ISSN 1870-4069

Método de corrección ortográfica basado en la frecuencia de las letras

Edgar Moyotl-Hernández

Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias Físico Matemáticas, Puebla, México

emoyotl@fcfm.buap.mx

Resumen. En este trabajo se presentan los primeros resultados de un método de corrección ortográfica para la variante del español mexicano, basado en la distancia de edición propuesta por Levenshtein. Para mejorar el funcionamiento de este algoritmo, se propone asignar costos diferentes a las operaciones de edición tomando en cuenta la frecuencia de las letras. Los resultados obtenidos en la evaluación son satisfactorios, especialmente si se considera que se trata de un corrector ortográfico de propósito general y que las palabras se analizan sin ningún tipo de información contextual. Además, este enfoque es capaz de detectar errores que otros correctores no identifican.

Palabras clave: Corrector ortográfico, errores ortográficos, distancia de edición, algoritmo de Damerau-Levenshtein.

Spell Checking Method Based on the Frequency of Letters

Abstract. This paper presents the first results of a spell checking method for variant of Mexican Spanish based on the Levenshtein edit distance. To improve the performance of this algorithm, we propose to assign different costs to editing operations taking into account the frequency of the letters. The results of the evaluation are good, especially if you consider that this algorithm is a spell checker general purpose and that the words are analyzed without any contextual information. In addition, this approach detects errors that other spelling checkers are not able to identify.

Keywords. Spell checker, orthographic errors, edit distance, Damerau-Levenshtein algorithm.

1. Introducción

Con el aumento en la cantidad de información textual generada por las personas y disponible en formato digital, se requieren herramientas que procesen

la mayor cantidad de información posible. Sin embargo, las personas en su redacción producen un sin número de errores y la presencia de estos errores en los textos reduce las posibilidades de éxito a las aplicaciones estándar de *Procesamiento de Lenguaje Natural (PLN)*, encargadas del análisis y procesamiento de documentos. Por ejemplo, un buscador de textos no podrá recuperar documentos para una consulta realizada por un usuario si la misma se escribió mal o los documentos contienen errores en las palabras con las que se está consultando.

Los algoritmos de corrección tienen como objetivos: detectar y corregir, ya sea de forma automática o interactiva, los errores de redacción generados por los humanos; esto con el fin de aumentar la calidad de los textos. Estos algoritmos son utilizados ampliamente por los procesadores de texto, y también se aplican en tareas como la normalización de textos, detección de plagios, reconocimiento óptico de caracteres, recuperación de información entre otras. Por ejemplo, algunos sistemas buscadores como Google[1] son capaces de detectar errores en las consultas mal escritas y brindar sugerencias de palabras correctas; si se teclea *"corector"* en la barra de búsqueda y se da la orden de buscar, Google instantáneamente mostrará *"Se muestran resultados de* ***corrector****"*.

Justamente, la mayoría de los correctores existentes se enfocan en corregir errores ortográficos; su función es identificar palabras que posiblemente estén mal escritas y presentar al usuario una serie de alternativas de corrección. Para saber si una palabra está escrita correctamente o no, el enfoque más sencillo consiste en utilizar como referencia un diccionario con la lista de (casi) todas las palabras válidas de la lengua tratada. Una palabra está escrita correctamente si se encuentra en el diccionario, en caso contrario es considerada un error. Para este trabajo la composición del diccionario se hizo con el *Corpus de Referencia del Español Actual (CREA)*[2] compuesto de una amplia variedad de textos, producidos en todos los países de habla hispana desde 1975 hasta 2004 y cubre el español mexicano.

Para obtener la lista de alternativas en el caso de que la palabra sea errónea, el corrector buscará en el diccionario las palabras que se obtengan a partir de la palabra incorrecta mediante las operaciones elementales de caracteres: inserción, borrado, sustitución o transposición, las operaciones permitidas en la distancia Damerau-Levenshtein [2], [5]. A diferencia del modelo original, el método propuesto asigna diferentes pesos a las operaciones de acuerdo con la frecuencia de las letras en el español de México[3]. Así mismo, el orden de las alternativas para elegir la mejor candidata se apoya en la probabilidad de ocurrencia en el corpus.

