteganográficos

En los apartados anteriores se ha descrito el ámbito de utilidad de la esteganografía lingüística y las dos grandes líneas de investigación para crear algoritmos robustos. Debido a los numerosos problemas de la generación automática de estegotextos en la presente década los esfuerzos van destinados a la modificación de textos existentes, principalmente mediante modificaciones léxicas, sintácticas y semánticas, intentando identificar y cuantificar el impacto de estas modificaciones (Bergmair 2007).

En este trabajo nuestro interés se centra exclusivamente en analizar la posibilidad de utilizar modificaciones sintácticas en frases en lengua española con fines esteganográficos, modificaciones que se han mostrado útiles en otras lenguas para marcado digital de textos o *Natural Language Watermarking* (NLW), en terminología inglesa. En teoría, la manipulación sintáctica de una frase con fines esteganográficos se basa en el hecho que las frases son combinaciones de sintaxis y semántica, y 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 frase o 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 de una u otra opción es lo que permitirá ocultar información. Este estudio se centra en analizar la conveniencia o no de utilizar transformaciones ya analizadas con éxito en otros idiomas e indagar sobre otras nuevas. Para ello se entiende por modificación sintáctica aquella modificación que no consiste exclusivamente en modificar, añadir o suprimir algún término de una frase. Estas modificaciones para nosotros son modificaciones léxicas a diferencia de otros autores (Murphy 2007).

Algunas de las transformaciones sintácticas (Bergmair 2007) más interesantes publicadas para la lengua inglesa y que nos sirven de referencia, al ser la lengua con más estudios publicados, son:

a) Transformación Activa-Pasiva. Esta transformación consiste en la posibilidad de convertir una frase de voz activa a voz pasiva y viceversa. Un ejemplo de transformación en inglés sería: *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.

b) Movimiento de complementos, frases preposicionales y frases adverbiales. La idea aparentemente sencilla detrás de estas transformaciones reside en la posibilidad de mover palabras dentro de una frase sin que esto afecte al significado de la misma. Un ejemplo de esta transformación en inglés sería: a) *Often the dog chased the cat, b) the dog often chased the cat y c) the dog chased the cat often.* En esta estructura un algoritmo de ocultación debería poder asignar códigos binarios diferentes a cada variación de una frase concreta. Al igual que la estructura anterior esto permitiría la ocultación de información. En la frase

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 229-247
Recibido: 30/12/2009
Aceptación comunicada:14/01/2010

anterior existen 3 formas de ocultar información luego sería posible ocultar $\log_2 3 = 1{,}58$ bits/frase. Se han documentado otras estructuras considerando complementos de tiempo, lugar, modo y contingencia (según Murphy 2001) algunos ejemplos son los siguientes: a) *[In the morning]$_{TIME}$ I went to work* o *I went to work [In the morning]$_{TIME,}$* b) *Life was better [in the home country]$_{PLACE}$* o *[In the home country]$_{PLACE}$ life was better* y c) *I'll call him [if I can find my phone]$_{CONTINGENCY}$* o *[If I can find my phone]$_{CONTINGENCY}$ I'll call him.*

c) Cambio de orden de los términos unidos por una conjunción. Esta transformación se basa en la posibilidad de intercambiar las palabras vinculadas por una conjunción, especialmente y/o. En el ejemplo más sencillo palabras unidas por la conjunción "y". Un ejemplo de transformación ideal en inglés sería: a) *Ali and Ayse* y *b) Ayse and Ali.* En este ejemplo simple un algoritmo de ocultación debería establecer un criterio para conocer qué orden de palabras codificaría un bit 0 o un bit 1, al existir sólo dos posibilidades de la frase.

d) Movimiento de partículas de Phrasal Verbs. En lengua inglesa muchos verbos pueden llevar asociados preposiciones y adverbios. Algunos verbos permiten movimiento de la preposición o el adverbio después del objeto directo. En este caso un ejemplo de transformación que ocultaría un bit sería: *a) Alfonso took the mafia on* (por ejemplo, ocultaría un bit 0) *y b) Alfonso took on the mafia* (ocultaría un bit 1).

e) Inserción artículo It (extraposición). La extraposición es un fenómeno mediante el cual un sujeto puede sustituirse por *it* y moverse al final de una oración convirtiéndose en un complemento (Murphy 2001). Ejemplos de esta transformación sería: a) *[That he leaves the top off the toothpaste]$_{subject}$ annoys me intensely* o *It annoys me intensely [that he leaves the top off the toothpaste]* y b) *To believe that is difficult o It is difficult to believe that.*

Como se ha podido observar en las transformaciones sintácticas anteriores, existen diversas formas de alcanzar el objetivo de ocultar información. La mayoría de los procedimientos más interesantes se basan en el movimiento de palabras y en general permiten ocultar muy poca información por cada frase de un texto, en la práctica menos de un bit en media por frase. Esto fuerza a que el texto portador que se quiere modificar tenga un tamaño mínimo determinado para poder añadir algún tipo de información con utilidad en marcado digital de textos, por ejemplo, una decena de bits. Independientemente de la validez o no de los resultados publicados en lengua inglesa y aprovechando la idea general expuesta en estos trabajos previos, a continuación, se analizan algunas transformaciones sintácticas en su aplicación esteganográfica a la lengua española.

