MODIFICACIONES SINTÁCTICAS BASADAS EN LA REORDENACIÓN DE COMPLEMENTOS DEL VERBO CON UTILIDAD EN ESTEGANOGRAFÍA LINGÜÍSTICA

ALFONSO MUÑOZ
IRINA ARGÜELLES
UNIVERSIDAD POLITÉCNICA DE MADRID

Resumen. En la actualidad, los avances en la ciencia de la esteganografía lingüística en español abren nuevas líneas de investigación en su aplicación a la protección / privacidad de las comunicaciones digitales y en el marcado de textos. El presente artículo profundiza en el interés del uso de la reordenación de complementos del verbo en textos existentes en lengua española con utilidad en esteganografía lingüística y en el marcado digital de textos (marca de agua).

Palabras clave: *esteganografía lingüística, marcado digital, complementos del verbo, modificaciones sintácticas.*

Abstract. At present the advances in the science of linguistic steganography in Spanish open new lines of research in its application for the protection / privacy of digital communications and in the marking of texts. This article studies the possible interest of reordering complements of the verb in existing texts in Spanish language with regard to its usefulness in linguistic steganography and in digital marking of texts (watermarks).

Keywords: *linguistic steganography, watermarking, verb complements, syntactic modifications.*

1. Esteganografía lingüística. Ocultando información en el lenguaje natural

En la última década, bajo el paraguas de la lingüística computacional se ha aprovechado el conocimiento disponible en áreas como el análisis del discurso, la lingüística de corpus, la lexicografía o la estadística de palabras para su aplicación a tecnologías muy diversas como los algoritmos de reconocimiento del habla, los sistemas de traducción automática, sistemas de *data mining*, algoritmos de análisis ortográficos, resumen automático de textos, etc. Todo este conocimiento lingüístico y algorítmico ha permitido que en este tiempo se haya avanzado en una nueva aplicación, especialmente en lengua inglesa: la posibilidad de ocultar información en lenguaje natural, es decir, 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 de dicha información (Carracedo 2004: 123-131). 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.

Cuando la cubierta o tapadera es un texto en lenguaje natural se habla de un tipo concreto de esteganografía, la esteganografía lingüística, y el texto que se modifica o se genera se denomina estegotexto. 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. 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 los principios de no oscuridad de Kerckhoffs (Kahn

REVISTA ELECTRÓNICA DE LINGÜÍSTICA APLICADA (ISSN 1885-9089)
2011, Número 10, páginas 31-54
Recibido: 09/12/2010
Aceptación comunicada: 14/03/2011

1967). La aplicación de este principio a la 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 el emisor y el 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). Esta nueva línea de investigación tiene aplicación en dos grandes áreas: a) anonimato y privacidad de las comunicaciones y b) marcado digital de textos.

El uso de la esteganografía lingüística con aplicación en la privacidad de las comunicaciones abre la posibilidad de ocultar información en textos en lenguaje natural permitiendo intercambiar datos a la vez que se dificulta su detección a personas y sistemas de monitorización automáticos (Echelon, Carnivore, Sitel, etc.) y por tanto tiene utilidad en términos de libertad de expresión. Pero, sin duda, es en el área del marcado digital de textos donde esta tecnología podría tener suficiente interés para profundizar en su investigación.

En la actualidad, la integridad y la autenticidad de una información pueden ser garantizadas mediante la utilización de la denominada firma digital. En general, se trata de 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 de que la información generada no está autocontenida en la información que se quiere proteger, luego pueden ser separadsa 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 y si fuera difícil que un atacante pudiera eliminar la marca sin afectar significativamente al texto portador, permitiría la inclusión autocontenida de firmas que podrían tener utilidad en su aplicación a la 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 además se puede 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.

Los mejores procedimientos documentados hasta la actualidad para alcanzar este objetivo (Bergmair 2007) hacen referencia a la posibilidad de modificar textos originales para crear estegotextos con la información oculta/autocontenida, por ejemplo, una firma digital. Algunos de los procedimientos útiles con este fin son:

a) Modificaciones Léxicas. Estos procedimientos consisten en la ocultación de información mediante la sustitución/modificación de palabras. El método más analizado 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

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 idiomas. 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 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 (Shirali-Shahreza 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 estegotexto 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 necesitarían 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 describe brevemente los resultados obtenidos en el trabajo previo (Muñoz *et al.* 2010) y se enlaza con los nuevos resultados.

a) 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.[1]

[1] En el apartado 3.2, más adelante, volveremos a este corpus y su descripción.

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.

Si se consideran las medidas observadas en LEXESP se podrían ocultar (a razón de 1 bit por frase donde fuera posible aplicar esta transformación) 3,13 bits/100 frases (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. A estos datos deben sumarse las limitaciones al considerar la estadística de la aparición de estas estructuras, así por ejemplo en LEXESP la estructura ser+participio+por sucede en media cada 174,36 frases y 31,9417 frases por cada frase con ser+participio. En términos de aceptabilidad los ejemplos observados en este corpus no permiten afirmar que la automatización de esta transformación sea sencilla sin perjuicios.[2]

b) Transformación basada en movimientos de los adjetivos y otros complementos dentro del sintagma nominal. Esta línea de investigación, avanzada en Muñoz *et al.* (2010), tiene como intención detectar estructuras lingüísticas que complementen/modifiquen un sintagma nominal que pudieran tener utilidad en esteganografía lingüística y marcado digital de textos. De las estructuras analizadas, se centró el interés en la posibilidad de mover un adjetivo que modifica a un sustantivo al ser una estructura que aparece con una frecuencia alta**.** Se midió la posibilidad de anteponer o posponer un adjetivo que modifica un sustantivo y por tanto, ocultar un bit de información (bit 0/1 adjetivo delante, bit1/0 adjetivo detrás).[3]

Transformación basada en movimientos de complementos dentro del sintagma nominal	**Corpus LEXESP**	
	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 2: Cuantificación de estructuras útiles en movimientos de complementos del sintagma nominal. CORPUS LEXESP.

Midiendo en el corpus LEXESP se buscaron parejas sustantivo|adjetivo (por ser su presencia elevada) que aparecieran en sus dos formas con adjetivo antepuesto (adj+sust) y postpuesto (sust+adj). En estas condiciones se encontraron en LEXESP 3.520 parejas que cumplen esta condición. Las mediciones indican que según se va igualando la co-ocurrencia de las dos estructuras posibles (adj+sust o sust+adj) más fácil es que la aceptabilidad lingüística sea mayor al alternar arbitrariamente la posición del adjetivo y por tanto tendrían utilidad esteganográfica.

[2] Para más detalles véase (Muñoz *et al.* 2010).
[3] Para más detalles véase (Muñoz *et al.* 2010).

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

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

Si se fija la atención en las 1.537 parejas que sólo se encuentran una vez en el corpus, es más sencillo ver la flexibilidad de posponer o anteponer el adjetivo que califica al sustantivo (2,3).

(2) Pareja: rectángulo pequeño – pequeño rectángulo

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

(3) Pareja: campaña feroz – feroz campaña

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

La investigación en curso está determinando en corpus más amplios qué parejas presentan definitivamente una aceptabilidad interesante en el campo de la esteganografía lingüística y se está analizando la estadística de aparición de las parejas para cuantificar cual sería el volumen de la información que se podría ocultar. Es posible que para implementar una herramienta práctica fuera necesario combinar la aplicación de varias estructuras como esta (adj+sust – sust+adj).

