Revista Signos. Estudios de Lingüística ISSN 0718-0934
© 2016 PUCV, Chile · DOI: 10.4067/S0718-09342016000100005 · 49(90) 94-118

Un corpus de bigramas utilizado como corrector ortográfico y gramatical destinado a hablantes nativos de español

A bigram corpus used as a grammar checker for Spanish native speakers

Alicia San Mateo

Universidad Nacional de Educación a Distancia
España

asanmateo@flog.uned.es

Recibido: 04-I-2014 / **Aceptado:** 19-VI-2015

Resumen

Este artículo describe el funcionamiento de un algoritmo de corrección ortográfica y gramatical para textos escritos en español, destinado a hablantes nativos competentes que realizan labores de corrección de textos. Los posibles errores se identifican por medio de análisis estadísticos (en vez de emplear el sistema de 'etiquetado' y análisis sintáctico que utiliza la mayor parte de correctores), comparando las combinaciones de palabras utilizadas con un corpus de referencia de cien millones de vocablos. De esa manera, se señalan los pares de palabras (bigramas) poco o muy poco frecuentes, y que, en muchas ocasiones, lo son porque contienen algún error. La limitación fundamental es que no se detectan errores que no puedan ser deducidos del análisis de palabras adyacentes. Pero, como hemos comprobado aquí en el análisis de diferentes textos, el algoritmo es capaz de localizar errores que otros correctores no identifican.

Palabras Clave: Bigramas, corrección de textos, corrector gramatical, detección de errores, pares de palabras.

Abstract

This paper describes the performance of an algorithm of spelling and grammar checker for texts written in Spanish by proficient native speakers during proof-reading. Possible mistakes are not detected by tagging and parsing but by statistical analysis, comparing combinations of two words used in the text to a hundred-million-word *corpus*. Those pairs of words (bigrams) which do not occur or which are suspiciously infrequent in the corpus are highlighted; and such pairs often contain errors. The main limitation is that mistakes that arise from non-adjacent words are not detected. Nevertheless, as we have seen in this study with some different texts, the algorithm detects many errors that other grammar checkers are not able to identify.

Key Words: Bigrams, error detection, grammar checker, pairs of words, proof-reading.

INTRODUCCIÓN

En este trabajo presentamos las posibilidades que, para la detección y corrección de errores ortográficos y gramaticales en textos escritos por nativos, ofrece el análisis estadístico de la frecuencia de las palabras y de los pares de palabras ('bigrama' o 'digrama'; en inglés, *bigram*) utilizados en un texto en comparación con un corpus de textos escritos en español, de cien millones de vocablos, compilado expresamente para desarrollar la herramienta CorrectMe (Universidad Nacional de Educación a Distancia), que implementa este método de análisis estadístico.

En los últimos años se han diseñado algoritmos basados en la estadística para detectar errores gramaticales, que han venido a complementar a los correctores automáticos que precisan del etiquetado y del análisis morfosintáctico previo para llevar a cabo su función. El funcionamiento de esos nuevos instrumentos es sencillo, como veremos en el apartado siguiente, y la mayor parte se ha creado para corregir textos escritos en inglés; si bien también hay una propuesta para el español utilizando la base de datos de Google Books. Nuestro objetivo es comprobar si este método basado en la estadística funciona también para detectar errores en textos extensos escritos en español por hablantes competentes en la lengua, y verificar su utilidad durante el proceso de corrección y edición. Para evaluarlo utilizaremos fragmentos de textos escritos por nativos, así como un texto extenso, el capítulo de un libro destinado a ser publicado por una editorial, y compararemos su eficacia con la de otros correctores automáticos.

El artículo está dividido en cinco apartados: en el primero presentamos diferentes correctores de textos que emplean análisis estadísticos; después, expondremos la metodología del estudio, que incluye la explicación del funcionamiento del corrector y los tipos de errores que encontramos frecuentemente en textos escritos por hablantes nativos, que son con los que trabajaremos. En el tercer apartado, incluimos el resultado de la evaluación en comparación con otros correctores; y en el cuarto, las limitaciones

que hemos observado. Finalmente, las conclusiones y algunas sugerencias que podrían hacer que la corrección fuera más precisa.

1. Correctores automáticos de textos basados en análisis estadísticos

En la bibliografía disponible, como veremos a continuación, hay numerosos informes positivos sobre el uso de bigramas, trigramas, etc. (*n-gram*) para detectar errores en textos escritos, sobre todo, en inglés. Este tipo de correctores estudia el contexto que rodea a cada palabra utilizando análisis estadísticos. Agrupan las palabras del texto de dos en dos, de tres en tres, etcétera, y comprueban si cada una de esas combinaciones aparece en un gran corpus de textos escritos; si se registra menos veces de lo que sería esperable, hay que preguntarse si la razón es que contiene un error. También hay correctores que combinan el etiquetado y la sintaxis con la información procedente del análisis de corpus. Estos son algunos de ellos.

ALEK (Assessing Lexical Knowledge) está pensado para corregir redacciones en inglés del examen TOEFL (Title of English as a Foreign Language); Chodorow y Leacock (2000: 140) explican así su funcionamiento:

> "A major objective of this research is to avoid the laborious and costly process of collecting errors (or negative evidence) for each word that we wish to evaluate. Instead, we train ALEK on a general corpus of English and on edited text containing example uses of the target word. The system identifies inappropriate usage based on differences between the word's local context cues in an essay and the models of context it has derived from the corpora of well-formed sentences."

Sjöbergh (2006) presenta un programa que puede ser utilizado con textos de cualquier lengua (en su caso, el sueco), siempre que se cuente con un corpus lo suficientemente grande que sirva de referencia y con un programa que segmente textos (*chunker*) escritos en la lengua en cuestión. Así lo describe:

> "Our proposed method is to run the chunker on a reference corpus with known correct text. This training step collects and saves a lot of statistics on what chunk sequences occur in normal language use. When a new text needs to be checked, simply run the chunker on the new text. If it contains chunk sequences that never occurred in the reference texts, these passages are marked as possible errors." (Sjöbergh, 2006: 180).

Chen (2009), por su parte, compara la eficacia de un corrector gramatical de inglés, basado en métodos estadísticos, diseñado por la National Taiwan Normal University (NTNU), con otro de Microsoft, el Microsoft ESL Assistant. Aquel funciona así:

> "The procedures used to detect the violations of English grammar in a statistical grammar checker are not complicated. The grammar

> checker system is first trained on a large corpus of edited texts, from which it extracts and counts bigrams that consist of sequences of adjacent words. British National Corpus (BNC) was used as the reference corpus for the NTNU checker." (Chen, 2009: 165).

Encontramos más información sobre esta clase de programas y sus aplicaciones en los siguientes trabajos. Algunos están pensados específicamente para identificar un tipo de errores concretos, como los relacionados con el uso de las preposiciones en inglés (Wu & Su, 2006); otros para corregir el orden de las palabras en la oración (Athanaselis, Bakamidis & Dologlou, 2006), bien como complemento de un traductor automático que produce errores que tienen que ver con el orden de las palabras en sus traducciones del birmano al inglés (Lin, Soe & Thein, 2011); otros correctores están diseñados para detectar errores en redacciones de estudiantes de segunda lengua (Briscoe, Medlock & Andersen, 2010; Yannakoudakis, 2013); otros están destinados a hablantes de inglés como L2 para que corrijan sus propios textos (Naber, 2003; Moré, 2006; Lawley & Martin, 2006; García-Heras, 2007; Gamon, Leacock, Brockett, Dolan, Gao, Belenko & Klementiev, 2009; Islam & Inkpen, 2011; Hernández García, 2012; Lawley 2015; entre otros). Estos trabajos dan cuenta del buen funcionamiento y eficacia de diferentes correctores que utilizan análisis estadísticos de palabras adyacentes y un corpus de referencia, pero estos programas están diseñados para corregir textos escritos fundamentalmente en inglés. Sobre el español contamos con el estudio de Nazar y Renau (2012), que utiliza Google Books como corpus de referencia para construir una base de datos de secuencias de hasta cinco palabras, que aparecen como mínimo 40 veces en el corpus (2012). Con el fin de evaluar la eficacia del algoritmo propuesto, los autores lo utilizan para completar ejercicios de un manual de gramática destinado a estudiantes de español y para detectar errores en textos producidos por aprendices de español. Los resultados indican que el método es capaz de señalar:

> "difficult mistakes such as **informes conteniendo* (instead of *informes que contenían* […]) or **máscaras antigases* (instead of *máscaras antigás* […]), which are errors that were not detected by MS Word" (Nazar & Renau, 2012: 32).