4. Modificaciones sintácticas en lengua española con fines esteganográficos

4.1. Metodología de análisis

Las transformaciones sintácticas que se van a estudiar con fines esteganográficos son tres: a) transformación basada en la conversión de una frase activa-pasiva, b) transformación basada en movimientos de los adjetivos y otros complementos dentro del sintagma nominal y c) transformación basada en la reordenación de complementos del verbo. Para cada una de estas estructuras se presenta una hipótesis lingüística de partida y su comprobación mediante

mediciones reales en corpora. Carreter (1980) entiende que "En todo conflicto entre formulación gramatical y sentimiento espontáneo de la lengua, es éste el que, por principio, merece mayor confianza" luego se asume que aunque las teorías gramaticales pueden y deben servir para predecir posibles problemas a la hora de manipular estructuras sintácticas con fines esteganográficos, es en definitiva la intuición del hablante la que decidirá sobre qué cambios o alteraciones son de hecho posibles y cuáles no, por esto resulta de interés cuantificar la utilidad de esa transformación en esteganografía, así como la aceptabilidad por lectores humanos.

Junto con las transformaciones que se han propuesto, se adjuntan una serie de medidas para distintas estructuras que se piensan pueden tener utilidad esteganográfica. Estas medidas permitirían, en el caso que dichas transformaciones fueran realmente útiles, calcular la cantidad máxima de información por frase que se podría ocultar usando una estructura concreta. La aceptabilidad de una frase modificada con una transformación determinada y la cantidad de información que permita ocultar dicha frase permitirá concluir la productividad de esas transformaciones con fines esteganográficos. Respecto a este último punto es importante recordar que el tamaño de la firma digital "tradicional" en documentos electrónicos suele rondar la centena de bits de información (128 bits o 256 bits típicamente). Si se desea marcar digitalmente textos con procedimientos esteganográficos sería necesario ocultar al menos unas cuantas decenas de bits, esto debe considerarse para determinar la cantidad de texto original que sería necesario para incluir dicha firma. Esta limitación de longitud hará que en la práctica ciertas estructuras no sean productivas esteganográficamente.

4.1.1. Descripción de los corpora bajo estudio

En este trabajo se utilizan dos corpora reales para cuantificar la utilidad de estas transformaciones para esteganografía lingüística en español.

a) Corpus LEXESP. Existen numerosos corpus en lengua española que por sus características serían de interés para realizar diferentes consultas lingüísticas, como puede ser el corpus CREA (corpus de referencia del español actual de la Real Academia de la lengua Española), el corpus CUMBRE (Almela et al. 2005) u otros. En este estudio, por su disponibilidad, se trabaja con el corpus LEXESP. Un corpus lematizado de 5.020.930 palabras (Sebástian et al. 2000) cuya composición mezcla estilos variados como la narrativa (40%), la divulgación científica (10%), el ensayo (10%), la prensa (25%), semanarios (10%) y prensa deportiva (5%). La homogeneidad de este corpus que prescinde de las variaciones dialectales del castellano, salvo aproximadamente un 10% de textos de autores hispanoamericanos, y la heterogeneidad del corpus, donde la selección de los textos pretende ser heterogénea y representativa, hace que se utilice como corpus de referencia en este trabajo para analizar la utilidad de las transformaciones esteganográficas bajo estudio.

b) Corpus Twitter. El uso del lenguaje podría variar en función del canal de comunicación donde se deseara transmitir una información. Es posible que una técnica esteganográfica no fuera permisible en un medio de comunicación tradicional, como un artículo en un periódico, y sin embargo sí lo fuera en un canal más informal como pudiera ser una red social en Internet. En este sentido se decide crear un corpus de una red concreta, la red de microblogging Twitter, y cuantificar la presencia de las transformaciones bajo estudio en este corpus. En general, la red Twitter permite enviar mensajes (frases) basados en texto,

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 229-247
Recibido: 30/12/2009
Aceptación comunicada:14/01/2010

denominados “tweets”, de una longitud máxima de 140 caracteres, que se publican en la página de un usuario de la red twitter. Cada usuario twitter puede tener un número determinado de “followers” (seguidores) y “following” (personas a las que sigue) que definen en cierta forma su popularidad en la red. Considerándose estos parámetros se construye un corpus seleccionando 103 usuarios de España. Los criterios de selección de estos usuarios se establecen según un número mínimo de seguidores (1000) y un número tweets de 3200, que es el máximo al que se puede acceder por usuario (en algunos usuarios este número es inferior), de forma que el lenguaje que se va a analizar no sea una jerga propia de un conjunto reducido de usuarios y sí la expresión de mensajes que desean ser leídos por una mayoría pero que pueden tener las características o “comodidades” propias del canal donde se emiten. El corpus Twitter creado está constituido por 319.381 mensajes Twitter y un total de 4.029.257 palabras. Para el etiquetado de este corpus se implementa un programa en lenguaje JAVA que utiliza el etiquetador TreeTagger (Schmid 2009) considerando un mensaje Twitter como una frase. Los resultados en las mediciones en este corpus deben considerar las limitaciones o imprecisiones de este etiquetador, además aunque las medidas en este corpus no son comparables con las del corpus LEXESP, sin embargo sí sirven como una primera aproximación de la presencia de estas estructuras en redes sociales y para una posible cuantificación inicial de la capacidad máxima de ocultación por estructura. Los razonamientos lingüísticos se realizarán exclusivamente con el corpus LEXESP para evitar, de momento, posibles errores derivados del etiquetador. Este corpus puede descargarse de (Muñoz 2009b).