3. Transformación basada en la reordenación de complementos del verbo

En lo que sigue, se va a profundizar en la posibilidad de realizar modificaciones sintácticas con utilidad en esteganografía lingüística en lengua española basada en la reordenación de complementos del verbo. La metodología de análisis empleada consistirá en primer lugar en la descripción de una hipótesis lingüística con la que trabajar, a continuación se realizará la experimentación de diversas transformaciones basadas en el estudio en dos corpus con características distintas.

3.1. Hipótesis Lingüística

En lengua inglesa se ha documentado (Murphy y Vogel 2007) la posibilidad del movimiento de complementos, palabras o frases adverbiales sin que esto afecte al significado de la misma. Según esto, un algoritmo esteganográfico debería poder asignar códigos binarios diferentes a cada variación de una frase concreta, por ejemplo en lengua inglesa podrían establecerse 3 variantes del siguiente ejemplo: a) *Often the dog chased the cat, b) the dog often chased the cat y c) the dog chased the cat often*, pudiendo ocultar $\log_2 3 = 1{,}58$ bits/frase. En este sentido,

Murphy (2001) adelantó la posibilidad de utilizar complementos de tiempo y lugar con esta utilidad, por ejemplo: 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*.

En lengua española no es obligatoria, en general, la presencia en la oración de adverbios y otros complementos no subcategorizados por el verbo, como complementos circunstanciales externos al sintagma verbal. Esto debería suponer que estos elementos podrían gozar de mayor movilidad y seguir siendo aceptables ante la manipulación. En nuestro caso concreto, nos centramos en los adverbios atendiendo a su significado. Estos, según Seco (1985) son de dos tipos: 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, pueden apreciarse diferencias de aceptabilidad ante la manipulación sintáctica que dependen del adverbio y de la cercanía con la que este modifica al verbo (4a, 4b, 4c, 5a, 5b, 5c, 5d):

(4) a. Alfonso analizará esa posibilidad mañana.
b. Mañana Alfonso analizará esa posibilidad.
c. Alfonso analizará mañana esa posibilidad.

(5) a. Alfonso analizará esa posibilidad detenidamente.
b. *Detenidamente Alfonso analizará esa posibilidad.
c. Alfonso analizará detenidamente esa posibilidad.
d. Alfonso analizará esa posibilidad.

El segundo tipo de adverbios se separara del resto de la oración con una coma en la mayoría de los casos. Su movilidad no parece afectar a la aceptabilidad (6a, 6b) y puede sustituirse por otro adverbio con un significado parecido (7a, 7b). Su omisión no afecta al significado de la oración ni a la sintaxis pero la posición del hablante, es decir, su actitud recogida en el adverbio omitido, no puede recuperarse (7c).

(6) a. Indudablemente, podemos creerlo o no.
b. Podemos creerlo o no, indudablemente.

(7) a. Evidentemente, podemos creerlo o no.
b. Podemos creerlo o no, evidentemente.
c. Podemos creerlo o no.

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

3.2. Experimentación de la transformación

A continuación se van a medir diferentes estructuras para cuantificar la utilidad del cambio de posición de un adverbio dentro de una oración en español con utilidad esteganográfica en dos corpora diferentes que se describen a continuación.

3.2.1. Descripción de los corpus en los que se experimentarán diferentes transformaciones.
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 géneros distintos como la narrativa (40%), la divulgación científica (10%), el ensayo (10%), la prensa (25%), semanarios (10%) y prensa deportiva (5%). La homogeneidad del registro de este corpus que prescinde de las variaciones dialectales del castellano, salvo aproximadamente un 10% de textos de autores hispanoamericanos y la heterogeneidad de los textos que integran el corpus, donde la selección quiere ser representativa de distintos géneros, hace que se utilice como corpus de referencia en este trabajo para analizar la utilidad de las transformaciones esteganográficas bajo estudio.

b) Corpus Wikipedia. En lingüística computacional a menudo un problema a resolver es disponer de recursos léxicos, como puede ser un corpus grande, con el cual realizar las medidas necesarias para la investigación. En la investigación en curso resulta interesante utilizar un corpus enorme de palabras para afinar las medidas de algunas estructuras de interés detectadas. En esta primera aproximación se utiliza LEXESP como corpus de referencia y un corpus mucho más grande basado en la Wikipedia en castellano para medidas más amplias, construyendo un corpus libremente accesible que puede permitir fácilmente a otros investigadores reproducir los resultados aquí expuestos. Para ello se decide construir un corpus real muy grande utilizando la Wikipedia en español (19GB) descargable de (Wikipedia 2008). De este volumen de datos se filtran 1.362.460 ficheros y sólo se extrae el texto disponible entre las etiquetas de los artículos principales (ignoramos comentarios, borradores y similares). Con estos datos se construye un corpus de 745MB con un total de 129 millones de palabras (entendemos por palabra cualquier elemento separado por un espacio). Para el etiquetado de este corpus se implementa un programa en lenguaje JAVA que utiliza el etiquetador TreeTagger (Schmid 2009), separando el texto por frases. Los resultados en las mediciones en este corpus deben entenderse con las limitaciones o imprecisiones de este etiquetador.

3.2.2. Medidas y experimentación
Este apartado analiza la presencia de diferentes estructuras lingüísticas en los corpus bajo estudio. Todas las mediciones se dirigen al objetivo de precisar el grado de movilidad de un adverbio o elementos adyacentes a él a diferentes posiciones dentro de la oración, lo cual permitiría establecer un procedimiento esteganográfico y por tanto ocultar información. A continuación, los razonamientos se realizarán considerando como corpus de referencia el corpus LEXESP al ser éste, como ya adelantábamos, más representativo (representación homogénea y heterogénea en cuanto a géneros que lo integran). Para ciertas estructuras de interés se comparará la presencia de las mismas en otro corpus más grande, el construido basado en la enciclopedia Wikipedia.

Las estructuras que se van a medir porque se parte de que podrían tener utilidad esteganográfica, son dos: a) parejas verbo+adv y adv+verbo, de tal forma que pueda analizarse el grado de aceptabilidad del movimiento de un adverbio respecto del verbo al que

modifica y b) movimiento de adverbios del principio al final de la frase y viceversa según las hipótesis documentadas anteriormente.

La tabla 3 y la tabla 4 cuantifican la aparición de estas estructuras en general y por tipo de adverbio en el corpus LEXESP y en el corpus Wikipedia respectivamente. Para estas mediciones se ha tenido en cuenta la siguiente clasificación: **Adverbio de tiempo** (*pronto, tarde, siempre, jamás, próximamente, prontamente, anoche, durante, ahora, antaño, antes, aún, ayer, cuando, constantemente, después, enseguida, hogaño, hoy, luego, mientras, mañana, nunca, recién, recientemente, temprano, todavía, ya, posteriormente, primeramente, primero, respectivamente*), **adverbio de lugar** (*adelante, adonde, ahí, aquí, allí, allá, abajo, arriba, cerca, delante, detrás, dónde, encima, lejos, atrás, debajo, fuera, junto, alrededor, acá, enfrente*), **adverbio de modo** (*regular, aprisa, peor, fielmente, estupendamente, fácilmente, así, bien, como, cual, igual, mal, mejor, según, adrede, despacio, rápido*), **adverbio de cantidad** (*apenas, bastante, casi, cuanto, demasiado, justo, más, menos, mucho, muy, nada, poco, sobremanera, tan, todo, sólo, algo, solamente, tanto, aproximadamente*), **adverbio de afirmación** (*bueno, claro, efectivamente, naturalmente, seguro, sí, también, verdaderamente, ciertamente, indudablemente, cierto, exacto),* **adverbio de negación** *(jamás, nada, no, nunca, tampoco*) y **adverbio de duda** (*quizá, quizás, acaso, probablemente, posiblemente, seguramente*).[4]