Es decir, que el análisis estadístico también resulta de utilidad para los textos redactados en español.

Nuestro propósito aquí es ampliar la idea de Nazar y Renau (2012) y emplear el análisis estadístico para detectar errores en el proceso de corrección y edición de

textos realizados por hablantes competentes de la lengua, en este caso, el español, ya sean correctores profesionales, periodistas, estudiantes, profesores, etc. La corrección de pruebas tiene sus propias particularidades: se trata normalmente de textos más extensos que las redacciones de un estudiante de español y, en general, la corrección automática no es tan necesaria, pues el hablante competente lo que requiere es ayuda para detectar el error y no tanto opciones para corregirlo. Como señalan Nazar y Renau (2012: 29), creemos que "a very simple grammar checker based on corpus statistics could prove to be helpful, at least as a complement to the standard procedures", y eso es lo que hace CorrectMe.

2. Metodología

El funcionamiento del corrector es sencillo: contrasta las combinaciones de palabras utilizadas en el texto con los datos de esas mismas palabras obtenidos en el corpus de textos y, dependiendo del resultado del algoritmo, nos avisa de si los bigramas son poco o nada frecuentes (y deberían serlo más, dada la frecuencia de las palabras que los forman), lo cual puede ser un indicio de que contienen un error. Para evaluar la eficacia del algoritmo, hemos recopilado textos escritos por hablantes nativos que incluyen errores que encontramos con cierta frecuencia –muchos de ellos están recogidos en la sección de 'preguntas frecuentes' de la página web de la Real Academia Española (http://www.rae.es/consultas-linguisticas/preguntas-frecuentes)–. En general, las palabras usadas erróneamente existen en español y son correctas en otros contextos –de ahí la dificultad de detectar sus usos incorrectos o, en muchos casos, impropios–.

Primero, vamos a explicar la composición del corpus de referencia; luego, el algoritmo y, finalmente, los tipos de errores que previsiblemente serán detectados.

2.1. Funcionamiento del corrector: Corpus y algoritmo

El corrector utiliza como material de referencia para identificar errores un corpus de cien millones de palabras, procedentes de textos escritos por nativos. Este corpus se ha compilado con el único propósito de facilitar la corrección de escritos mediante la detección de bigramas que normalmente no se combinan en la lengua; p. ej., la preposición 'a' y 'abierto', el participio del verbo 'abrir', en una secuencia como '*Mi amigo a abierto una librería en Granada'.

A diferencia de otros corpus como, por ejemplo, el Corpus del español actual (CREA), en este no se han incluido transcripciones de textos orales, ya que estos contienen autocorrecciones, interrupciones, incisos, repeticiones, etc., que no son modelos apropiados para la mayor parte de textos escritos. Para seleccionar el material del corpus se ha tenido en cuenta, en primer lugar, el criterio cronológico, ya que es un corpus sincrónico y se pretende recoger el uso actual de la lengua. Está formado por extractos de textos narrativos, es decir, novelas, cuentos y ensayos, escritos en

español (no traducidos) desde 1980 hasta 2012, y disponibles en formato electrónico; además de noticias y artículos de opinión publicados en periódicos nacionales y revistas de actualidad durante el año 2012 principalmente; y no contiene poemas, obras de teatro ni novelas dialogadas, con el fin de que no aparezcan expresiones propias de la oralidad que se alejan de la lengua escrita. El empleo de múltiples fragmentos de diversas obras permite incluir en el corpus una gama amplia de textos y evitar la presencia excesiva de textos idiosincrásicos. En definitiva, el objetivo fue construir un corpus de textos sobre temas generales y no excesivamente técnicos, tanto de ficción como de no ficción, que reflejaran el español contemporáneo. El equipo que lo ha creado es consciente de que un corpus de cien millones de palabras no ofrece demasiada información sobre el comportamiento de términos científicos, técnicos o poco habituales como por ejemplo, 'cizalla', que no aparece nunca, o como 'ornitorrinco', cuya frecuencia es seis; pero sí posee suficiente información para detectar anomalías en el comportamiento de vocablos relativamente frecuentes en el idioma, tal y como veremos a continuación.

En cuanto a la procedencia geográfica, los textos del corpus provienen de los diferentes países de habla española, si bien se han elegido aquellos desprovistos de las peculiaridades que aportan las variaciones diatópicas; es decir, el objetivo ha sido evitar localismos y otros particularismos o divergencias, procedentes también de variedades diastráticas concretas, para que el corpus contenga muestras de lengua de nivel formal, que identificamos como lengua estándar:

> "Es por ello la expresión culta formal la que constituye el español estándar: la lengua que todos empleamos, o aspiramos a emplear, cuando sentimos la necesidad de expresarnos con corrección; la lengua que se enseña en las escuelas; la que, con mayor o menor acierto, utilizamos al hablar en público o emplean los medios de comunicación; la lengua de los ensayos y de los libros científicos y técnicos. Es, en definitiva, la que configura la norma, el código compartido que hace posible que hispanohablantes de muy distintas procedencias se entiendan sin dificultad y se reconozcan miembros de una misma comunidad lingüística" (RAE, 2005: XIV-XV).

Así, por ejemplo, no se incluyen textos representativos de variedades geográficas como el extremeño, el andaluz, el canario o el murciano. De todos modos, aunque un texto contenga algún uso atípico o incluso alguna errata –a pesar de que han sido redactados por periodistas y escritores profesionales y han sido convenientemente corregidos y editados, siempre se puede haber deslizado alguna (igual que Nazar & Renau, 2012: 27), asumimos que los textos "go through different phases of revision and correction with high standars" y que, por lo tanto, "can be used as a reference corpus for inferring the grammar rules of a language")–, su efecto no será estadísticamente relevante dado el tamaño del corpus. Por ejemplo, si escribimos 'el yo', a pesar de que esta combinación se da en el corpus porque este incluye alguna obra sobre psicología freudiana, el número de veces que aparece es tan bajo que el corrector nos alertará

de que quizá hayamos cometido un error. También es importante saber que el corpus solo se usa para extraer la información sobre la frecuencia de las palabras que en él se emplean, pero que en el proceso de corrección de un texto no se accede a los documentos completos.

El algoritmo que se utiliza aquí para detectar errores no es nuevo (véase Sinclair, 1991). Por un lado, se analiza la frecuencia en el corpus de cada una de las palabras utilizadas en el texto y, por otra parte, de cada bigrama. Además se calcula el número de veces que cada combinación de dos palabras aparecería en el corpus si estas se dieran de manera aleatoria; es decir, se estima la probabilidad del par teniendo en cuenta la frecuencia de cada una de las dos palabras, por separado, en el corpus, tal y como refleja la fórmula siguiente, donde *P* es la probabilidad; (a) y (b) representan, respectivamente, a la palabra 1 y a la palabra 2; *T* es el número total de palabras que componen el corpus (cien millones) y *F*, la frecuencia. Conviene tener en cuenta que, como veremos a continuación, la escala de la probabilidad empleada no es la de 0 a 1.

$$P(ab) = \frac{T}{\left(\frac{T}{F(a)}\right) \times \left(\frac{T}{F(b)}\right)}$$

Sinclair (1991) explica el sentido de utilizar el concepto de 'probabilidad de la combinación de palabras' con el ejemplo del verbo *set*, seguido de la preposición *off*. La frecuencia del verbo en un corpus de 7,3 millones de palabras es 1.855; por lo que su probabilidad de aparecer al azar es de 250 por millón de palabras, o lo que es lo mismo, *set* sería una de cada 3.935 palabras. Por otro lado, la probabilidad de aparición de la preposición *off* es 0.00055. Así que la probabilidad de que *off* aparezca tras *set* es 0.00025 por 0.00055, cuyo resultado es 0.0000001375; y en un corpus de 7,3 millones de palabras distribuidas al azar, "we might expect 0.0000001375 x 7300000 occurrences of *set off*, that is, one only" (Sinclair, 1991: 70). El autor recuerda que:

> "The assumption behind this calculation is that the words are distributed at random in a text. It is obvious to a linguist that this is not so, and a rough measure of how much *set* and *off* attract each other is to compare the probability with what actually happens".