4.2. Transformación sintáctica basada en el cambio activa-pasiva

4.2.1. Hipótesis lingüística

Aunque en lengua inglesa esta transformación ha resultado ser productiva con fines esteganográficos, partimos de la hipótesis de que no lo será tanto en el caso del español donde la oración pasiva tiene una fuerte componente semántica. Si bien no cabe duda de que aparecerán muchos ejemplos en los que la transformación implica una variación semántica mínima (1_1, 1_2), es cierto que es relativamente fácil dar con otros en los que la transformación es dudosa, por ejemplo por la estructura argumental y temática del verbo (2_1, 2_2, 2_3), o por las características del sintagma que recibe el argumento sujeto en la construcción activa y que condicionan el papel semántico (o temático) que recibe el mismo SN en la construcción pasiva (3_1, 3_2, 4_1, 4_2).

(1_1) El atracador golpeó a Luis.
(1_2) Luis fue golpeado por el atracador.
(2_1) El atracador pegó un puñetazo a Luis.
(2_2) ? Un puñetazo fue pegado por el atracador a Luis.
(2_3) *Luis fue pegado un puñetazo por el atracador.
(3_1) El capitán tranquilizó al marinero.
(3_2) ?El marinero fue tranquilizado por el capitán [agente].
(4_1) El cambio en el tiempo tranquilizó al marinero.
(4_2) * El marinero fue tranquilizado por el cambio en el tiempo [causa].

Otros ejemplos producen una variación semántico-pragmática (5_1,5_2) más notoria.

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 229-247
Recibido: 30/12/2009
Aceptación comunicada:14/01/2010

(5_1) María$_i$ acogió a su$_i$ amiga.
(5_2) Su$_i$ amiga fue acogida por María$_j$. (¿La amiga de María?)

Además de los posibles problemas con esta construcción debe considerarse otras más frecuentes como las construcciones de verbo con se (6,7):

(6) La puerta se abrió (forma media).
(7) Se admiten animales (pasiva refleja)[3].

En general, el uso de la pasiva (Lázaro 1980) en lengua española no es común. Además, no queda claro si la alternancia de frases en voz activa y voz pasiva en un mismo texto o fragmento medio delataría la presencia de información oculta. Salvo pocas modificaciones, unos pocos bits por texto, parece que el uso automático y masivo de esta transformación crearía estegotextos que no pasarían inadvertidos.

4.2.2. Experimentación de la transformación

A continuación se va a cuantificar la presencia de tres estructuras en los corpora bajo estudio: a) ser+participio, b) ser+participio+por, c) se + verbo en 3ºpersona para medir la presencia de oraciones en pasiva y llegar a conclusiones sobre su utilidad esteganográfica.

Transformación Activa/Pasiva	**Corpus LEXESP**	
	Nº frases	Ocurrencias (%) (total 214.400 frases)
SER+PARTICIPIO	**6712**	**3,1305%**
SER+PARTICIPIO+POR	**1229**	**0,5732%**
SE+VERBO	**40006**	**18,6595%**

Tabla 1: Cuantificación de estructuras típicas en oraciones en pasiva para corpus LEXESP.

Transformación Activa/Pasiva	**Corpus TWITTER**	
	Nº frases	Ocurrencias (%) (319.483 frases)
SER+PARTICIPIO	**1207**	**0,3777%**
SER+PARTICIPIO+POR	**37**	**0,0115%**
SE+VERBO	**17864**	**5,5915%**

Tabla 2: Cuantificación de estructuras típicas en oraciones en pasiva para corpus TWITTER.

Las medidas reflejadas en la Tabla1 indican (recordamos que de momento nuestras conclusiones serán a partir del corpus LEXESP) que las transformaciones ser+participio y ser+participio+por tienen una ocurrencia baja. En el apartado 3 se justificó cómo la transformación activa-pasiva permite ocultar 1 bit de información por frase modificada. Según este criterio, si el texto que queremos modificar tuviera las propiedades estadísticas del corpus LEXESP esto significaría que se podrían ocultar 3,13 bits/100 frases

3 Siguiendo la distinción que propone Cano Aguilar (1987: 276, 278, 288). No se tratará aquí sobre consideraciones relativas a la clasificación y denominación de éstos y otros tipos de construcciones con se.

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 229-247
Recibido: 30/12/2009
Aceptación comunicada:14/01/2010

(0,0313bits/frase) y 0,57 bits/100 frases (0,0057 bits/frase) para el caso de la estructura ser+participio y la estructura ser+participio+por respectivamente. Si se deseara ocultar, por ejemplo, 40 bits de una firma se necesitarían 1277,95 frases y 7017,54 frases respectivamente. Esto a priori, ya haría que muchos textos no cumplieran esa condición inicial de longitud media. Además, existen otros aspectos que deben considerarse como son la presencia de patrones estadísticos que podrían delatar la existencia de información oculta. Por ejemplo, en el corpus LEXESP la aparición de estas estructuras sigue una media de 174,36 frases por cada frase con ser+participio+por y 31,9417 frases por cada frase con ser+participio; estas consideraciones deberían tenerse en cuenta junto al criterio anterior. En términos de aceptabilidad, los ejemplos extraídos del corpus no permiten afirmar que la automatización de esta transformación sea sencilla sin perjuicios (1_1,1_2,2_1,2_2). De las medidas obtenidas puede concluirse que la estructura se+participio y ser+participio+por no son productivas en términos de esteganografía lingüística en lengua española.

(1_1) El libro queda en su extraña perfección y en el exquisito cuidado con_que *fue alumbrado.*

(1_2) [Transformación Manual] El libro queda en su extraña perfección y en el exquisito cuidado con_que *[el autor]*$_{agente}$ *lo alumbró.*

(2_1) Sin mayores comprobaciones, *Bubber es catapultado por Gale* hacia la cima y se embolsa el millón.

(2_2) [Transformación Manual] Sin mayores comprobaciones, *Gale catapulta a Bubber* hacia la cima y se embolsa el millón.