Este trabajo está organizado de la siguiente forma. En la sección 2., se revisan las características de la corrección ortográfica. Luego, en la sección 3. se describe el concepto de distancia de edición y se muestra su uso en los correctores ortográficos. Posteriormente, en la sección 4. se presentan las ideas utilizadas en el método propuesto para modificar el costo de las operaciones de edición.

[1] https://www.google.com
[2] http://www.rae.es/recursos/banco-de-datos/crea
[3] http://dem.colmex.mx

Después en la sección 5., se describen los datos, los experimentos y los resultados de esta propuesta para corregir textos. Finalmente, se presentan las conclusiones y el trabajo futuro que se espera desarrollar.

2. Corrección ortográfica

Los errores de redacción se pueden clasificar de diferentes maneras. En [6] se propone la siguiente clasificación:

- Errores ortográficos (palabras no presentes en el idioma).
- Errores gramaticales (palabras del idioma pero no correctas en el contexto).
- Errores de estilo (palabras redundantes, ambiguas o repetidas).

De la misma forma, los errores de redacción se pueden clasificar, según su naturaleza, en dos clases: *errores ortográficos*, palabras no presentes en el idioma; y *errores gramaticales*, palabras del idioma usadas incorrectamente en el contexto [1]. Por lo tanto, el objetivo de un corrector ortográfico es señalar al usuario las palabras del texto que se encuentran escritas de manera errónea, y sugerir la palabra más apropiada dentro de una serie de palabras válidas en el idioma. En cambio, la tarea del corrector gramatical es más compleja, ya que se encarga de verificar la correcta construcción de una oración, como la concordancia de género y número, tiempos verbales, etcétera; por lo que no siempre pueden descubrir errores en los que la palabra es ortográficamente correcta pero su uso es incorrecto en un contexto específico (por ejemplo: *"una obra de teatro popular* ***mexicana****"*).

Por otra parte, los errores ortográficos producidos por las personas durante la composición de un texto pueden producirse por alguna de las siguientes razones:

- por equivocaciones al teclear los caracteres, los que son *tipográficos*,
- o por desconocimiento de las reglas de ortografía, los llamados *cognitivos*.

Así, por ejemplo, la palabra *"teirra"* es un error tipográfico causado por transponer los caracteres *"e, i"*. Esto se debió a que se presionaron en orden inverso las teclas correspondientes a los caracteres mencionados. Y un ejemplo en el que aparece un error cognitivo es *"canpo"* cuando la palabra correcta es *"campo"*. La causa de este error es el desconocimiento de la regla ortográfica que establece que antes de *p* o *b* se escribe *m*.

Del mismo modo, los errores de carácter ortográfico ya sean tipográficos o cognitivos se pueden diferenciar, de acuerdo con el error cometido, en las categorías siguientes [1]:

- Error por sustitución de una letra por otra (por ejemplo, *"elije"* en lugar de *"elige"*).
- Error por inserción de una letra extra (por ejemplo, *"dirrección"* en lugar de *"dirección"*. Un caso particular es la separación de palabras cuando se introduce un espacio entre ellas, tal como *"video juegos"* por *"videojuegos"*).

- Error por eliminación de una letra (por ejemplo, *"bibioteca"* en lugar de *"biblioteca"*. De igual manera, un caso particular es la unión de palabras, como se muestra en *"denuevo"* por *"de nuevo"*).
- Error por transposición de dos letras adyacentes (por ejemplo, *"haora"* en lugar de *"ahora"*).

En [2] se definió a los errores simples como las palabras que presentan una sola ocurrencia de uno de los errores anteriores: sustitución de una letra, inserción de una letra, eliminación de una letra o transposición de dos letras. Por consiguiente, se pueden catalogar los errores ortográficos como errores simples o errores múltiples según el número de errores que los diferencia de las palabras correctas, esto es, la cantidad de caracteres erróneos presentes en el error. Por lo tanto, los errores simples presentan una sola transformación y los errores múltiples más de una.