Y este es precisamente el siguiente paso: el par *set off* aparece casi 70 veces en el corpus de 7,3 millones de palabras "as against the random prediction of only one occurrence. The 70 instances give us enough evidence of the main patterning" (Sinclair, 1991: 70).

Volviendo al algoritmo, tras aplicar la fórmula anterior, se analiza si el par de palabras aparece en el corpus más (o menos) veces de lo que sería esperable según su probabilidad –es decir, se calcula el umbral (*U*)– mediante esta fórmula: U = F(ab) / P(ab).

A modo de ejemplo, analizaremos esta oración, procedente de una noticia del periódico *El Mundo* (http://www.elmundo.es/elmundo/2013/10/28/gentes/1382945093.html):

(1) […] *y su tiene intención de visitar la tumba del que fue su amigo […].

En ella 'su' debería ser 'si'; este error se debe a la cercanía en el teclado de las letras u e i. El corrector ortográfico del procesador de textos no detecta la confusión porque 'su', aunque es incorrecto en este contexto –donde debería utilizarse la conjunción condicional 'si'–, es una palabra existente en español.

Para aplicar el algoritmo, se comprueba primero la frecuencia de cada palabra por separado y se confirma que todas están en el corpus (si no fuera así, se señalarían); después se segmenta la oración en estas doce combinaciones: 1. y su, 2. su tiene, 3. tiene intención, 4. intención de, 5. de visitar, 6. visitar la, 7. la tumba, 8. tumba del, 9. del que, 10. que fue, 11. fue su, y 12. su amigo.

El resultado, en primer lugar, mostrará gráficamente las combinaciones poco y muy poco frecuentes y tendremos los datos estadísticos sobre esas combinaciones. En este caso se indica que 'su tiene' es un par muy poco frecuente. Después, podremos obtener los datos que recogemos en la Tabla 1 sobre cada uno de los pares de palabras adyacentes del texto, los frecuentes y los no frecuentes. Trabajaremos ahora con todos ellos para compararlos y ver las diferencias.

Tabla 1. Análisis de las combinaciones de (1).

Frecuencia combinación	Frecuencia palabra 1	Frecuencia palabra 2	Probabilidad combinación	Umbral
y su=34793	y= 2813605	su= 937618	26380.87	1.32
su tiene=0	su= 937618	tiene= 69592	652.51	0
tiene intención=131	tiene= 69592	intención= 9245	6.43	20.37
intención de=5176	intención= 9245	de= 5012014	463.36	11.17
de visitar=407	de= 5012014	visitar= 3514	176.12	2.31
visitar la=287	visitar= 3514	la= 3333903	117.15	2.45
la tumba=2403	la= 3333903	tumba= 4086	136.22	17.64
tumba del=121	tumba= 4086	del= 752923	30.76	3.93
del que=7091	del= 752923	que= 3295376	24811.64	0.29
que fue=7686	que= 3295376	fue= 122589	4039.77	1.9
fue su=1310	fue= 122589	su= 937618	1149.42	1.14
su amigo=4911	su= 937618	amigo= 28896	270.93	18.13

En la primera columna de la Tabla 1, aparece reflejada la frecuencia del par; así, por ejemplo, la combinación 'intención' de se registra 5.176 veces en este corpus. Después, en las dos columnas siguientes, tenemos la frecuencia de cada una de las dos palabras por separado: 'intención' se usa en el corpus más de nueve mil veces (9.245) y 'de',

más de cinco millones (5012014); o lo que es lo mismo, una de cada 10817 palabras del corpus (100000000/9245 = 10817) es el sustantivo 'intención' y una de cada 20, la preposición 'de' (100000000/5012014 = 20). La cuarta columna incluye la probabilidad de que esta combinación de palabras, teniendo en cuenta la frecuencia de cada una de ellas por separado (10817*20 = 215815), si todas las palabras apareciesen al azar en el corpus (recuérdese aquí lo que decía Sinclair, 1991: 79): 'intención de' se registraría 463.36 veces en el banco de cien millones de palabras (100000000/215815 = 463.36). Sin embargo, en el corpus encontramos el bigrama 'intención de' un número de veces que es 11.17 mayor que la probabilidad de combinarse al azar (5176/463.36 = 11.17). Esta cifra es el umbral e indica el grado de atracción que poseen dos palabras; es decir, si los nativos suelen utilizarlas juntas o no. Cuanto mayor es el umbral, mayor es el grado de atracción que demuestran las palabras. Solo si el umbral es mayor de uno, podemos inferir que dos palabras se atraen; es decir, tienden a usarse juntas. Si es menor de uno, significa que esos dos vocablos tienden a rechazarse entre sí.

La frecuencia real del par y el umbral (5.176 y 11.17, respectivamente, en el caso de 'intención de') son dos buenos indicativos de la corrección o incorrección de una combinación. Lógicamente, si los hablantes han utilizado un par más de cinco mil veces, podemos estar bastante seguros de su corrección, sobre todo si esta cifra (5.176) es más alta de lo que se esperaría si las palabras en los textos apareciesen al azar (463.36).

Ahora bien, si nos fijamos en las cifras del par 'su tiene' (véase Tabla 1), veremos que la frecuencia es cero, lo que quiere decir que en el corpus de cien millones de palabras no aparece nunca. Una de cada 107 palabras del corpus es el posesivo 'su' (100000000/937618 = 107) y una de cada 1.437 es la forma verbal 'tiene' (100000000/69592 = 1437); de ahí que la probabilidad del par de aparecer al azar sea 652.51 (107*1437 = 153255; 100000000/153255 = 652.51); pero, como la frecuencia real del par es cero, el umbral también es cero (0/652.51 = 0). Estas cifras nos indican que, a pesar de que la probabilidad de combinación de este par de palabras es bastante alta (652.51), en el español correcto no se usan nunca juntas; por lo que podemos deducir que tienden a rechazarse entre sí. Gracias a la información estadística extraída del corpus, detectamos errores que ni el corrector ortográfico y gramatical del procesador de textos de Microsoft Word ni otros correctores como Stilus® (Villena, González, González & Muriel, 2002) o SpanishChecker® (Nadasdi & Sinclair, 2001-2015) identifican como tales. Si bien es cierto que este último recomienda verificar la combinación 'su tiene', el comentario que ofrece podría desorientar al usuario, sobre todo si es un estudiante de L2, pues dice: "Debería verificar. La palabra 'su' es el posesivo *'his/her'*. Si quiere decir *'she',* debería escribir 'ella'; si quiere decir *'he'*, debería escribir 'él' (SpanishChecker®, en línea).

No obstante, en la recomendación no se menciona nada sobre la posibilidad de utilizar la conjunción condicional 'si'. Los otros dos correctores no señalan de ninguna forma que podría haber un error en el par de palabras 'su tiene'.

2.2. Tipos de errores

Para ver si el algoritmo es eficaz detectando errores, vamos a centrarnos en los cinco casos que presentamos a continuación, y vamos a probar primero con fragmentos y, después, con un texto extenso de 9.000 palabras, que contienen algún error.

1. Errores gramaticales u ortográficos que el corrector del procesador de textos no detecta; por ejemplo, la tilde en la conjunción 'o'.

2. Problemas de paronimia o confusiones entre palabras correctas –es decir, que existen en la lengua–, que son homófonas, tales como 'aya/haya/halla', 'haber/a ver'; o que una de ellas lleva tilde diacrítica como, por ejemplo, 'dé/de', 'él/el', 'más/mas'; o entre dos palabras cuya sílaba tónica no es la misma, como 'acorde/acordé, 'articulo/artículo', 'deseo/deseó'.