Respecto a la construcción se+verbo aunque su presencia es más alta, se necesita un estudio más profundo, ya que la medida reflejada en Tabla1 y 2 no tiene en cuenta que esta estructura se repite en verbos reflexivos, recíprocos, etc. No obstante, suponiendo el 18,6595% como una cota superior de capacidad de ocultación estaríamos hablando de 0,1865 bits/frase. Este valor, sin considerar de momento condiciones de aceptabilidad, limitaría en la práctica su utilidad esteganográfica en cualquier texto, al menos de forma única.

4.3. Transformación basada en movimientos de los adjetivos y otros complementos dentro del sintagma nominal

4.3.1. Hipótesis lingüística

Cabe preguntarse si en relación con los adjetivos en una oración, existe en español algún orden determinado. Al margen del tema ya muy estudiado del efecto de la anteposición o posposición de los adjetivos calificativos existen otras cuestiones que impiden dar por hecho que los adjetivos que modifican el nombre son intercambiables. Según Jackendoff (1977) sintácticamente no se distingue una jerarquía entre distintos modificadores de un núcleo lo que nos llevaría a deducir que sí pueden serlo, y que el intercambio de adjetivos podría tener utilidad esteganográfica. Pero otros investigadores defienden que unos modificadores están más ligados al nombre que otros y, por lo tanto, deben aparecer más cerca (1_1, 1_2):

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 229-247
Recibido: 30/12/2009
Aceptación comunicada:14/01/2010

(1_1) El cuadro rojo roto.
(1_2) ?El cuadro roto rojo.

Además de los adjetivos, existen otras construcciones de modificación del nombre frecuentes en español que podrían considerarse para la ocultación sintáctica pero que, atendiendo a los estudios de aceptabilidad de algunos investigadores (Hornstein y Lightfoot, 1981) también podrían presentar algunos problemas desde el punto de vista de la esteganografía. Mientras que las estructuras 1_1, 1_2 se consideran aceptables, otras que podrían en un primer momento parecer estructuralmente paralelas a las anteriores tienen un efecto de inaceptabilidad (2_1, 2_2).

(1_1) El estudiante de Valladolid con gafas.
(1_2) El estudiante con gafas de Valladolid.
(2_1) El estudiante de periodismo de Valladolid.
(2_2) *El estudiante de Valladolid de periodismo.

Al margen de las fundadas razones por las que esto es así (Hornstein y Lightfoot, 1981), todos los ejemplos anteriores brindan elementos suficientes de juicio para no asumir que estructuras sintácticamente paralelas sufren idénticas restricciones con respecto a su orden dentro de la oración. Mencionar por último que las construcciones en las que el núcleo del SN es un nombre deverbal presentan tantas posibilidades que no cabe aquí ocuparse de ellas. Sólo por dejar apuntadas otras posibles circunstancias contextuales, véase la propuesta de Fernández Soriano (1993) sobre la influencia de la longitud de los constituyentes en la ordenación de los argumentos del nombre y su aceptabilidad:

(3_1) La crítica a los planes de estudios de Juan.
(3_2) La crítica de Juan a los planes de estudios.
(4_1) *La crítica a los planes de estudios elaborados por la comisión correspondiente del departamento de Juan.
(4_2) La crítica de Juan a los planes de estudios elaborados por la comisión correspondiente del departamento.

Surge de nuevo y quizás es aquí mayor la duda sobre en qué grado los hablantes de una lengua producen o, en su caso, reconocen como aceptables estos tipos de transformaciones.

4.3.2. Experimentación de la transformación

A continuación se va a cuantificar las siguientes estructuras en los corpora bajo estudio: a) sust+adj=adj+sust, b) preposición+[art|det]+nombre+preposición+nombre, c) nom+adj1+adj2. Estas estructuras se han seleccionado con la intención de determinar la libertad de movimiento de los adjetivos respecto del nombre al que califican, así como del movimiento de palabras en frases preposicionales. El número de bits que se podría ocultar dependerá de las formas diferentes en las que puedan escribirse la misma frase, en general, $\log_2$(Nºformas diferentes). Por ejemplo, en el caso de poder situar un adjetivo sin perjuicios delante o detrás de un sustantivo hablaríamos de $\log_2 2=1$ bit/pareja sust-adj.

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 229-247
Recibido: 30/12/2009
Aceptación comunicada:14/01/2010

dentro del sintagma nominal		
	Nº frases	Ocurrencias (%) (total 214.400 frases)
SUST+ADJ|ADJ+SUST	**92301|59673**	**43,0508% | 27,8325%**
PREP+NOM+PREP+NOM	**16491**	**7,6916%**
NOM+ADJ+ADJ	**2031**	**0,9472%**
PREP+ART|DET+NOM+PREP+NOM	**26210**	**12,2248%**
	Nº parejas únicas	
SUST+ADJ=ADJ+SUST	**3520**	

Tabla 3: Cuantificación de estructuras útiles en movimientos de complementos del sintagma nominal. CORPUS LEXESP.