Transformación basada en la reordenación de los complementos del verbo.	**Corpus LEXESP**	**Palabras total** = 5.020.930
	Nº frases	Ocurrencias (%) (total 214.400 frases)
VERBO+ADV|ADV+VERBO	**56.803 | 62.195**	**26,4939% | 29,0088%**
Verbo-adverbio total= 69.390 ocurrencias totales		
Adverbio-verbo total= 69.758 ocurrencias totales		
Verbo+adv lugar | Adv lugar + verbo	3.248 | 1.859	1,5149% | 0,8670%
3326 ocurrencias totales | 1901 ocurrencias totales		
Verbo+adv modo | Adv modo+verbo	4.049 | 2.472	1,8885% | 1,1529%
4148 ocurrencias totales | 2566 ocurrencias totales		
Verbo+adv cantidad | Adv cantidad + verbo	17.108 | 8.312	7,9794% | 3,8768%
18387 ocurrencias totales | 8678 ocurrencias totales		
Verbo+adv tiempo | Adv tiempo + verbo	9.726 | 11.797	4,5363% | 5,5023%
10140 ocurrencias totales | 12506 ocurrencias totales		
Verbo+adv negación | Adv negación + verbo	4.383 | 34.242	2,0443% | 15,9710%
4571 ocurrencias totales | 38579 ocurrencias totales		
Verbo+adv afirmación | Adv afirmación + verbo	2.932 | 2.319	1,3675% | 1,0816%
2975 ocurrencias totales | 2342 ocurrencias totales		
Verbo+adv duda | Adv duda + verbo	174 | 506	0,0811% | 0,2360%
177 ocurrencias totales | 509 ocurrencias totales		

[4] En las mediciones realizadas se ha añadido los adverbios *jamás, nunca y nada* en la categoría de adverbios de negación. Adicionalmente se ha considerado interesante medir los adverbios *jamás* y *nunca* en la categoría de adverbios de tiempo y el adverbio *nada* en la categoría de adverbios de cantidad.

VERBO+COMPLEMENTO+ADV		
a) Complemento = art/det+nom+[adj]	6.053	2,8232%
b) Complemento = prep+nom	1.723	0,8036%
c) Complemento = prep+ Infinitivo	2.061	0,9612%
ADV PRINCIPIO FRASE | ADV FINAL FRASE	**32.741 | 11.428**	**15,2709% | 5,3302%**
Adv lugar principio | Adv lugar final	1.006 | 1.056	0,4692% | 0,4925%
Adv modo principio | Adv modo final	1.182 | 725	0,5513% | 0,3381%
Adv cantidad principio | Adv cantidad final	2.492 | 1.373	1,1623% | 0,6403%
Adv tiempo principio | Adv tiempo final	5.445 | 1.364	2,5396% | 0,6361%
Adv negación principio | Adv negación final	2.492 | 1.317	1,1623% | 0,6142%
Adv afirmación principio | Adv afirmación final	1.457 | 482	0,6795% | 0,2248%
Adv duda principio | Adv duda final	518 | 18	0,2416% | 0,0083%
ADV+COMA|COMA+ADV	**33.544|31.822**	**15,6455% | 14,8423%**
ADV principio frase + coma | Coma + ADV final frase	**9.589|1.044**	**4,4724% | 0,4869%**
Adv lugar inicio + coma | coma + Adv lugar final	207|9	**0,0965% | 0,0041%**
Adv modo inicio + coma | coma+ adv modo final	490|10	**0,2285% | 0,0046%**
Adv cantidad inicio + coma | coma + adv cantidad final	36|43	**0,0167% | 0,0200%**
Adv tiempo inicio + coma |coma + adv tiempo final	884|50	**0,4123% | 0,0233%**
Adv negación inicio + coma | coma + adv negación final	336|151	**0,1567% | 0,0704%**
Adv afirmación inicio + coma | coma + adv afirma final	419|174	**0,1954% | 0,0811%**
Adv duda inicio + coma | coma + adv duda final	25|4	**0,0116% | 0,0018%**
ADV total ocurrencias = 276.703		
ADV lugar | ADV tiempo	**9.895 | 33.330**	**4,6152% | 15,5457%**
Total ocurrencias = 10.847| Total ocurrencias = 39.756		
ADV modo | ADV cantidad	**9.840 | 41.438**	**4,5895% | 19,3274%**
Total ocurrencias= 10.473 | Total ocurrencias = 54.445		
ADV negación| ADV afirmación	**43.997 | 10.247**	**20,5209%|4,7793%**
Total ocurrencias= 61.359 | Total ocurrencias= 10.774		
ADV de duda o dubitativos	**2.052**	**0,9570%**
Total ocurrencias= 2.120		

Tabla 3: Cuantificación de estructuras útiles en reordenación de complementos del verbo. CORPUS LEXESP.

Transformación basada en la reordenación de los complementos del verbo.	**Corpus Wikipedia**	**Palabras= 128.453.788**
	Nº frases	Ocurrencias (%) (total 5.110.956 frases)
VERBO+ADV|ADV+VERBO	**771.527|584.921**	**15,0955%|11,4444%**
Verbo-adverbio total= 860.205		
Adverbio- verbo total= 636.772		
Verbo+adv lugar | Adv lugar + verbo	38.584|79.825	0,7549%|1,5618%

39202 ocurrencias totales | 82179 ocurrencias totales		
Verbo+adv modo | Adv modo+verbo	63.943|35.602	1,2510%|0,6965%
64900 ocurrencias totales | 36270 ocurrencias totales		
Verbo+adv cantidad | Adv cantidad + verbo	140.691|61.707	2,7527%|1,2073%
145656 ocurrencias totales | 62835 ocurrencias totales		
Verbo+adv tiempo | Adv tiempo + verbo	123.188|139.532	2,4102%|2,7300%
126.531 ocurrencias totales | 143.812 ocurrencias totales		
Verbo+adv negación | Adv negación + verbo	4.277|17.267	0,0836%|0,3378%
4324 ocurrencias totales | 17535 ocurrencias totales		
Verbo+adv afirmación | Adv afirmación + verbo	61.895|68.721	1,2110%|1,3445%
62338 ocurrencias totales | 69181 ocurrencias totales		
Verbo+adv duda | Adv duda + verbo	6921|7047	0,1354%|0,1378%
6948 ocurrencias totales | 7083 ocurrencias totales		
ADV PRINCIPIO FRASE | ADV FINAL FRASE	**200.132|60.905**	**3,9157%|1,1916%**
Adv lugar principio | Adv lugar final	5.571|4.755	0,1090%|0,0930%
Adv modo principio | Adv modo final	10.504|2.072	0,2055%|0,0405%
Adv cantidad principio | Adv cantidad final	9.385|3.077	0,1836%|0,0602%
Adv tiempo principio | Adv tiempo final	55.256|15.918	1,0811%|0,3114%
Adv negación principio | Adv negación final	1.847|498	0,0361%|0,0097%
Adv afirmación principio | Adv afirmación final	21.140|1.685	0,4136%|0,0329%
Adv duda principio | Adv duda final	2.542|37	0,0497%|0,0007%
ADV+COMA|COMA+ADV	**329|1.183|**	**0,0064%|0,0231%**
ADV principio + coma | Coma + ADV final	**39.485|2.870**	**0,7725%|0,0561%**
Adv lugar inicio + coma | coma + Adv lugar final	214|21	0,0041%|0,0004%
Adv modo inicio + coma | coma+ adv modo final	844|42	0,0165%|0,0008%
Adv cantidad + coma | coma + adv cantidad final	104|256	0,0020%|0,0050%
Adv tiempo + coma |coma + adv tiempo final	7.370|1.810	0,1442%|0,0354%
Adv negación inicio + coma | coma + adv negación final	15|8	0,0002%|0,0001%
Adv afirmación inicio + coma | coma + adv afirma final	1.034|36	0,0202%|0,0007%
Adv duda inicio + coma | coma + adv duda final	299|10	0,0058%|0,0001%
ADV total ocurrencias = 2.680.602		
ADV lugar | ADV tiempo	**5567|55251**	**0,1089%|1,0810%**
total ocurrencias = 265.676 | total ocurrencias = 592.753		
ADV modo | ADV cantidad	**10500|9383**	**0,2054%|0,1835%**
total ocurrencias = 194.955 | total ocurrencias = 392.494		
ADV negación| ADV afirmación	**1846|21140**	**0,0361%|0,4136%**
total ocurrencias = 30.079 | total ocurrencias =		

176.954		
ADV de duda o dubitativos	**10**	**0,0001%**
total ocurrencias = 25.313		

Tabla 4: Cuantificación de estructuras útiles en reordenación de complementos del verbo. CORPUS WIKIPEDIA.