3. Omisión de alguna palabra (o inclusión donde no corresponde) como, por ejemplo, la preposición 'a' en *ayudar cocinar, *empezar estudiar o *ir comer; o la preposición 'de' en *terminar hacer.

4. Erratas provocadas por la rapidez de la escritura en el teclado del ordenador, como, por ejemplo, el cambio de orden de las letras ('al/la', 'el/le', 'otro/toro', 'pato/apto'); o el uso de una letra cuya tecla está al lado de la correcta ('conde/donde', 'lana/lama', 'no/ni'); o por la confusión entre dos letras: 'boda/bola', 'cadenas/caderas', 'cara/capa', 'casa/cada'.

5. Olvido e inclusión de alguna letra (o palabra) de forma incorrecta, como, por ejemplo, en *él dices, *tú va, 'ambas/amas', 'consejo/conejo', 'había/habían'.

En la tipología de errores que establece Díaz Villa (2005), la causa de los cinco anteriores es cognitiva; solo la omisión de un elemento se debe a la falta de atención (error fortuito). De una manera o de otra, en todos los casos se trata de errores frecuentes cuya detección no siempre resulta fácil para el corrector de textos, pero que, con una simple llamada de atención, el hablante competente corregiría fácilmente.

3. Resultados y evaluación

Los resultados de la detección de los errores incluidos en las oraciones realizada con tres correctores y con CorrectMe se resumen en la Tabla 7. Veremos a continuación los ejemplos concretos y la respuesta obtenida en cada caso. Téngase presente que, a menos que se indique lo contrario, el corrector de Microsoft Word no detecta los errores. No obstante, iremos aportando los resultados que ofrecen el corrector del procesador de textos, Stilus® y SpanishChecker® –estos últimos están destinados específicamente al hablante de español no nativo–, al analizar cada uno de los fragmentos. Después de estudiar los resultados del análisis de los fragmentos, pasaremos a los de un texto extenso de 9.000 palabras.

3.1. Errores gramaticales u ortográficos

En ocasiones, en textos publicados, encontramos ejemplos como los siguientes: Fundación Vicente Ferrer: "*Hay bienes para erradicar la pobreza tres ó cuatro veces" (http://www.diariodeavisos.com/2013/06/fundacion-vicente-ferrer-hay-bienes-para-erradicar-pobreza-tres-o-cuatro-veces/), o Ramos: "*Hemos tenido cuatro ó cinco ocasiones pero no las hemos aprovechado" (http://latribunamadridista.com/ramos-hemos-tenido-cuatro-o-cinco-ocasiones-pero-no-las-hemos-aprovechado/), en los que la conjunción disyuntiva 'o' aparece con tilde. A pesar de que, como se recoge en el Diccionario panhispánico de dudas (s. v. o[2], § 3), tradicionalmente se recomendaba tildar la *o* cuando iba entre números para distinguirla del cero –y no confundir, por ejemplo, '3 ó 4' con '304'–; sin embargo, nunca debe llevar tilde cuando va entre un número y una palabra o entre dos palabras, como en los ejemplos propuestos. Actualmente la RAE no justifica la tilde en esta conjunción en ningún contexto (RAE 2010: 270) e incluye este asunto en la sección de 'preguntas frecuentes': 'La conjunción o siempre sin tilde, incluso entre cifras'. El corrector de Microsoft Word no señala error alguno en oraciones comoz.

(2) *¿Quieres té ó café?

(3) *Hemos tenido cuatro ó cinco ocasiones pero no las hemos aprovechado;

en cambio, al aplicar el algoritmo tendremos evidencias claras de que en estas combinaciones de palabras hay un error (véase Tabla 2):

Tabla 2. Análisis de las combinaciones menos frecuentes de (2-3).

Frecuencia combinación	Frecuencia palabra 1	Frecuencia palabra 2	Probabilidad combinación	Umbral
té ó=0	té= 4033	ó= 530	0.02	0
ó café=0	ó= 530	café= 7548	0.04	0
cuatro ó=0	cuatro= 28452	ó= 530	0.15	0
ó cinco=2	ó= 530	cinco= 23204	0.12	16.67

La frecuencia de los pares, como vemos en los datos anteriores, no supera los dos casos, y en tres de ellos es cero, al igual que el umbral. Si tenemos en cuenta que no se trata de tecnicismos, la baja frecuencia es un claro indicativo de que la combinación es incorrecta. En el caso del último par ('ó cinco') al aparecer dos veces en el corpus y ser la probabilidad muy baja (0.12), el umbral es superior a cero; si bien, gracias a las cifras del par anterior ('cuatro ó') no será difícil detectar el error. Como hemos dicho antes, la conjunción 'o' nunca ha llevado tilde en ese contexto (entre dos cifras escritas con letra); por lo que podemos concluir que esos 530 casos registrados en el corpus son errores.

En este caso, Stilus® identifica el error en (2-3); pero SpanishChecker® indica, refiriéndose a la conjunción, que "No es muy usual que una letra esté sola así.

¿Está seguro que sea su intención eso (sic)?"; lo cual puede confundir al usuario y, además, no dice nada acerca del error cometido.

Por otro lado, encontramos frecuentemente errores relacionados con la concordancia de los determinantes y adjetivos que acompañan a los sustantivos femeninos que comienzan por /a/ (gráficamente 'a'- o 'ha'-) tónica, como 'aula', 'agua', 'arma', etc. ("error de concordancia con femeninos débiles" (Díaz Villa, 2005: 415). Ante estos nombres, debemos usar la forma 'el' del artículo, pero solo cuando va inmediatamente antes; si entre el artículo y el sustantivo se interpone un elemento, como en (4), un adjetivo ('dichoso'), entonces se utiliza la forma 'la' (*DPD*, s. v. *el*, § 2.1). Si al sustantivo lo precede algún indefinido, es posible emplear cualquiera de las dos formas: 'un' o 'una', 'algún' o 'alguna', etc. Los demás adjetivos determinativos deben aparecer siempre en su forma femenina ('esta', 'esa', 'aquella'…).

(4) *Ya dentro de la particular situación, y de la alarma de los agentes por el dichoso arma […] (http://www.elmundo.es/madrid/2013/11/21/528e77a661fd3d6a758b4587.html).

Cuando aplicamos el algoritmo comprobamos que la combinación 'dichoso arma' es muy poco frecuente (véase Tabla 3); sin embargo, en el bigrama 'el dichoso' no encontraremos nada anómalo, como es lógico, pues la concordancia es correcta.

Tabla 3. Análisis de las combinaciones menos frecuentes de (4).

Frecuencia combinación	Frecuencia palabra 1	Frecuencia palabra 2	Probabilidad combinación	Umbral
el dichoso=89	el= 2412037	dichoso= 964	23.25	3.83
dichoso arma=0	dichoso= 964	arma= 7140	0.07	0

Ni el corrector de Microsoft Word ni Stilus® identifican el error de (4). En cambio, SpanishChecker® señala el par 'dichoso arma' y aconseja verificar si la concordancia de género entre el adjetivo y el sustantivo es correcta, lo cual da una pista acerca del error cometido.

3.2. Confusión entre palabras correctas

El algoritmo detecta muchos casos de confusión entre dos homófonos. Es frecuente confundir el infinitivo del verbo 'haber' y la combinación de la preposición 'a' seguida del infinitivo del verbo 'ver' (caso que recoge la RAE entre sus 'preguntas frecuentes'), como vemos en estos ejemplos:

(5) *Vete haber qué nota te han puesto

(6) *Tiene que a ver sucedido algo

Nótese que en (6) el infinitivo 'ver' va seguido de un participio ('sucedido'), lo cual es una combinación imposible es español.

Tabla 4. Análisis de las combinaciones menos frecuentes de (5-6).