Transformación basada en movimientos de complementos dentro del sintagma nominal	**Corpus TWITTER**	
	Nº frases	Ocurrencias (%) (total 319.483 frases)
SUST+ADJ|ADJ+SUST	**75793|43884**	**23,7236% | 13,7359%**
PREP+NOM+PREP+NOM	**14514**	**4,5429%**
NOM+ADJ+ADJ	**6738 (*)**	**2,1090% (*)**
PREP+ART|DET+NOM+PREP+NOM	**11082**	**3,4687%**
	Nº parejas únicas	
SUST+ADJ=ADJ+SUST	**1662**	

*Tabla 4: Cuantificación de estructuras útiles en movimientos de complementos del sintagma nominal. CORPUS TWITTER. Las medidas con * hacen cuestionable la fiabilidad del etiquetador.*

a) Estructura Sust+adj=adj+sust. En la Tabla1 puede observarse que la presencia del adjetivo es más común después del sustantivo que antes. Cabe preguntarse si las parejas de palabras existentes en el corpus que cumplan la condición sust+adj=adj+sust permitiría más libertad de movimiento anteponiendo o posponiendo el adjetivo al sustantivo en la frase donde aparezca esta pareja. Las mediciones en LEXESP indican que las colocaciones (frecuencia de aparición de una palabra respecto a la otra) condicionan totalmente la flexibilidad de un adjetivo respecto de un sustantivo. Pueden verse la "libertad de intercambio" en los siguientes ejemplos del corpus (1,2):

(1) buena parte (179 veces) y parte buena (1 vez). Total -180 apariciones.

(1_1) Él tenía una *parte buena*, mansa, y otra muy violenta, ciega.
(1_2) En el último festival de cinema mediterráneo celebrado en Montpellier *buena parte* de las películas presentadas a concurso por los países mediterráneos del Norte, el Sur y el Este […]
(1_3) *Buena parte* de nuestros primeros trabajos consistieron en ese tipo de filmaciones.

(2) semana pasada (79 veces) y pasada semana (20 veces). Total – 99 apariciones.

(2_1) Hoy está un punto más cerca_del Deportivo que la *semana pasada.*
(2_2) La *semana pasada* en Orlando obtuvo un buen resultado.
(2_3) Lo único que se sabe del Madrid, tras los entrenamientos secretos de la *pasada semana* y su aparición pública ante el Málaga […]

En función de que las frecuencias de las parejas adj+sust y sust+adj para un mismo adjetivo y sustantivo se igualan, la apariencia de flexibilidad es mayor en el corpus LEXESP. De las 3520 parejas encontradas (sust+adj=adj+sust) podemos fijarnos, por ejemplo, en alguna de las 1537 parejas que sólo se encuentran una vez en el corpus. En estos ejemplos es más sencillo ver la flexibilidad de posponer o anteponer el adjetivo que califica al sustantivo (3,4).

(3) Pareja: rectángulo pequeño – pequeño rectángulo

(3_1) Allí enfrente, ese *rectángulo pequeño* que lleva en el centro una d.
(3_2) La habitación estaba completamente a_oscuras, pero una luz helada estallaba en los cristales enmarcando con timidez extraña el *pequeño rectángulo* de la ventana.

(4) Pareja: campaña feroz – feroz campaña

(4_1) La muerte de los tres guardias civiles desató una *campaña feroz* contra la acción de Alajuela.
(4_2) No se olvide que cuando Merino entró en "Four_Roses", los dominicanos estaban siendo objeto de una *feroz campaña* en_contra de su presencia.

Este hecho podría hacer necesario un estudio más detallado para concretar qué parejas tendrían más aceptabilidad. En el caso de parejas cuya frecuencia de aparición es muy baja (como las 1537 parejas anteriores) sería necesario comparar en otros corpus la presencia real de estas parejas para definir si son útiles o no esteganográficamente y cómo se distribuyen.

b) Nom+Adj1+Adj2. Con respecto a la ocultación basada en el intercambio del orden de dos adjetivos que siguen a un nombre en una estructura del tipo Nom+Adj1+Adj2, las mediciones indican que esta estructura no es útil esteganográficamente. En primer lugar por su baja capacidad de ocultación (0,009472 bits/frase) y en segundo lugar porque las colocaciones observadas de esta estructura hace que el orden de las palabras sea muy estricto. Algunos ejemplos de estas colocaciones son: autoridades sanitarias estadounidenses (6 veces), conversación especializada compulsiva (4 veces), escuchas telefónicas ilegales (3) y trampa electrónica imprevista (1 vez). Un ejemplo concreto extraído del corpus donde puede verse es (5):

(5) […] la investigación que nuestro compañero Melchor_Miralles estaba realizando sobre la trama de *escuchas telefónicas ilegales* descubierta en Barcelona en el entorno del editor de La_Vanguardia.

c) Prep+nom+prep+nom. Al igual que la estructura anterior, el movimiento de los elementos que constituyen frases preposicionales del tipo prep+nom+prep+nom o prep+art|det+nom+prep+nom no se muestran tampoco de utilidad en esteganografía lingüística. El estudio de colocaciones refleja que es difícil justificar reglas generales para el movimiento de estos elementos. Algunos ejemplos de colocaciones para la estructura prep+nom+prep+nom son: (26) a finales del siglo, (26) del presidente del gobierno, (21) de par en par, (16) del cuarto de baño, (15) al presidente del gobierno, (15) en tela de juicio, (1) en cuestiones de rango y (1) con olor a jazmines. Algunos ejemplos de colocaciones para la

estructura prep+art|det+nom+prep-nom: (107) desde el punto de vista, (32) de los medios de comunicación, (32) con el paso del tiempo, (31) desde un punto de vista, (21) en los medios de comunicación, (2) de los servicios de inteligencia, (1) en un núcleo de helio y (1) de la perdida de presión.