La columna "ocurrencias" de las tablas anteriores refleja a primera vista una baja capacidad de ocultación basada en las estructuras medidas. Supongamos un caso ideal de transformación que fuera posible con el valor de ocurrencia más alta obtenido en LEXESP para las estructuras analizadas, es decir un 29,008% (suponiendo también que todas las modificaciones tienen aceptabilidad lingüística). Estaríamos hablando que de cada 100 frases sólo 29 tendrían utilidad esteganográfica. Si ocultamos como mínimo 1 bit por frase estaríamos hablando de 29 bits cada 100 frases o 0,29 bits por frase en media. Si hacemos un promedio del número total de frases en LEXESP y el número total de palabras tenemos que cada frase son unas 23,41 palabras. Luego 0,0123 bits de ocultación/palabra de media. Si quisiéramos ocultar sólo 20 bits necesitaríamos un texto de al menos 1.626,01 palabras, para 40 bits 3.252 palabras, etc. Estos valores a primera vista invitan a pensar que este tipo de estructuras no son productivas esteganográficamente o no lo son al menos para textos de tamaño medio. Por ejemplo, sería difícil aplicar esto en artículos en prensa que pueden no llegar a las 1.000 palabras. En cualquier caso, y pese a este resultado inicial, se va a profundizar más en las dos estructuras lingüísticas principales medidas para ver en qué medida tendrían aceptabilidad lingüística. Si se consigue encontrar diferentes estructuras con aceptabilidad, podría ser que combinándolas, su ocurrencia en un texto pudiera hacer útil su uso en esteganografía lingüística.

3.2.3. Movimiento de adverbio respecto del verbo inmediato al que modifica

En este apartado se analiza la presencia de parejas verbo+adv y adv+verbo que cumplan la siguiente condición VERBO+ADV=ADV+VERBO, es decir que faciliten el movimiento del adverbio. Esto es así porque de las medidas realizadas son las estructuras que más aparecen y por tanto si tuvieran utilidad esteganográfica permitirían ocultar más información. Además, con la tecnología actual (etiquetadores) es relativamente sencillo localizar estas estructuras y alterarlas moviendo las palabras en la frase mediante software. Analizando este patrón se encuentra en el corpus LEXESP 3.449 parejas (verbo+adv=adv+verbo) y en el corpus WIKIPEDIA 36.848 parejas. A continuación, se observan algunos ejemplos del corpus LEXESP[5], se puede observar la aceptabilidad intercambiando en cada oración el adverbio delante y detrás del verbo.

(1093 veces) no era, (7 veces) era no

(8) a. **No era** mal bagaje si se une al saber, a la vocacion y a la capacidad para estimular a quienes confían.
b. Al_fin_y_al_cabo **no era** para tanto.

(9) a. Guardiola: "Ganaremos a Bolivia" "Lo que importaba de verdad en este encuentro **era no** perder el partido, ya_que hubiera sido muy duro para nosotros no puntuar, y, sobre_todo, dar una buena imagen para hacer olvidar el nefasto partido contra Corea.

[5] Los ejemplos están extraídos directamente del corpus y se muestra el texto literal de los extractos originales.

b. Hace pocos días, lúcidamente Juan_Goytisolo venía a decir que el mejor modo de no pasar nunca de_moda **era no** haber estado nunca ni estar jamás en esa ridícula situación.[6]

(177 veces) ya está, (37 veces) está ya

(10) a. Sea como fuere, el mal **ya está** hecho.
b. Con la primavera, que **ya está** encima, todas nos ponemos monísimas y hay mucho psicópata suelto.

(11) a. Tras tantos esfuerzos para convencernos de que todo está ya determinado y de que nada auténticamente imprevisto puede ocurrir […]
b. Todo el mundo **está ya** diciendo que en verano puede producirse una horrorosa saturación turística y un exceso de contrataciones de reservas de plazas […]

(46 veces) sólo era, (68 veces) era sólo

(12) a. Para los antiguos griegos, Selene, diosa que personificaba a la Luna, no **sólo era** dadora de vida, sino también destructora de ella […]
b. Pero eso **sólo era** el inicio de los problemas, porque en los próximos meses en otras provincias como Jaén y Córdoba habrá asamblea para elegir comités.

(13) a. Tal Pascua como hoy, hace ya años, vivíamos el alivio inconmensurable de una resurrección que no **era sólo** la del Cristo, sino de la ciudad entera, pasadas ya las agonías de la Semana_Santa.

b. La goma no era elegante, **era sólo** un mal menor del que apenas se hablaba en aquellos años de excomunión y misa diaria.

La observación de ejemplos del corpus parece indicar que si las frecuencias de aparición se igualan es más fácil mover el adverbio en ambos sentidos y por tanto, tener una mayor aceptabilidad lingüística. Con esta condición se miden las parejas, verbo+adv y adv+verbo, que tienen la misma frecuencia obteniendo en LEXESP **1.323** parejas (1182 con ocurrencia 1), **2.805** parejas si la diferencia de frecuencias es menor o igual que 5 y **3.105** parejas si la diferencia es menor o igual que 10. En Wikipedia **9.748** parejas iguales (7.690 ocurrencia 1), **26.572** parejas cuya diferencia de frecuencia es menor o igual que 5 y **31.044** parejas menor o igual que 10. A continuación se adjuntan ejemplos de parejas con igual frecuencia en LEXESP y WIKIPEDIA para observar la aceptabilidad lingüística en estas circunstancias.

CORPUS LEXESP:

(28 veces) son ya, (28 veces) ya son

[6] las parejas con adverbios de negación requieren una consideración adicional que se razonará en las conclusiones.

(14) a. Los amigos de don Darío **ya son** (**son ya**) harina de otro costal.
b. Diré día 2, aunque **ya son (son ya)** más de las doce y media.
c. Las posibilidades de Moracho y Sala, especialistas de 60 metros vallas, **son ya (ya son)** menores.
d. Puesto que unos y otros **son ya (ya son)** veteranos reincidentes con diversas investigaciones y estudios de campo -- y me refiero a trabajos de campo serios y no sólo a paseos etnográficos – […]

(13 veces) estaba sólo, (13 veces) sólo estaba

(15) a. Es hombre seguro de sí mismo y es ubicuo; no **sólo estaba (estaba sólo)** en todos los campos, sino en todas las casas, para mi desgracia.
b. **Sólo estaba (estaba sólo)** a mi lado el dueño de la voz, que me conocía, aunque yo lo desconociera.
c. El Barça regaló tres goles por fallos; el medio campo es de ellos, en el Barça **estaba sólo (sólo estaba)** Guardiola y el coraje de Bakero.
d. La puerta, que **estaba sólo (sólo estaba)** entrecerrada, se abrió con brusquedad.