Frecuencia combinación	Frecuencia palabra 1	Frecuencia palabra 2	Probabilidad combinación	Umbral
Vete haber=0	Vete= 2665	haber= 53686	1.43	0
haber qué=0	haber= 53686	qué= 263847	141.65	0
ver sucedido=0	ver= 77172	sucedido= 5839	4.51	0

Como vemos en la Tabla 4, estas combinaciones no se registran ni una sola vez en el corpus de cien millones de palabras, a pesar de que la probabilidad de aparecer juntas es, en todos los casos, superior a 1.4 (incluso en 'haber qué' esa probabilidad es bastante alta: 141.65). Esos son claros indicios de que contienen errores. En cambio, sí se encuentra en el corpus esta otra combinación ('haber si'), pero solo siete veces; mientras que la probabilidad de que se dé este par de palabras es mucho mayor: 183.29 (por separado, la frecuencia de 'haber' es 53686 y la de 'si', 341403); de ahí que el umbral sea tan bajo: 0.04 y que nos deba hacer pensar en un error:

(7) *Dice que haber si le sacan.

Solo Stilus® identifica y corrige correctamente el error de (7). El corrector de Word no señala ningún error en (5-7), y tanto Stilus® como SpanishChecker® plantean una corrección equivocada en dos casos en los que detectan una falta: el primero, en (5), plantea incluir la preposición *a*, lo cual resultaría en una combinación incorrecta (*vete a haber), y SpanishChecker®, en (6), propone eliminar la preposición 'a' ('tiene que ver 'no es lo mismo que 'tiene que haber', que es lo que se quiere decir).

Por otro lado, en ocasiones se omite la tilde diacrítica o se coloca en la palabra que no la lleva. Este es un ejemplo del primer error:

(8) *Quienes beban te o café con moderación no tienen por qué preocuparse por el consumo de cafeína […] (http://www.alimentacion-sana.org/informaciones/novedades/cafeina2.htm)

La palabra 'té', la infusión, lleva tilde; en cambio, en (8) se ha omitido, con lo cual se ha convertido en el pronombre de 2.ª persona de singular. El análisis de las cifras (la frecuencia del par es cero; por separado, la de 'beban' es 47 y la de 'te', 175840; la probabilidad del bigrama es 0.08 y el umbral, cero) nos advierte de la anomalía existente en 'beban te', que no se registra en el corpus ni una sola vez.

En el siguiente ejemplo tenemos varios errores, entre ellos uno relacionado con la asignación errónea de la sílaba tónica ('ingles-inglés'):

(9) *Playa del ingles abrio sus puertas en 1984 y desde ahi es uno de los mas populares Disco-Pubs del sur de Gran Canaria (http://www.pachaplayadelingles.com/pacha/es/index.php)

El corrector de Microsoft Word indica que se ha omitido la tilde en 'abrió' y en 'ahí' (al igual que Stilus® y SpanishChecker®); sin embargo, no señala el adverbio 'mas', pues existe la forma sin tilde, que es una conjunción adversativa; ni tampoco el adjetivo 'inglés', ya que el sustantivo 'ingles' también figura en el léxico del español –como indican Ariza y Tapia (1997-1998), el corrector no marca como error las formas que figuran en los diccionarios que utiliza como referencia– (en cambio Stilus® y SpanishChecker® sí que detectan estos dos errores). Además, este corrector subraya la forma plural de *pub*, a pesar de que, al ser una voz procedente de otra lengua y terminar en be, la regla es añadirle -s (*DPD*, s. v. *plural*, § 1. h).

Tabla 5. Análisis de las combinaciones menos frecuentes de (9).

Frecuencia combinación	Frecuencia palabra 1	Frecuencia palabra 2	Probabilidad combinación	Umbral
del ingles=0	del= 752923	ingles= 172	1.3	0
ingles abrio=0	ingles= 172	abrio= 1	0	0
abrio sus=0	abrio= 1	sus= 370971	0	0
desde ahi=0	desde= 97874	ahi= 54	0.05	0
ahi es=0	ahi= 54	es= 544084	0.29	0
mas populares=0	mas= 16546	populares= 950	0.16	0

Si nos fijamos en los resultados (véase Tabla 5) comprobamos que la frecuencia de todas las combinaciones que contienen errores es cero; con lo cual, queda demostrada la eficacia del algoritmo basado en el análisis estadístico: está por encima de la del corrector del procesador.

3.3. Omisión o inclusión de palabras

En ocasiones, la rapidez y la falta de atención nos llevan a omitir alguna palabra mientras escribimos o a incluirla donde no corresponde; las preposiciones son las más olvidadas. Este es un ejemplo en el que se ha suprimido la preposición *a* tras el verbo 'empezar':

(10) […] *para quien quiere empezar entrenar […] (http://www.elenamalova.com/2013/10/ejercicios-para-obesos-2-exercise-for.html)

CorrectMe, a diferencia del corrector del procesador de textos y de Stilus®, detecta la baja frecuencia de esta combinación ('empezar estrenar' no se registra en el corpus; la probabilidad de aparición del par es 0.01 y el umbral, cero). Por su parte, SpanishChecker® también señala la omisión de la preposición 'a'.

3.4. Confusiones entre letras

Otro caso de confusión es la que se da entre 'tubo' y 'tuvo', bien por la cercanía en el teclado de las dos letras (la be y la uve) o bien porque son dos palabras homófonas; como, por ejemplo, en:

(11) *No tubo suerte (http://www.corazonblanco.com/no_tubo_suerte-fotos_del_real_madrid-igfpo-642766.htm)

Tabla 6. Análisis de las combinaciones menos frecuentes de (11).

Frecuencia combinación	Frecuencia palabra 1	Frecuencia palabra 2	Probabilidad combinación	Umbral
no tubo=0	no= 1435889	tubo= 2110	30.3	0
tubo suerte=0	tubo= 2110	suerte= 15238	0.32	0

En este caso, también el corrector del procesador de textos nos alerta del error. Como vemos en la Tabla 6, la frecuencia y el umbral de los dos pares ('no tubo' y 'tubo suerte') es cero. Aquí, al igual que en otros ejemplos como (5), incluso el programa identifica un error en dos combinaciones contiguas, lo que nos debe llevar a pensar que el problema está precisamente en la palabra que comparten los dos bigramas. Sin embargo, ni SpanishChecker® ni Stilus® indican que haya error alguno en (11).

3.5. Omisión e inclusión de letras incorrectas

La omisión y la inclusión de letras en el lugar equivocado dan lugar a no pocas erratas. Las combinaciones resultantes no siempre, afortunadamente, forman pares frecuentes y, gracias a ello, tras aplicar el algoritmo, se pondrá de manifiesto su baja frecuencia. Por ejemplo, en (12) se ha confundido 'consejo' con 'conejo':

(12) *¿Un buen conejo para dejar de fumar?

(http://espanol.answers.yahoo.com/question/index?qid=20130208160646AAcVIP3)

En el análisis observamos la baja frecuencia de este par: 'buen conejo' no se registra en el corpus (la probabilidad de aparición del par es 0.27 y el umbral, cero); un primer indicio de error. Si bien la combinación 'buen conejo' será correcta en determinados contextos, el hecho de que no esté registrada en el corpus es lo que ayuda a descubrir que se ha producido una confusión entre dos palabras. Ni el corrector de Microsoft Word, ni Stilus® ni SpanishChecker® la identifican.

Hasta ahora el análisis estadístico de la frecuencia ha demostrado su eficacia detectando errores en combinaciones de palabras dado que no se registran en el corpus de referencia; de manera que CorrectMe es más que un corrector ortográfico

y gramatical, ya que consigue identificar faltas relacionadas con el contexto en el que se insertan las palabras. Y, como vemos en la Tabla 7, supera con creces el número de errores descubiertos por el corrector del procesador de textos y por los correctores SpanishChecker® y Stilus®.

Tabla 7. Resumen del número de errores detectados por los diferentes correctores (*Entre paréntesis se indican los ejemplos en los que aparecen los errores. **Número total de errores incluidos en los ejemplos).