4.4. Transformación basada en la reordenación de los complementos del verbo.

4.4.1. Hipótesis lingüística

En general, la presencia en la oración de adverbios y otros complementos no subcategorizados por el verbo, como complementos circunstanciales externos al sintagma verbal, no es obligatoria y por lo tanto se puede predecir que podrían gozar de mayor movilidad y aceptabilidad ante la manipulación en general. Según Seco (1985: 174, 175), atendiendo a su significación hay dos clases de adverbios: el tipo 1, los de lugar, tiempo, modo e intensidad y el tipo 2 que "se refieren a la existencia misma, a la realidad, a la sustancia de lo significado por la palabra o grupo de palabras acompañado por aquellos". Con respecto al primer tipo en la clasificación de Seco, sí pueden apreciarse diferencias de aceptabilidad ante la manipulación sintáctica (1_1, 1_2,1_3, 2_1,2_2, 2_3)

(1_1) Alfonso estudió ese asunto ayer.
(1_2) Ayer Alfonso estudió ese asunto.
(1_3) Alfonso estudió ayer ese asunto
(2_1) Alfonso estudió ese asunto a fondo.
(2_2) *A fondo Alfonso estudió ese asunto.
(2_3) Alfonso estudió a fondo ese asunto.

El segundo tipo de adverbios suele separarse del resto de la oración con una coma y en principio su movilidad no parece afectar a la aceptabilidad (3_1,3_2), puede sustituirse por otro adverbio con significado similar (4_1,4_2) y su omisión no afecta a la sintaxis ni al significado de la oración aunque la actitud del hablante asociada al adverbio omitido no es recuperable (4_3).

(3_1) Desgraciadamente, todo ha terminado.
(3_2) Todo ha terminado, desgraciadamente.
(4_1) Lamentablemente, todo ha terminado.
(4_2) Todo ha terminado, lamentablemente.
(4_3) Todo ha terminado.

Dado que el movimiento de un adverbio parece que tiene cierta libertad, aparentemente más libertad de movimiento que la de un sintagma adverbial, cabe analizar hasta que punto esto es así y para qué tipo de adverbios.

4.4.2. Experimentación de la transformación

A continuación se va a medir la presencia en los corpus bajo estudio de diferentes estructuras. Todas las estructuras se basan en la posibilidad de mover un adverbio o elementos adyacentes a él en diferentes posiciones.

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 229-247
Recibido: 30/12/2009
Aceptación comunicada:14/01/2010

Transformación basada en la reordenación de los complementos del verbo.	**Corpus LEXESP**	
	Nº frases	Ocurrencias (%) (total 214.400 frases)
ADV+COMA|COMA+ADV	**33544|31822**	**15,6455% | 14,8423%**
VERBO+ADV|ADV+VERBO	**56803|62195**	**26,4939% | 29,0088%**
ADV principio frase | ADVfinal frase	**32741|11428**	**15,2709% | 5,3302%**
VERBO+COMPLEMENTO+ADV		
a) Complemento = art/det+nom+[adj]	**6053**	**2,8232%**
b) Complemento = prep+nom	**1723**	**0,8036%**
c) Complemento = prep+ Infinitivo	**2061**	**0,9612%**
ADV lugar | ADV tiempo	**9292|27030**	**4,3339% | 12,6072%**
ADV modo | ADV cantidad	**9800|36112**	**4,5708% | 16,8432 %**

Tabla 5: Cuantificación de estructuras útiles en reordenación de complementos del verbo. CORPUS LEXESP.

Transformación basada en la reordenación de los complementos del verbo.	**Corpus TWITTER**	
	Nº frases	Ocurrencias (%) (total 319.483 frases)
ADV+COMA|COMA+ADV	**5824 | 7468**	**1,8229% | 2,3375%**
VERBO+ADV|ADV+VERBO	**31173 | 25750**	**9,7573% | 8,0598%**
ADV principio frase | ADVfinal frase	**13388 | 8849**	**4,1905% | 2,7697%**
VERBO+COMPLEMENTO+ADV		
a) Complemento = art/det+nom+[adj]	**2206**	**0,6904%**
b) Complemento = prep+nom	**29**	**0,0090%**
c) Complemento = prep+ Infinitivo	**592**	**0,1852%**
ADV lugar | ADV tiempo	**4845 | 39950**	**1,5165% | 12,5045%**
ADV modo | ADV cantidad	**19279 | 18148**	**6,0344% | 5,6804%**

Tabla 6: Cuantificación de estructuras útiles en reordenación de complementos del verbo. CORPUS TWITTER.

Dada la cantidad ingente de datos a analizar las medidas de la Tabla5 y Tabla6 indicarían la cota superior de capacidad de ocultación que permitiría cada estructura si tuviera utilidad esteganográfica en un caso ideal. Por motivos de espacio, se presentan solo los resultados del análisis de un tipo de adverbio (el que a priori parece más productivo con fines esteganográficos) en algunas de las estructuras indicadas.

Se presenta a continuación el análisis del movimiento de los adverbios de tiempo (ahora, antaño, antes, aún, ayer, cuando, constantemente, después, enseguida, hogaño, hoy, luego, mientras, mañana, nunca, recién, recientemente, temprano, todavía, ya) por dos motivos: a) la presupuesta utilidad esteganográfica basada en las hipótesis lingüísticas realizadas y los trabajos en lengua inglesa (Murphy 2001) y b) por ser uno de los adverbios que permite más capacidad de ocultación (el 12,60% de las frases del corpus LEXESP tiene un adverbio de tiempo). Para ello, se va a medir la presencia de adverbios de tiempo al principio de frase y al final de frase (con o sin coma adyacentes) para definir si es posible intercambiar el adverbio de tiempo del principio al final de la frase o viceversa con el fin de ocultar de esta manera información.