(7 veces) tiene todavía, (7 veces) todavía tiene

(16) a. Se especula sobre si hay paz o no hay paz, se dice que EtA **todavía tiene (tiene todavía)** fuerza para seguir golpeando, se asegura que la cúpula parisina del Grapo está a_favor_del cese de la violencia, pero no responde de la gente que todavía pulula por España […]
b. Tampoco parece preocuparles la puesta en marcha de una Reforma_Laboral que **todavía tiene (todavía tiene)** pendiente de desarrollo una parte importante: la liberalización de la negociación colectiva.
c. ¿Tiene tanto poder la difamación - porque se cuenta siempre para mal - o tanto poder **tiene todavía (todavía tiene)** la sociedad cristiana, puritana, ancestral? No creo que haya que incitar a nadie a hacer nada.
d. Armadin, entre premoniciones inciertas, **tiene todavía (todavía tiene)** tiempo de ser un anfitrión ameno: "una vez tuve un sueño y estaba en un cementerio.

(1 veces) sentado cómodamente, (1 veces) cómodamente sentado

(17) a. Así lo hacemos los adictos a la barra de bar; así lo niegan quienes se acercan al cliente -**cómodamente sentado (sentado cómodamente)** a la mesa, que es la prolongación de la abuelesca mesa camilla- con una amplia carta en las manos y la pretensión de que el invierno puede conjurarse con un tournedó.
b. Hasta sus 94 anos regentó su restaurante, al final vestido de_paisano y **sentado cómodamente (cómodamente sentado)** debajo_de una bóveda provenzal de piedra, trabajo que compaginaba con ser alcalde del pueblo.

CORPUS WIKIPEDIA

Puede observarse la aceptabilidad invirtiendo el orden al igual que en los ejemplos anteriores:

(51 veces) estudió también, (51 veces) también estudió

(18) a. **Estudió también** 'Artes plásticas', pero no logró concluir los estudios.
b. **Estudió también** la música barroca en un curso en la universidad, prestando atención a Couperin, Rameau and Bach.

c. Trabajó como modelo mientras estudiaba la secundaria en el colegio San Mateo de California, **también estudió** en Attended Old Dominion University en Norfolk Virginia.

d. La Policía **también estudió** las sincronizaciones de las explosiones del metro: los informes iniciales habían indicado que ocurrieron en un período de casi treinta minutos.

(19 veces) era originariamente, (19 veces) originariamente era

(19) a. Según la Enciclopedia Británica, **originariamente era** un traje agrícola piamontés.
b. **Originariamente era** el sustento de algunas poblaciones costeras o isleñas.
c. **Era originariamente** un dios agrícola, del campo, la vegetación y la fecundidad, por lo que su ritual comprendía una serie de ritos de muerte y resurrección cíclicos anuales [..]
d. Poco se conoce de sus primeros años de vida, salvo que **era originariamente** un mercenario de los Hunos.

(1 veces) delegaba también, (1) también delegaba

(20) a. La política exterior era una esfera específica del faraón que controla todas las expediciones externas y la diplomacia, aunque **también delegaba** sus responsabilidades a terceros que dependían directamente del faraón.
b. No obstante, en la práctica se **delegaba también** en ellos la iniciativa política, en ausencia de control efectivo de la sociedad civil.

A continuación por motivos de espacio se añade exclusivamente una lista reducida de las parejas detectadas con el mismo valor de ocurrencia en el corpus LEXESP.

(28) son ya-ya son, (13) **estaba sólo-sólo estaba**, (11) uno solo-solo uno, (9) para tanto - tanto para, (7) tiene todavía - todavía tiene, (7) quedó sólo - sólo quedó, (7) llevaba ya - ya llevaba, (6) tenemos también - también tenemos, (6) sería ya - ya sería, (6) **podía más - más podía**, (5) van siempre - siempre van, (5) **como ya - ya como**, (4) serán siempre - siempre serán, (4) da también - también da, (4) resulta aún - aún resulta, (4) comprendí entonces - entonces comprendí, (4) es a_menudo - a_menudo es, **(4) quiere más - más quiere**, (4) queda también - también queda, (4) saben bien - bien saben, (3) estamos siempre - siempre estamos, (3) quedo también - también quedo, (3) da ya - ya da, (3) resulta ya - ya resulta, (3)

sabré nunca - nunca sabré, **(3) dices no - no dices**, (3) podrían también - también podrían, (3) era acaso - acaso era, (3) es mañana - mañana es, (3) somos también - también somos, **(3) afirma no - no afirma**, (3) sintió de_pronto - de_pronto sintió, (3) produce sólo - sólo produce, (3) estará aquí - aquí estará, (3) era por_lo_menos - por_lo_menos era, (3) aparecen ya - ya aparecen, (3) lleva ahora - ahora lleva, (3) tienen aún - aún tienen, (3) acaba siempre - siempre acaba, (3) viene después - después viene, (3) hacen siempre - siempre hacen, (3) quieren ahora - ahora quieren, (3) entrado ya - ya entrado, **(3) vivo más - más vivo**, (2) estaba entonces - entonces estaba, (2) rodeado siempre - siempre rodeado, (2) tiene bastante - bastante tiene, (2) descubrí de_pronto - de_pronto descubrí, (2) fue posteriormente - posteriormente fue, (2) **cambiado tan - tan cambiado**, (2) creo personalmente - personalmente creo, (2) sirve ahora - ahora sirve, (2) daba siempre - siempre daba, (2) parecía entonces - entonces parecía, (2) hablo solo - solo hablo, (2) permite sólo - sólo permite, **(2) temía no - no temía**, (2) tenía todavía - todavía tenía, (2) ganó ayer - ayer ganó, (2) ocurrió entonces - entonces ocurrió, (2) parecía también - también parecía, **(2) temo no - no temo**, (2) trabajaba ya - ya trabajaba, (2) salía siempre - siempre salía, (2) tuvo aquí - aquí tuvo, (2) dieron nunca - nunca dieron, (2) deja solo - solo deja, (2) estaba arriba - arriba estaba, (2) sufren también - también sufren, **(2) beber no - no beber**, (2) pone hoy - hoy pone, (2) son probablemente - probablemente son, (2) puede hoy - hoy puede, (2) vi también - también vi, (2) conoció jamás - jamás conoció, (2) utilizan habitualmente - habitualmente utilizan, (2) falta todavía - todavía falta, **(2) sentimos no - no sentimos**, (2) tuvieron nunca - nunca tuvieron, (2) constituye también - también constituye, (2) este solo - solo este, **(2) media no - no media**, (2) estuvo ahí - ahí estuvo, (2) vemos ahora - ahora vemos, (2) viene también - también viene, (2) así sólo - sólo así, (2) son difícilmente - difícilmente son, (2) encuentra siempre - siempre encuentra, (2) paso ahora - ahora paso, (2) entran aquí - aquí entran, (2) estaba así - así estaba, (2) voy allá - allá voy, (2) pensaba entonces - entonces pensaba, (2) dijo solo - solo dijo, (2) es quizás - quizás es, (2) pensé entonces - entonces pensé, (2) hay probablemente - probablemente hay, (2) son por_lo_general - por_lo_general son, (2) siguen ahí - ahí siguen, (2) llevamos ya - ya llevamos, **(2) víctimas más - más víctimas**, (2) creo más_bien - más_bien creo, (2) tuvo aún - aún tuvo, (2) llegan nunca - nunca llegan, (2) quiere nada - nada quiere, (2) hablaba así - así hablaba, (2) vestido siempre - siempre vestido, (2) **escrito no - no escrito**, (2) fue finalmente - finalmente fue, (2) **parte no - no parte, (2) resultados más - más resultados, (2) quedaba nada - nada quedaba,** (2) publicados recientemente - recientemente publicados, (2) buscando siempre - siempre buscando, **(2) es cuanto - cuanto es**, (2) vinculado estrechamente - estrechamente vinculado, (2) decía también - también decía, (2) podrá siempre - siempre podrá, (2) ser tampoco - tampoco ser, (2) estaban antes - antes estaban, (2) son aproximadamente - aproximadamente son, (2) fue quizás - quizás fue, (2) terminan siempre - siempre terminan, (2) sugiere también - también sugiere, (2) tiene solamente - solamente tiene, (2) sigue todavía - todavía sigue, (2) aparecen siempre - siempre aparecen, (2) recuerda hoy - hoy recuerda, (2) es desde_luego - desde_luego es, (2) siento ahora - ahora siento, (2) gusta tanto - tanto gusta, (2) actúa también - también actúa, **(2) conseguirlo no - no conseguirlo**, (2) iban siempre - siempre iban, (2) parados más - más parados, (2) habla siempre - siempre habla, (2) tenían poco - poco tenían, (2) comenzó también - también comenzó, (2) pueden todavía - todavía pueden, (2) tenía casi - casi tenía, (1) lanzo ayer - ayer lanzo, (1) resulte finalmente - finalmente resulte, (1) habituado ya - ya habituado, (1) están nunca - nunca están, (1) sigue allí - allí sigue, (1) repetía machaconamente - machaconamente repetía, (1) dirigió apenas - apenas dirigió, (1) vio también - también vio, (1) pasaba así - así pasaba, (1) conserva todavía - todavía conserva, (1) quiero hoy - hoy quiero, (1) entendía entonces - entonces entendía, (1) buscar más - más buscar, (1) demuestra además - además demuestra, (1) perdió después - después perdió, (1) saldrá jamás - jamás saldrá, (1) entendemos también - también entendemos, (1)