Tipo de error*	Microsoft Word	SpanishChecker®	Stilus®	CorrectMe
1. Errores gramaticales u ortográficos (2-4)	0/3** (0%)	1/3 (33%)	2/3 (67%)	3/3 (100%)
2. Problemas de paronimia (5-9)	0/6 (0%)	3/6 (50%)	3/6 (50%)	6/6 (100%)
3. Omisión de palabras (10)	0/1 (0%)	1/1 (100%)	0/1 (0%)	1/1 (100%)
4. Confusión entre letras (11)	1/1 (100%)	0/1 (0%)	0/1 (0%)	1/1 (100%)
5. Omisión/inclusión de letras (12)	0/1 (0%)	0/1 (0%)	0/1 (0%)	1/1 (100%)
Promedio	**25%**	**37%**	**23%**	**100%**

El corrector de Microsoft Word y Stilus® son los que menos errores han identificado (un 25% y un 23%, respectivamente). SpanishChecker® ha sido algo más eficaz, pues ha señalado el 37% de los errores; pero, como hemos comprobado, no detecta fallos cuando se han producido confusiones, omisiones o inclusiones de letras que dan lugar a palabras existentes en la lengua. Con todo, los resultados quedan muy lejos de los de CorrectMe.

Hasta ahora nos hemos limitado a probar el algoritmo con fragmentos breves, pero quizá la mayor utilidad es la que proporciona al corrector de un texto extenso si lo ayuda a localizar aquellos errores que han podido pasar desapercibidos en una primera lectura. Para comprobar la eficacia en estos casos, hemos analizado el capítulo de un libro destinado a ser incluido en una publicación universitaria, y que consta de 9.000 palabras. Estos son los fragmentos del capítulo en los que se han encontrado bigramas muy poco frecuentes (se indican subrayados) y todos contienen algún error, ya sea ortográfico, como, por ejemplo, la omisión de la tilde en 'módulo' y 'diálogo':

(13) […] *deberán tomar café en el mismo modulo.

(14) […] *pueden regresar a su modulo de origen para prepararse para la comida.

(15) […] *asunción de las propias responsabilidades frente al grupo y a la comunidad, hábitos de participación y dialogo.

O bien un error gramatical, como la falta de concordancia que se produce al escribir 'tiene' en lugar de 'tienen' (§ 3.1):

(16) *Los primeros tiene como fin la consecución de los objetivos de la organización […].

O se ha confundido una palabra por otra; p. ej., al utilizar *porque* en vez de *por qué* (§ 3.2):

(17) […] *tiene sus propias estructuras de liderazgo que no tienen porque corresponder con las de la organización formal.

O bien se ha incluido una palabra en la posición equivocada; por ejemplo, se dice 'grupo de social' en lugar de 'grupo social' (§ 3.3):

(18) […] *tanto para caracterizar la conducta en sí misma como para captar el valor de significación que tal conducta toma en el grupo de social de referencia.

O se ha omitido una letra dando lugar a otra palabra también existente en español; por ejemplo, se usa 'platean' en vez de 'plantean'; 'mimas' por 'mismas' (§ 3.5):

(19) *Para la consecución de estos objetivos de carácter general se platean estructuras […].

(20) *Se seguirán las mimas normas que para la comida […].

El hecho de que la probabilidad de que estas combinaciones se den en español sea mayor que la frecuencia real de las mismas es lo que nos hace sospechar que no son correctas, y este es el dato que nos proporciona el algoritmo que emplea CorrectMe. Por otro lado, en la propuesta de Nazar y Renau (2012), en la que se utiliza como base de datos el corpus Google Books *N*-gram, se registran todas las combinaciones de palabras cuya frecuencia es igual o mayor de 40 y su objetivo es "detect any sequence of words that cannot be found in the *n*-gram data base" (Nazar & Renau, 2012: 28). En este corpus de Gooble Books, aparecen las secuencias de (15-20) –es decir: 'y dialogo', 'primeros tiene', 'tienen porque', 'de social', 'se platean' y 'las mimas'–, con lo cual no serían susceptibles de ser detectados los errores, y no sería de gran utilidad para la lectura y corrección de pruebas. En este sentido, creemos que el empleo del algoritmo propuesto aquí supone un avance en la detección de combinaciones erróneas. Además, también evita que palabras poco frecuentes –por ejemplo, determinados nombres propios o tecnicismos– se señalen como errores, puesto que la probabilidad de aparición y su frecuencia real serán igualmente bajas.

En concreto, en este texto de 9.000 palabras, en el que hay 8.100 bigramas, si el algoritmo llama la atención sobre los 100 pares con las menores puntuaciones, que son los que más probablemente pueden contener un error –y de hecho, ocho son incorrectos– será muy útil; sobre todo, porque se trata de errores que otros correctores no identifican como tales. Por ejemplo, de esos ocho errores, los correctores destinados específicamente al aprendiz de español como L2, identifican dos cada uno. Grammar

Checker®, por un lado, señala (15 y 16); y Stilus®, por su parte (13 y 16). De igual forma, el corrector de Microsoft Word identifica dos de los ocho errores: (14 y 17).

Ahora bien, para realizar una evaluación más completa de los resultados obtenidos con la aplicación del algoritmo que presentamos aquí, vamos a compararlos con los del corrector de Microsoft Word, pues, en principio, ambos están destinados al hablante nativo. El algoritmo basado en el análisis estadístico de frecuencias indica ocho bigramas de los nueve que contienen errores en el texto de 9.000 palabras, pero además resalta 92 combinaciones más que son correctas –son falsos positivos porque los datos nos llevan a pensar que se ha cometido una falta cuando en realidad la secuencia es correcta–. Por su parte, Microsoft Word señala los dos errores que recogimos antes (14 y 17) y asimismo esta falta de concordancia:

(21) *La mayoría de las imágenes de la realidad sobre la que basamos nuestras acciones están realmente basados en la experiencia vicaria.

Y, por otro lado, señala seis términos, que son correctos, pero que no están incluidos en su diccionario; además de nombres propios y extranjerismos crudos. En total, los casos de falsos positivos suman 33.

Tabla 8. Resultados obtenidos con CorrectMe y Microsoft Word.

	Errores detectados	Falsos negativos	Falsos positivos	Precisión	Cobertura	Media armónica (F_1)
CorrectMe	8	1	92	8%	88.89%	14.68%
Microsoft Word	3	6	33	8.33%	33.33%	13.33%

Con estos datos podemos calcular el rendimiento de ambos sistemas de recuperación de información (Manning, Raghavan & Schütze, 2008). La precisión (P) indica el porcentaje de documentos recuperados que son relevantes; es decir, la proporción de entidades propuestas correctamente por el sistema de reconocimiento –en nuestro caso, el porcentaje de errores señalados que realmente lo son–, y se calcula con la siguiente fórmula: P = errores detectados / (errores detectados + falsos positivos).

En segundo lugar, la cobertura (C) muestra el porcentaje de documentos relevantes que son recuperados; o sea, la proporción de entidades existentes que el sistema de reconocimiento recupera correctamente –en nuestro caso, el porcentaje de errores cometidos que son detectados–, y se calcula aplicando esta fórmula: C = errores detectados / (errores detectados + falsos negativos). Y estas dos medidas se combinan en una sola, que es la media armónica entre la precisión y la cobertura (F_1).

Como vemos en la Tabla 8, la precisión de ambos sistemas al analizar el texto que nos ocupa no difiere demasiado (8 y 8.33%). La tendencia es clara: el algoritmo basado en la estadística de la frecuencia de los bigramas resalta un gran número de

combinaciones (92) que, en realidad, son correctas; en cambio, Microsoft Word señala menos falsos positivos (33), pero también menos errores de los que se han cometido. Así que, en términos de cobertura, el primer algoritmo ayuda a detectar un mayor número errores, si bien hay que extraerlos de entre todos los resultados obtenidos. Aunque no es el mismo caso que el que analizan Nazar y Renau (2012), pues ellos están trabajando con hablantes de L2, creemos que se puede aplicar el mismo argumento:

> "It can be argued that, in a task like this, it is preferable to have false positives rather than false negatives, because the difficult part of producing a text is to find the errors. However, a system that produces many false positives will lose the confidence of the user" (Nazar & Renau, 2012: 31).

El riesgo de que el hablante nativo pierda la confianza en el sistema probablemente es menor que en el caso del aprendiz de L2, pues su condición de nativo lo ayudará a rechazar los falsos positivos con mayor celeridad. A continuación exponemos con más detalle las carencias que hemos descubierto en la aplicación del algoritmo.