Las medidas muestran 4625 frases con un adverbio de tiempo al principio de frase (2,1571% del total) y 878 frases donde ese adverbio precede a una coma. En el caso de la presencia del adverbio de tiempo al final de frase se detectan 1016 frases (0,4738% del total) y 51 frases con coma que le antecede. En primer lugar destaca la mayor presencia de adverbios de tiempo al principio de frase que al final (unas cuatro veces más en este corpus). Esta primera característica ya establece limitaciones a la hora de intercambiar las posiciones principio-final de frase. Si nos fijamos en las frecuencias de aparición de cada adverbio de

tiempo en esta posición queda claro que ciertos adverbios tienen una presencia más clara al principio de frase y otros al final, esto puede implicar que en caso de utilidad esteganográfica, la transformación podría sólo ser interesante en una dirección, por ejemplo del final al principio y no al revés. La observación de ejemplos del corpus parece indicar que si las frecuencias de aparición del mismo adverbio al inicio y al final se igualan ésta condición se relajaría y sería más fácil el movimiento en ambas direcciones. Algunos ejemplos donde se puede ver la libertad de movimiento del adverbio de principio a fin o viceversa son:

(1097 veces) #ya, (66 veces) ya#[4]

($1_{1\#}$) Resulta entonces que no practica las historias del espejo, del tomavistas o de todos esos inventos que tan poca ilusión nos producen *ya.*
($1_{2\#}$) […] de los fracasos que no dan tiempo a complacerse en la esperanza porque hay que plantar *ya.*
($2_{\#1}$) *Ya* no tenía novia, ni leía prensa, ni salía de copas, ni iba a bailar, ni tan siquiera dormía.
($2_{\#2}$) *Ya* ves, comprando un periódico para mirar los cines.

(117 veces) #antes, (136 veces) antes#

($3_{\#1}$) *Antes* que Occidente cantara, la gente de nuestra cultura tenía la garganta quebrada de tanto alzar la voz.
($3_{\#2}$) *Antes* rezaba sin razón alguna.
($4_{1\#}$) Venía de Cantón y fue la última nave mercantil que siguió una ruta abierta doscientos cincuenta años *antes.*
($4_{2\#}$) Ha crecido de_golpe_y_porrazo esta Ana, y en sólo tres días habla de cosas de las que nunca había hablado *antes*.

En el caso de que el adverbio de tiempo esté seguido o precedido de coma su movimiento y por tanto la aceptabilidad parecen ser mayores. Véase algunos ejemplos de LEXESP para los adverbios con coma más frecuentes (ahora, luego, después, hoy):

($5_{\#1}$) *Ahora*, este hombre, que sigue tan loco como antes, es feliz. *[, ahora]*
($5_{\#2}$) *Hoy*, cuerpo joven es igual a imagen cotizada. *[, hoy]*
($5_{\#3}$) *Luego*, viene la recaída, el paso atrás, el retroceso.
($5_{\#4}$) *Después*, en la estrategia de equipo, se hablará de los puntos más peligrosos.

Por otro lado algunos ejemplos de coma más adverbio de tiempo al final de frase para los adverbios más frecuentes (ya, nunca, todavía) serían:

($6_{1\#}$) Tú dile a la gallina cosas feas y ya verás, *ya*.
($6_{2\#}$) Nadie quiso hablar, *nunca*.
($6_{3\#}$) Algunos maliciosos afirman que en_realidad no le hizo caso porque no era lo suficientemente rico, *todavía*.

[4] El # a la izquierda del adverbio significa que está situado al principio de frase. El # a la derecha que está situado al final.

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 229-247
Recibido: 30/12/2009
Aceptación comunicada:14/01/2010

En los ejemplos anteriores parece existir una cierta flexibilidad al intercambiar adverbio y coma al principio o final de frase, no obstante debe considerarse la estadística de aparición de los adverbios en cada posición para no crear patrones esteganográficos detectables. Por ejemplo, en el caso de LEXESP la estructura coma más los adverbios luego, mientras, cuando, enseguida, antaño o aún nunca aparecen al final de frase.

Visto lo anterior, parece que puede ser provechoso un estudio en profundidad de la utilidad del movimiento de adverbios con utilidad esteganográfica, analizando los tipos de adverbios útiles y dentro de estos, si existen subconjuntos interesantes o no en determinadas estructuras como las medidas en la Tabla1. El estudio de colocaciones facilitará descartar determinadas combinaciones.

5. Conclusión. Trabajo Futuro

Este trabajo aborda una nueva línea de investigación relacionada con el uso de los avances en el procesamiento del lenguaje natural en lengua española aplicado al mundo de la protección de comunicaciones, enmascarado de información y marcado digital de textos. Los resultados iniciales sintetizados en este artículo se centran en la posibilidad de realizar modificaciones sintácticas en textos escritos en lengua española con fines esteganográficos, tarea nada sencilla. De hecho, el orden de las palabras determina hasta tal punto la "hechura" de una lengua, como demostró Greenberg (1966), que no admite trivialidades en su manipulación.

En primer lugar, dado la poca información existente respecto a este tipo de investigación, se han seleccionado una serie de estructuras siguiendo estudios previos en otras lenguas y se ha seleccionado otras estructuras nuevas basadas en hipótesis lingüísticas. En segundo lugar, se ha medido la presencia de estas estructuras en dos corpora y se ha profundizado en el análisis de alguna de ellas basándose en mediciones del corpus LEXESP. La intención principal de esta primera aproximación es describir un proceso para la detección de estructuras con posible utilidad esteganográfica y alcanzar algunas conclusiones iniciales para ver hasta qué punto resulta interesante realizar estudios más profundos sobre algunas de las estructuras detectadas y medidas. Queda para trabajo futuro detectar nuevas estructuras con utilidad esteganográfica y realizar estudios más exhaustivos para las estructuras analizadas y medidas.