dar también - también dar, (1) vimos antes - antes vimos, (1) prevé ya - ya prevé, (1) verla solo - solo verla, (1) eran tampoco - tampoco eran, (1) quedaba ahora - ahora quedaba, (1) hay sin_embargo - sin_embargo hay, (1) es por_lo_común - por_lo_común es, (1) destacado especialmente - especialmente destacado, (1) utilizado también - también utilizado, (1) hable así - así hable, (1182 parejas con ocurrencia 1) […]

Algunas de las parejas en la lista anterior han sido resaltadas en negrita para destacar peculiaridades de ciertas parejas cuya modificación podría afectar notoriamente al significado final. Un ejemplo de ello son las parejas con adverbio de negación como: temía no, beber no, etc. En este caso, el adverbio detrás del verbo suele modificar al verbo siguiente y por tanto, este movimiento, afecta al significado final de la oración, por ejemplo: no beber es malo, beber no es malo, temía no comer, no temía comer, etc. En estas circunstancias podrían considerarse criterios para hacer excepciones en ciertas oraciones, por ejemplo, en aquella que tuvieran un patrón verbo+adv+verbo, etc.

En cualquier caso, para analizar si algún tipo de adverbio concreto favorece más este movimiento dentro de las parejas detectadas se adjuntan las tablas 5 y 6. En la tabla 5 se muestra las parejas encontradas en LEXESP que cumplen la condición adv+verbo=verbo+adv para un tipo concreto de adverbio (lugar, modo, cantidad, tiempo, negación, afirmación, duda y acabados en mente). En la primera columna se muestra el número de ocurrencias en estas condiciones cuando en LEXESP se encuentra que para un mismo adverbio y verbo la ocurrencia de adv+verbo y verbo+adv son iguales. En la segunda columna el número de ocurrencias cuando la diferencia es menor o igual a 5 y en la tercera columna cuando son menores o iguales a 10. El mismo razonamiento se siguiente en la tabla 6 para el corpus Wikipedia.

En la tabla 5 se observa la siguiente frecuencia de los adverbios según su tipo y la ocurrencia de adv+verbo y verbo+adv. Según esto, si adv+verbo=verbo+adv se observa que: **tiempo > negación > cantidad > mente > modo > duda > lugar > afirmación**) y en los otros casos: **tiempo > negación > cantidad > mente > modo > lugar > duda > afirmación.** En Wikipedia para todos los casos analizados: **mente > tiempo > cantidad > modo > lugar > duda > negación > afirmación.**

TIPO ADV	Adv+Verbo=Verbo+A Corpus LEXESP Frecuencias Iguales	Adv+Verbo=Verbo+Adv Corpus LEXESP Frecuencias <=5	Adv+Verbo=Verbo+Adv Corpus LEXESP Frecuencias <=10
Lugar | Modo	6|32	23|76	24|87
Cantidad | Tiempo	112|348	215|762	243|838
Negación | Afirmación	181|5	402|10	477|10
Duda	11	20	21
Acabados en "mente"	108	177	188

Tabla 5: Tipo de adverbio encontrado en parejas que cumplen la condición adv+verbo = verbo+adv en el corpus LEXESP.

TIPO ADV	Adv+Verb=Verb+Adv Corpus Wikipedia Frecuencias Iguales	Adv+Verb=Verb+Adv Corpus Wikipedia Frecuencias <=5	Adv+Verb=Verb+Adv Corpus Wikipedia Frecuencias <=10
Lugar | Modo	224|301	623|978	795|1187
Cantidad | Tiempo	752|1764	2192|4885	2666|5757
Negación | Afirmación	137|71	346|161	418|174
Duda	183	456	510
Acabados en "mente"	5269	13887	15886

Tabla 6: Tipo de adverbio encontrado en parejas que cumplen la condición adv+verbo=verbo+adv en el corpus WIKIPEDIA.

En este punto, sería interesante analizar si las parejas encontradas se reproducen también en otros corpora. Así, por ejemplo, si se analiza la lista de parejas con frecuencias iguales en LEXESP y WIKIPEDIA sólo se encuentran **12** coincidentes cuando las frecuencias son iguales, **125** cuando las frecuencias son menores de 5 y **194** cuando las frecuencias menores de 10. Ejemplo de parejas adv+verbo y verbo+adv que presentan la misma frecuencia en cada corpus de manera individual y además se encuentran en ambos corpus analizados (12 parejas):

tenían poco | poco tenían, había únicamente | únicamente había, quedó ya | ya quedó, permanecer sólo | sólo permanecer, probar simplemente | simplemente probar, están nunca | nunca están, dio ya | ya dio, parecía realmente | realmente parecía, advirtió entonces | entonces advirtió, apareció enseguida | enseguida apareció, ocurría siempre | siempre ocurría, avanzó ya | ya avanzó.

En cualquier caso, las parejas detectadas, en general, muestran un conjunto de palabras donde sería viable el movimiento de los adverbios con utilidad esteganográfica y con una aceptabilidad lingüística razonable (esta aceptabilidad es mayor si las ocurrencias de las parejas se igualan). A pesar de estos resultados no es viable construir un procedimiento basado exclusivamente en esta estructura ya que su aparición, en los corpus analizados, es muy baja. Por ejemplo, si se fija la atención en el corpus LEXESP (Tabla 3) se observa que el patrón verbo+adv (69.390 ocurrencias) y el patrón adv+verbo (69.758 ocurrencias) aparece en un 26,4939% y 29,0088% de las frases. Realizando un cálculo simple tendríamos que en LEXESP existen 5.020.930/2 = 2.510.465 parejas de palabras. Si nos fijamos en las 1.323 parejas (adv-verbo=verbo+adv) en LEXESP que aparecen con la misma frecuencia vemos que en total, aparecen en todo el corpus 1.613 veces (la mayoría de las parejas tienen ocurrencia 1). Según esto, si el número total de parejas en LEXESP es 2.510.465 (100%), 1.613 parejas sería 0,0642% del total del corpus. En media, de cada 2.510.465/1.613=1556,3949 parejas una sería de utilidad esteganográfica. Si hacemos un promedio del número total de frases en LEXESP y el número total de palabras tenemos que cada frase son unas 23,41 palabras, luego 23,41/2=11,705 parejas/frase. De media y en LEXESP, serían necesarias 1556,3949/11,705=132,9683 frases para ocultar 1 bit de información, lo que descartaría la utilidad de utilizar este tipo de procedimientos esteganográficos. Este valor podría ser mayor para otros textos analizados.