4. Limitaciones

En primer lugar, hay que tener en cuenta que hay palabras que son poco frecuentes, así que, aunque el banco de datos con el que trabaja el programa contiene cien millones de vocablos, habrá formas cuya frecuencia sea muy baja. Por otro lado, existe una limitación derivada del análisis de palabras adyacentes que realiza el programa, que es, como ya sabemos, la esencia misma de su funcionamiento. Veremos ahora, con algunos ejemplos, en qué se traduce esta carencia y en qué casos la hemos detectado.

4.1. Error de tipo II o falso negativo

Los análisis de los textos que hemos realizado revelan que no se marcan siempre los errores relacionados con la concordancia, ya sea entre el sujeto y el verbo o entre sustantivos y adjetivos, cuando los elementos que deben concordar no son contiguos (22), ni los errores relacionados con el uso de los tiempos verbales o con la correlación verbal (24). Por ejemplo, en las subordinadas de relativo, donde el verbo de la proposición subordinada no va precedido de su antecedente sino del pronombre relativo ('que', en este caso), como vemos en (22):

(22) […] *con la gente que <u>están</u> lejos de ti,

el corrector no identifica que el sujeto es un sustantivo en singular ('la gente') mientras que el verbo aparece en plural ('están'). Las cifras del análisis no nos ayudarán a percatarnos del error: el par 'que están' es una combinación muy frecuente en los textos del corpus (aparece en 4.531 ocasiones), incluso se registra tres mil veces más de lo que sería esperable, teniendo en cuenta la probabilidad de la combinación: 1119.7 (por su parte, la frecuencia de 'que' es 3295376 y la de 'están', 33978; y el umbral es

4.05); y es correcta si el antecedente es plural, como aquí:

(23) […] con las personas que están lejos de ti.

Estamos ante un caso de lo que en estadística se conoce como error de tipo II o falso negativo (también llamado error beta (β)). El programa, al solo analizar pares de palabras adyacentes, no consigue identificar una combinación errónea que está más allá de los límites del par.

Así mismo, al igual que otros muchos correctores –véase el análisis que realiza Burston (1996) de varios correctores de francés, y el de Chen (2009) de correctores de inglés–, no está preparado para identificar algunos errores relacionados con el uso de tiempos verbales o con la correlación verbal; sí detectará, en cambio, una mala formación de un tiempo perfecto, por ejemplo, pues el auxiliar 'haber' deberá ir seguido de un participio –como comprobamos en (6)–. El primer tipo de error (*Mañana Luis no fue al cine) no es común que un nativo lo cometa; sin embargo, los errores de concordancia verbal sí son más frecuentes; como en este ejemplo:

(24) *Nos deseó que tengamos suerte al día siguiente.

El análisis estadístico no nos lleva a pensar que la oración contenga falta alguna, pues la frecuencia de 'que tengamos' es 515 (por separado, la de 'que' es 3295376 y la de 'tengamos', 1346), la probabilidad del bigrama, 44.36 y el umbral, 11.61. Es otro caso de falso negativo, como (22).

4.2. Error de tipo I o falso positivo

En las oraciones con sujeto compuesto, el verbo debe utilizarse en plural, como en esta:

(25) Tanto Lola como su hermana aprobaron el examen.

Ahora bien, si analizamos esta oración, comprobaremos que el par 'hermana aprobaron 'no se registra en el corpus de referencia (la frecuencia de 'hermana 'es 13132 y la de 'aprobaron', 66; la probabilidad del par es 0.01 y el umbral, cero), con lo cual podríamos pensar que se trata de un error. Es un falso positivo o error de tipo I (también llamado error alfa (α) o 'falsa alarma').

No sería lógico, por otra parte, que el programa nos indujera a pensar que esa oración es correcta: la herramienta no resultaría eficaz si diera por buena esa combinación en otros muchos contextos, como en este, por ejemplo,

(26) *Su hermana aprobaron el examen.

Estos son algunos de los contextos en los que el algoritmo no nos será de utilidad. Es conveniente que el usuario tenga claro que el programa solo da información sobre

dos palabras contiguas para así ser consciente del alcance y de las limitaciones de la herramienta.

CONCLUSIONES

En este trabajo hemos presentado un estudio sobre la detección de diferentes tipos de errores que con frecuencia aparecen en textos redactados por hablantes nativos de español. La herramienta utilizada para ello emplea métodos estadísticos, siguiendo las propuestas de otras usadas para corregir textos escritos, sobre todo, en inglés. Estos correctores consiguen localizar rápidamente errores relacionados con las combinaciones de palabras; errores que los correctores que utilizan los análisis de palabras y de oraciones no detectan siempre (como hemos comprobado en los resultados de las pruebas realizadas aquí; véase el apartado 3). En cambio, estos últimos identifican algunos errores que pasan inadvertidos para los métodos basados en estadística. La conclusión es, por lo tanto, que ambos sistemas de detección de errores son complementarios y resultan eficaces en el proceso de edición y corrección de textos.

El punto débil de los correctores gramaticales basados en el análisis estadístico de la frecuencia de las palabras del texto es que, como vimos en el apartado 4, solo identifican errores que puedan ser inferidos analizando información procedente de las combinaciones de palabras adyacentes –lo cual nos puede llevar a pensar que hay fallos donde en realidad no los hay (falsos positivos) o a no detectar otros que sí lo son (falsos negativos)–, como bien resume Chen (2009), tras su análisis de varios correctores de inglés:

> "To sum up, the statistical grammar checker will fail to capture errors if the errors are not word combination problems or they involve problems of non-adjacent word strings or conflicts across different clause boundaries" (Chen, 2009: 175).

Una forma de mejorar el funcionamiento de la aplicación del algoritmo que hemos empleado aquí para que sea todavía más útil es ampliar la extensión de la combinación de palabras analizadas; es decir, que en vez de ser solo pares, fueran grupos de tres, cuatro o, incluso, cinco palabras; de esta manera sería posible detectar algunos de los errores que no pueden ser descubiertos solo con el análisis de dos palabras adyacentes. Así lo han constatado Wu y Su (2006) en su estudio de un corrector utilizado para detectar el uso erróneo de preposiciones en inglés, en el que se ha empleado un 'modelo' de análisis de pares y otro de grupos de tres palabras (trigramas): 'The experiment results show that tri-gram language model can find most of the correct prepositions'.

Otra manera de aumentar la eficacia del corrector es ampliar el tamaño del corpus de referencia, de forma que, al menos, se reduzcan los falsos positivos relacionados con vocablos de baja frecuencia. Si bien, la ampliación del corpus habría que realizarla

con cautela, pues es absolutamente necesario que sea con muestras de lengua correctas, ya que si no es así, tendríamos un problema añadido, que es el que se plantea en Moré (2006) como emplea la web como corpus, el buscador no discrimina por sí mismo 'badly written pages'. Sin duda, una de las bases del buen funcionamiento de un corrector es la calidad del corpus (Athanaselis et al., 2006).

Otra posibilidad es utilizar un corpus donde las partes del discurso (*PoS = Parts of Speech*) estén etiquetadas y existan reglas que codifiquen su funcionamiento, y así a los datos estadísticos se podría sumar este tipo de información.

Finalmente, podría ser útil que el programa proporcionara al usuario algún tipo de retroalimentación sobre el error cometido e incluso algunas opciones para corregirlo. Aunque tampoco es fácil que un programa sea capaz de ofrecer siempre alternativas correctas, como vimos en los ejemplos (1-3, 5-6). La retroalimentación sería de gran utilidad, sobre todo, para el aprendiz de L2, pues al hablante nativo, si el error se ha producido por un descuido (otra cosa bien diferente sería que se debiera al desconocimiento de una regla, del género de un sustantivo, etc.), le basta con verlo señalado para saber cómo corregirlo. Su mayor problema es detectarlo. Recuérdese que aquí se está analizando la utilidad del corrector sobre todo para el hablante nativo competente que escribe textos extensos o que debe corregirlos como parte del proceso de edición.