Respecto de las estructuras medidas y analizadas puede concluirse que en general, la modificación sintáctica de frases en lengua española con utilidad esteganográfica es muy cuestionable, ninguna de las estructuras analizadas tendrían a priori y según las medidas que se han descrito una utilidad práctica por sí sola. Por ejemplo, estructuras sintácticas como la pasiva tiene una fuerte carga semántica en español lo que hace que en principio no sea productiva para la ocultación de información en texto escrito, mientras que estructuras sintácticas menos complejas, como pueden ser el movimiento de adjetivos alrededor de un sustantivo o adverbios en una oración, tengan una aceptabilidad mayor y puedan permitir, si mediciones futuras así lo confirman, combinarlas para crear una aplicación útil con fines esteganográficos. De las estructuras analizadas queda pendiente profundizar más en el estudio de los movimientos de adverbios ya que es presumible que tengan más utilidad esteganográfica.

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 229-247
Recibido: 30/12/2009
Aceptación comunicada:14/01/2010

Referencias bibliográficas

Almela, R., P. Cantos, A. Sánchez, R. Sarmiento y M. Almela. eds. 2005. *Frecuencias del español. Diccionario y estudios léxicos y morfológicos*. Madrid: Editorial Universitas, S.A.

Bergmair, R. 2007. A comprehensive Bibliography of Linguistic Steganography. *International Conference on Security and Steganography*. Vol. 6505, January 2007.

Calvo, H. e I. Bolshakov. 2004. Using selectional preferences for extending a synonymous paraphrasing method in steganography. En J. H. S. Azuela. *Avances en Ciencias de la Computación e Ingeniería de Cómputo*. *CIC'2004*: 231–242.

Cano Aguilar, R., ed. 1987. *Estructuras sintácticas transitivas en el español actual.* Madrid: Gredos.

Carracedo, J. 2004. *Seguridad en Redes Telemáticas.* Madrid: Mc-Graw Hill InterAmericana de España.

Chapman, M. 1997. *Hiding the hidden: a software system for concealing ciphertext as innocuous text*. Master of Science in Computer Science, University of Wisconsin-Milwaukee.

Fernández, S. 1993. Sobre el orden de palabras en español. *Dicenda. Cuadernos de Filología Hispánica* 11: 113-152.

Greenberg, J. 1966. *Some universals of grammar with particular reference to the order of meaningful elements*. Cambridge, Mass: MIT Press.

Horstein, N. y D. Lightfoot. 1981. *Explanation in linguistics*. Londres: Longman.

Khan, D. 1967. *The code breakers. The comprensive History of Secret Communication from Ancient Times to the Internet*. New York: Scribner.

Jackendoff, R. 1977. *X'syntax. A study of phrase structure*. Cambridge, Mass: MIT Press.

Lázaro, F. 1980. Sobre la pasiva en español. *Estudios de lingüística*. 61-70. Barcelona: Crítica.

Muñoz, A. 2009a. Corpus Twitter de 103 usuarios españoles. [Disponible en http://stelin.sourceforge.net/corpustwitter.zip]

Muñoz, A. 2009b. Generating Spanish Stegotext for fun and profit. Congreso de Seguridad Informática RootedCon 2010. [Comunicación Aceptada] [Disponible en: http://stelin.sourceforge.net/rooted.pdf]

Muñoz, A. y J. Carracedo. 2007. StegSecret: Una herramienta de estegoanálisis pública. Anales del IV Congreso Iberoamericano de Seguridad Informática. [Disponible en http://stegsecret.sourceforge.net]

Muñoz, A., J. Carracedo y S. Sánchez. 2008. Detection of distributed steganographic information in social networks. *EATIS 2008. ACM-DL Proceedings ISBN: 978-1-59593-988-3.*

Muñoz, A. y J. Carracedo. 2009a. Stelin. Una herramienta pública para generación automática de estegotextos en lengua española. JITEL 2009. [Disponible en http://stelin.source forge.net]

Muñoz, A. y J. Carracedo. 2009b. Estegoanálisis aplicado a la generación automática de estegotextos en lengua española. *Actas del V Congreso Iberoamericano de Seguridad Informática* CIBSI'09: 310-324.

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 229-247
Recibido: 30/12/2009
Aceptación comunicada:14/01/2010

Murphy, B. 2001. *Syntactic information hiding in plain text*. Master's thesis, Computer Science, Trinity College Dublin.

Murphy, B. y C. Vogel. 2007. The syntax of concealment: reliable methods for plain text information hiding. *International Conference on Security*, January 2007.

Wayner, P. 1992. Mimic functions. *Cryptologia XVI*: 193–214.

Wayner, P. 1995. Strong theoretical steganography. *Cryptologia XIX:* 285–299.

Schmid, H. 2009. TreeTagger - a language independent part-of-speech tagger. Institute for Computational Linguistics of the University of Stuttgart. [Disponible en http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/]

Sebástian, N., M. Martín, M. Francisco, y F. Cuestos, eds. 2000. *LEXESP Léxico informatizado del español*. Barcelona: Edicions de la Universitat de Barcelona.

Seco, M., ed. 1985. *Gramática esencial del español.* Madrid: Aguilar.

Zhi-li, C, H. Liu-Shen, Y. Zhen-shan, Z. Xin-xin, y Z. Xue-ling. 2008. Effective Linguistic Steganography Detection. *IEEE 8th International Conference on Computer and Information Technology Workshops.*

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 229-247
Recibido: 30/12/2009
Aceptación comunicada:14/01/2010