3.2.4. Movimiento del adverbio desde el principio al final de frase y viceversa

Como se observó en la Tabla 3 la presencia de las estructuras adverbio al principio de frase (15,2709% de las frases), adverbio al final de frase (5,2202% de las frases), adverbio al principio de frase+coma (4,4724% de las frases) y coma+adverbio al final de frase (0,4869% de las frases) es muy poco interesante de manera individual con utilidad esteganográfica debido a su baja ocurrencia. No obstante, al igual que sucedía en el apartado anterior puede que su utilidad en conjunto con otras estructuras permitiera construir procedimientos prácticos con utilidad esteganográfica.

En la Tabla 3, para el corpus LEXESP, se refleja la presencia de un tipo de adverbio concreto para estas estructuras. Vamos a profundizar algo más en el caso de las estructuras adverbio principio frase+coma y coma+adverbio al final de frase, al suponer que su manipulación supone un impacto menor en la oración resultante. La ocurrencia de cada tipo de adverbio en el corpus para estos patrones puede observarse en la siguiente lista, donde se encuentran ordenados por ocurrencia de mayor a menor respectivamente: *tiempo* > *modo* > *afirmación* > *negación* > *lugar* > *cantidad* > *duda* y *afirmación* > *negación* > *tiempo* >

cantidad > *modo* > *lugar* > *duda*. El orden, independientemente de la aceptabilidad lingüística, así como su frecuencia de aparición (es más probable la estructura a principio de oración que al final) debe ser tomado en cuenta con fines esteganográficos para no introducir patrones estadísticos detectables por un analista potencial.

Aunque sería posible aplicar el razonamiento que se ha seguido en el apartado anterior y buscar adverbios (con coma o sin coma) que aparezcan con frecuencias similares al principio y al final de las oraciones, se presupone que, en general, esta condición puede ser inconsistente dado que dependerá, al menos, de la longitud de la oración y del número de complementos. En su lugar se prefiere abordar el problema desde otro punto de vista simplificando las oraciones bajo estudio y analizando la hipótesis de la movilidad de adverbios que se separan de una oración mediante el uso de coma. Para ello, a continuación, se ha simplificado el corpus LEXESP dividiéndolo en oraciones simples para facilitar el análisis. Consideramos por oración simple en este punto aquella frase que sólo tiene una coma al principio o al final de oración, precedida de un adverbio o delante del mismo, y descartando oraciones con punto y coma. Este pequeño corpus extraído de LEXESP, tiene 3.335 frases que empiezan por adverbio+coma y 581 frases que terminan por coma+adverbio. Estas cantidades, nos hacen ya intuir que la posición del adverbio detrás de la coma y al final de frase es mucho menos frecuente en español lo que induce a pensar que el movimiento de principio de frase a final de frase no va a resultar productivo para nuestros fines.

A continuación se adjunta una lista de ejemplos de estas estructuras para analizar la aceptabilidad lingüística. En negrita se adjunta una posible transformación de la oración. Recuérdese que la ocultación sería de 1 bit por oración, así si se desea ocultar un bit 0/1 se dejaría la frase como está y un bit 1/0 se le aplicaría la transformación basada en el movimiento del adverbio.

[Ejemplos de adverbios con coma al principio de frase]

[Adverbio de lugar]

(21) a. Aquí, una mujer de ésas no se comería ni un rosco / Una mujer de ésas no se comería ni un rosco **aquí.**
b. Allí, el equipo de Miguel_Angel_Martín jugaba la fase final de la Copa_del_Rey / El equipo de Miguel_Angel_Martin jugaba la fase final de la Copa_del_Rey **allí.**
c. Lejos, ladra un perro / Ladra un perro **lejos.**
d. Abajo, hinchamiento de la célula en fase inicial y formación de un cilio gigante / Hinchamiento de la célula en fase inicial y formación de un cilio gigante **abajo.**[7]

[Adverbio de modo]

(22) a. Así, un artículo tras otro / Un artículo tras otro**, así.**
b. Bien, ya tenemos memorizadas dos palabras / Ya tenemos memorizadas dos palabras**, bien.**

[Adverbio de cantidad]

[7] Con adverbios de lugar la presencia de coma al final no sería adecuada.

(23) a. Nada, salvo la abstención del cigarrillo / Salvo la abstención del cigarrillo**, nada.**
b. Todo, hicieron dinero todo / Hicieron dinero todo**, todo.**

[Adverbio de tiempo]

(24) a. Luego, acaso los asalta el deseo de que alguien los conozca leyendo lo que escriben / Acaso los asalta el deseo de que alguien los conozca leyendo lo que escriben**, luego.**
b. Hoy, cuerpo joven es igual a imagen cotizada / Cuerpo joven es igual a imagen cotizada, **hoy.**
c. Antes, en muchos lugares los francfurts se servían calentados en agua o al_vapor / En muchos lugares los francfurts se servían calentados en agua o al_vapor**, antes.**

[Adverbio de negación]

(25) a. No, no pondría yo aquí sobre mi mesa los menudos regalos / No pondría yo aquí sobre mi mesa los menudos regalos**, no.**
b. Nunca, en esta materia hay que dar pábulo al triunfalismo / En esta materia [no] hay que dar pábulo al triunfalismo**, nunca.**[8]
c. Nunca, ninguno de ellos le haría a usted daño / Ninguno de ellos le haría a usted daño **, nunca.**

[Adverbio de afirmación]

(26) a. Sí, simplemente poesía / Simplemente poesía, **sí.**
b. Sí, más allá de cualquier cosa estaba el río / Más allá de cualquier cosa estaba el río, **sí.**
c. También, la entrega resignada y el miedo a_dentelladas / La entrega resignada y el miedo a_dentelladas, **también.**

[Adverbio de duda]

(27) a. Quizá, nuestros genes - como una forma de preservar la especie - nos indiquen quienes son las mujeres que más salud tienen o cuales pueden llegar a ser más fértiles / nuestros genes […] nos indiquen quienes son las mujeres que más salud tienen o cuales pueden llegar a ser más fértiles**, quizá.**[9]
b. Quizá, pero llevo muchos años en esta profesión y puedo asegurar que ningún gesto es inocente / Pero llevo muchos años en esta profesión y puedo asegurar que ningún gesto es inocente**, quizá.**[10]

[8] En ocasiones la coma no es adecuada y la doble negación mejoraría alguna frase: "*En esta materia NO hay que dar pábulo al triunfalismo, nunca*".
[9] En esta ocasión *quizá* condiciona al subjuntivo "*nos indiquen*" luego su transformación no sería posible.
[10] En este caso la palabra "*quizá*" hace referencia a información fuera de la oración.

c. Probablemente, al verlo sería consciente de haber olvidado comprar las pastas para el té / Al verlo sería consciente de haber olvidado comprar las pastas para el té**, probablemente.**

d. Seguramente, habrá algún otro heredero muy interesado en que la chica se presente / Habrá algún otro heredero muy interesado en que la chica se presente**, seguramente.**