El usuario no debería olvidar que no se trata de una herramienta infalible, sino que es, más bien, "a flagging tool which brings possible errors to their attention" (Jacobs & Rodgers, 1999: 523). De manera que, igual que cuando, por ejemplo, estamos escribiendo en español e introducimos una cita en inglés, el corrector del procesador de textos nos la subraya, y no por ello consideramos que hemos cometido un error; pues, CorrectMe también llamará la atención sobre combinaciones de palabras que, simplemente, son poco frecuentes, pero no por ello, erróneas. Este programa exige que el usuario no sea un mero receptor pasivo de la información, sino que sea capaz de intuir la causa de los avisos que le envía. Además, es conveniente que el usuario sea consciente de la necesidad de complementar la utilización del corrector con el Diccionario de la lengua española y el Diccionario panhispánico de dudas de la RAE, y con otras utilidades como el Corpus del español actual (CREA) o WebCorp (Renouf, Kehoe & Banerjee, 2007), mediante las cuales puede constatar la frecuencia de una determinada combinación de palabras en otros corpus de textos y los contextos en los que suele aparecer.

En definitiva, pensamos que la aplicación del algoritmo es una herramienta útil que puede contribuir a que los textos escritos en español, al menos los producidos por la comunidad universitaria, contengan un menor número de errores, ya que es capaz de detectar faltas que otros correctores no identifican. Asimismo, esa misma aplicación dispone de una versión adaptada para estudiantes de español como L2, la cual esperamos que se beneficie también del análisis aquí realizado.

REFERENCIAS BIBLIOGRÁFICAS

Ariza, A. & Tapia, A. M. (1997-1998). El corrector ortográfico y la presentación del texto escrito. *Cauce. Revista de Filología y su Didáctica*, 20-21, 375-412.

Athanaselis, T., Bakamidis, S. & Dologlou, I. (2006). An automatic method for revising ill-formed sentences based on *N -grams*. *Speech Prosody. ISCA Archive* [en línea]. Disponible en: http://www.isca-speech.org/archive/sp2006/papers/sp06_080.pdf

Briscoe, T., Medlock, B. & Andersen, O. (2010). Automated assessment of ESOL free text examinations. *Technical Report*. University of Cambridge Computer Laboratory, 790 [en línea]. Disponible en: http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-790.pdf

Burston, J. (1996). A comparative evaluation of French grammar checkers. *CALICO Journal*, *13*(2-3), 104-111.

Chen, H.-J. (2009). Evaluating two web-based grammar checkers - Microsoft ESL Assistant and NTNU Statistical Grammar Checker. *Computational Linguistics and Chinese Language Processing*, *14*(2), 161-180.

Chodorow, M. & Leacock, C. (2000). An unsupervised method for detecting grammatical errors. *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. Association for Computational Linguistics, 140-147 [en línea]. Disponible en: http://www.aclweb.org/anthology-new/A/A00/A00-2019

Díaz Villa, A.M. (2005). Tipología de errores gramaticales para un corrector automático. *Procesamiento del Lenguaje Natural*, 35, 409-416.

Gamon, M., Leacock, C., Brockett, C., Dolan, W. B., Gao, J., Belenko, D. & Klementiev, A. (2009). Using statistical techniques and web search to correct ESL errors. *CALICO Journal*, *26*(3), 491-511.

García-Heras Muñoz, A. (2007). Programas informáticos de corrección gramatical en el aprendizaje de una lengua extranjera (inglés): Expresión escrita. *Docencia e Investigación: Revista de la Escuela Universitaria de Magisterio de Toledo*, 17, 71-101.

Hernández García, F. (2012). Palabras problemáticas y frases incorrectas: una solución autónoma para detectar lo indetectable. *Revista electrónica de lingüística aplicada*, 1, 41-55.

Islam, A. & Inkpen, D. (2011). Correcting different types of errors in texts. En C. Butz & P. Lingras (Eds.), *Advances in Artificial Intelligence* (pp. 192-203). Berlín-Heidelberg: Springer.

Jacobs, G. & Rodgers, C. (1999). Treacherous Allies: Foreign language grammar checkers. *CALICO Journal*, *16*(4), 509-531.

Lawley, J. (2015). New software to help EFL students self-correct their writing. *Language Learning & Technology*, *19*(1), 23-33.

Lawley, J. & Martin, R. (2006). Corrector de gramática para estudiantes autodidactas de inglés como lengua extranjera. *Revista de Educación*, 340, 1171-1191.

Lin, N. Y., Soe, K. M. & Thein, N. L. (2011). Developing a chunk-based grammar checker for translated English sentences. *25th Pacific Asia Conference on Language, Information and Computation*, 245-254 [en línea]. Disponible en: http://www.newdesign.aclweb.org/anthology/Y/Y11/Y11-1026.pdf

Manning, Ch., Raghavan, P. & Schütze, H. (2008). *An introduction to information retrieval*. Cambridge: Cambridge University Press.

Moré, J. (2006). A grammar checker based on web searching. *Digithum*, 8, 1-5 [en línea]. Disponible en: http://www.uoc.edu/digithum/8/dt/eng/more.pdf

Naber, D. (2003). *A rule-based style and grammar checker*. Tesis doctoral, Universidad de Bielefeld, Bielefeld, Alemania [en línea]. Disponible en: http://www.danielnaber.de/languagetool/download/style_and_grammar_checker.pdf

Nadasdi, T. & Sinclair, S. (2001-2015). *SpanishChecker.com. Corrector de ortografía y gramática*. Nadaclair Language Technologies [en línea]. Disponible en: http://spanishchecker.com/

Nazar, R. & Renau, I. (2012). Google Books N-gram corpus used as a grammar checker. *Proceedings of EACL 2012: Second Workshop on Computational Linguistics and Writing*. Avignon, France [en línea]. Disponible en: http://www.aclweb.org/anthology/W12-0304

Real Academia Española: Banco de datos (CREA) (2008). *Corpus de referencia del español actual* [en línea]. Disponible en: http://www.rae.es

Real Academia Española (2005). *Diccionario panhispánico de dudas (DPD)*. Madrid: Santillana Ediciones Generales, S. L.

Real Academia Española (2010). *Ortografía de la lengua española*. Madrid: Espasa Calpe.

Real Academia Española (2014). *Diccionario de la lengua española*, 23.ª Ed. Madrid: Espasa Calpe.

Renouf, A., Kehoe, A. & Banerjee, J. (2007). WebCorp: An integrated system for web text search. En M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus Linguistics and the Web* (pp. 47-67). Ámsterdam: Rodopi.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sjöbergh, J. (2006). Chunking: An unsupervised method to find errors in text. *Proceedings of the 15th NODALIDA Conference, Joensuu 2005*. University of Joensuu electronic publications in linguistics and language technology, 1, 180-185 [en línea]. Disponible en: http://epublications.uef.fi/pub/urn_isbn_952-458-771-8/urn_isbn_952-458-771-8.pdf#page=187

Villena, J., González, B., González, J. C. & Muriel, M. (2002). STILUS: Sistema de revisión lingüística de textos en castellano. *Iberamia. VIII Conferencia Iberoamericana de Inteligencia Artificial* [en línea]. Disponible en: http://www.lsi.us.es/iberamia2002/confman/SUBMISSIONS/185-ulaededlle.PDF

Wu, S.-H. & Su, C.-Y. (2006). An evaluation of adopting language model as the checker of preposition usage. *Proceedings of the 18th Conference on Computational Linguistics and Speech Processing, ROCLING 2006, Taiwan*. Association for Computational Linguistics and Chinese Language Processing (ACLCLP) [en línea]. Disponible en: http://aclweb.org/anthology/O/O06/O06-1023.pdf

Yannakoudakis, H. (2013). Automated assessment of English-learner writing. *Technical Report*. University of Cambridge Computer Laboratory, 842 [en línea]. Disponible en: http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-842.pdf

* AGRADECIMIENTOS

Queremos dar las gracias a James Lawley por la gran ayuda prestada durante la elaboración de este trabajo y por todas sus sugerencias y observaciones; así como a Sergio Martín por programar CorrectMe y al Centro Superior de Enseñanza Virtual (CSEV) por financiar y apoyar este proyecto de investigación. También agradecemos a los evaluadores anónimos de la revista, sus comentarios y precisiones; y al Equipo de Edición sus indicaciones y correcciones.