[Adverbio terminado en "mente"]

(28) a. Afortunadamente, las conversaciones sobre la mili no son frecuentes / Las conversaciones sobre la mili no son frecuentes**, afortunadamente.**
b. Personalmente, su apuesta y su partida me conmueven / Su apuesta y su partida me conmueven**, personalmente.**
c. Igualmente, notorios homosexuales han debido matrimoniar con íntima violencia para así salir en la Prensa y ser recordados por los productores / Notorios homosexuales han debido matrimoniar con íntima violencia para así salir en la Prensa y ser recordados por los productores**, igualmente. (?)**
d. Desgraciadamente, el autor del relato que comento tiene ideas rotundas y desafortunadas sobre las personas discapacitadas / El autor del relato que comento tiene ideas rotundas y desafortunadas sobre las personas discapacitadas**, desgraciadamente.**
e. Evidentemente, los tejanos (rotos o no) forman parte del atuendo grunge / Los tejanos (rotos o no) forman parte del atuendo grunge**, evidentemente.**
f. Efectivamente, a_lo_largo_de los años se salvó de diversos peligros / A_lo_largo_de los años se salvó de diversos peligros**, efectivamente.**
g. Ciertamente, Kasparov tenía argumentos para convencerse a sí mismo de que aquí no ha pasado nada / Kasparov tenía argumentos para convenverse a sí mismo de que aquí no ha pasado nada**, ciertamente.**[11]

[Ejemplos de coma más adverbio al final de frase]
[Adverbio de lugar]

(29) a. Todo iba hacia lo alto, aquí / **Aquí,** todo iba hacia lo alto.
b. En la Plaza_de_los_Vientos, arriba / **Arriba,** en la Plaza_de_los_Vientos.

[Adverbio de Modo]

(30) a. El resto de la familia, bien / **Bien,** el resto de la familia. (*)

[Adverbio de Cantidad]

[11] Los adverbios de evaluación parecen más adecuados para estos fines. Indican la actitud del hablante sobre un hecho determinado. Esta "opinión" puede ser desplazada e incluso suprimida.

(31) a. La procesión que vimos y otros detalles superficiales son un envoltorio, apenas / **Apenas,** la procesión que vimos y otros detalles superficiales son un envoltorio.[12] (*)

b. Nunca una queja delante mío, nada / **Nada,** nunca una queja delante de mío.

[Adverbio de Tiempo]

(32) a. Todos unidos tras de la ley, siempre / **Siempre,** todos unidos tras de la ley.

b. Él no quería, antes / **Antes,** él no quería.

[Adverbio de Negación]

(33) a. No me obsesionan sus pupilas dilatadas por el asombro al recibir el golpe, no / **No,** no me obsesionan sus pupilas dilatadas por el asombro al recibir el golpe.

b. Nadie quiso hablar, nunca / **Nunca,** nadie quiso hablar.

[Adverbio de Afirmación]

(34) a. Murieron por temerarias y por abusar del abordaje, claro / **Claro,** murieron por temerarias y por abusar del abordaje.

b. Darán mucho que hablar "Los Solidarios", sí / **Sí,** darán mucho que hablar “Los Solidarios”.

[Adverbio de Duda]

(35) a. Las heladas no lo encogen, quizá / **Quizá,** las heladas no lo encogen

b. Esta noche tendrá su primer sueño erótico, quizás / **Quizás,** esta noche tendrá su primer sueño erótico.

[Adverbios acabados en “mente”]

(36) a Se levantó y se fue caminando al borde del agua siguiendo la corriente con los ojos, obstinadamente / **Obstinadamente,** se levantó y se fue caminando al borde del agua siguiendo la corriente con los ojos.[13]

b. Se acerca la mano al pecho y se encaja mejor el sostén, inconscientemente / **Inconscientemente,** se acerca la mano al pecho y se encaja mejor el sostén.

Al igual que sucedía con la estructura adv+verbo = verbo+adv, el movimiento de la estructura adv+coma y coma+adv del principio al final de la frase y viceversa no sería productiva esteganográficamente de manera individual. La aceptabilidad lingüística de las transformaciones realizadas desplazando esta estructura, al principio y al final de frase, depende mucho del tipo de adverbio, de la longitud de la oración y de los complementos. Aunque de los ejemplos anteriores podría dar la impresión que esta estructura tiene una cierta libertad de movimiento, no obstante, parece que sólo, por los resultados actuales, los

[12] apenas y son tienen una fuerte dependencia, los complementos intermedios afectan al significado final.

[13] El significado de la frase cambia completamente, no obstante la frase se podría considerar como válida si esta modificación no afecta a la cohesión/coherencia del texto en el que se incluyan.

adverbios terminados en mente serían de utilidad real en una implementación práctica sin considerar un número elevado de excepciones a tener en cuenta. Estos adverbios de evaluación pueden ser considerados como útiles en esteganografía lingüística.

4. Conclusiones

Este trabajo continúa una nueva línea de investigación relacionada con el aprovechamiento de los avances en el procesamiento del lenguaje natural en lengua española con utilidad en la protección de comunicaciones digitales, anonimato y marcado digital de textos. El artículo profundiza en la posibilidad de conseguir procedimientos esteganográficos utilizando textos en lengua española mediante modificaciones sintácticas basadas en la reordenación de algunos complementos del verbo.

Dada 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 han seleccionado otras estructuras nuevas basadas en hipótesis lingüísticas. Se analiza concretamente dos estructuras basadas en el movimiento de los adverbios: a) patrón adverbio-verbo=verbo-adverbio y b) patrón adverbio+coma y coma+adverbio. El movimiento del adverbio respecto de los elementos del patrón seleccionado permitirá ocultar información. Para medir la presencia de estas estructuras se ha seleccionado dos corpora en los que se han realizado las medidas (LEXESP y WIKIPEDIA) que permiten observar en un corpus real la frecuencia de aparición de estos patrones y extraer ejemplos concretos para observar la aceptabilidad lingüística de la transformación en las estructuras propuestas.

Las medidas justificadas en el artículo indican que sería viable utilizar las estructuras analizadas con utilidad esteganográfica. No obstante, su baja ocurrencia hace que en la práctica, independientemente que no todos los ejemplos tengan aceptabilidad lingüística, sea difícil construir un software que las utilice de manera única para ocultar una cantidad razonable de 40 o más bits en un texto de tamaño medio. Queda para futuras investigaciones detectar nuevas estructuras lingüísticas que combinadas con las analizadas pudieran favorecer la construcción de un software real de marcado digital de textos basado en ocultación por modificaciones sintácticas. Es precisamente sobre esta implementación real y sobre textos de temática concreta donde podrían focalizarse nuevos ataques estadísticos y lingüísticos a la propuesta implementada.

Referencias bibliográficas

Bergmair, R. 2007. A comprehensive Bibliography of Linguistic Steganography. *International Conference on Security and Steganography*. [Disponible en http://www.semantilog.org/biblingsteg/]

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

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

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

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. *Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents*.

Muñoz, A., I. Argüelles y J. Carracedo. 2010. Modificaciones sintácticas en lengua española con utilidad en esteganografía lingüística. *RAEL* 8: 229-247.

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.

Shirali-Shahreza, M. y M.H. Shirali-Shahreza. 2007. Text Steganography in SMS. *Proceedings of the International Conference on Convergence Information Technology* (*ICCIT 2007*): 2260-2265.

Wikipedia. 2008. http://static.wikipedia.org/ June 2008 edition Wikipedia Static HTML Dumps.