3. Distancia de edición

La medida básica utilizada para calcular la similitud entre palabras es la distancia de edición también conocida como *distancia de Levenshtein*[5]. El algoritmo cuenta la cantidad de operaciones requeridas para convertir una cadena de caracteres en otra, es por esto que, las únicas operaciones de edición permitidas son:

- Inserción de un carácter,
  alumo $->$ *alumno* (agregar *'n'* entre la *'m'* y la *'o'*).
- Eliminación de un carácter,
  transladan $->$ *trasladan* (quitar la primer *'n'*).
- Sustitución de un carácter por otro,
  numero $->$ *número* (reemplazar la *'u'* por *'ú')*.

Notese que la respuesta de este algoritmo es un valor numérico. Por ejemplo, la distancia entre *"tsetal"* y *"tzeltal"* es dos, dado que se necesitan dos operaciones elementales para transformar una palabra en otra:

1. *tsetal* $->$ *tzetal* (sustituir la *'s'* por *'z'*)
2. *tzetal* $->$ *tzeltal* (insertar la *'l'*)

Con el propósito de mejorar dicho algoritmo Damerau agregó una operación más, la transposición de dos caracteres adyacentes [2].

- Transposición de un carácter con otro,
  paelta $->$ *paleta* (permutar *'e'* con *'l'*).

Esta modificación de la distancia clásica de Levenshtein para contar la transposición de caracteres como una sola operación y no como dos distintas, produjo una medida de edición diferente conocida como *distancia Damerau-Levenshtein*. Damerau también encontró que el 80% de todos los errores ortográficos corresponde a palabras erróneas con distancia de edición igual a uno respecto a la palabra que originalmente debía escribirse. Esto significa que el 80% de los errores son de carácter tipográfico y se encuentran en una de las cuatro categorías de error descritas en la sección anterior.

3.1. Modelo de corrección

En general, el método de corrección basado en la distancia de Levenshtein trata de encontrar en el diccionario la palabra correcta c que es más similar a la palabra no reconocida w. En otras palabras, si la palabra es errónea entonces se obtendrán todas sus transformaciones posibles mediante la inserción, eliminación, sustitución y transposición de cada uno de sus caracteres. Después de generar todas las transformaciones de la palabra mal escrita, cada una de ellas se busca en el diccionario y las palabras que se encuentren en él se agregan a la lista de alternativas. La mejor sugerencia de corrección será la de menor distancia a la palabra errónea.

Como ejemplo, en la Tabla 1 se muestran todas las palabras generadas a partir de la palabra errónea *"**vidro**"* con una sola operación de edición. En este caso, las correcciones que se tienen son *"cidro"*, *"video"* y *"vidrio"*. La problemática consiste ahora en definir cuál es la corrección más apropiada, puesto que originalmente todas las operaciones de edición tienen el mismo costo y éste es unitario.

Tabla 1. Posibles transformaciones de la palabra *"vidro"* con distancia de edición uno.

Operación	Palabras generadas
Eliminación	*idro, vdro, viro, vido, vidr*
Inserción	*avidro, bvidro, cvidro, ..., zvidro, ..., vidroa, vidrob, ..., vidroz*
Sustitución	*aidro, bidro,* ***cidro****, ..., zidro, ..., vidra, vidrb, ..., vidrz*
Transposición	*ivdro, vdiro, virdo, vidor*

En estas condiciones, la distancia sólo es un valor numérico que cuenta el número de transformaciones, de modo que, mientras más transformaciones haya mayor será el valor del costo y viceversa. Por supuesto, esta limitación se corrige si se asignan costos distintos a las operaciones [7]. Por ejemplo, en el español mexicano es más probable encontrar la letra 'a', esto justificaría asignar un costo de sustitución menor cuando se cambie un carácter por 'a' que por otro menos frecuente. En la Tabla 2 se muestran las frecuencias de las letras en el vocabulario fundamental del español de México el cual tiene 842 vocablos y es considerado una muestra representativa del español mexicano (y de la lengua española en su conjunto) [4].

Por otra parte, para el idioma español cuyo alfabeto se compone de 27 letras: *a, b, c, d, e, f, g, h, i, j, k, l, m, n, ñ, o, p, q, r, s, t, u, v, w, x, y, z*; una palabra de longitud n produce: n eliminaciones, $n-1$ transposiciones, $26n$ sustituciones y $26(n+1)$ inserciones lo que da un total de $54n+25$ palabras generadas. Mas aún, de este conjunto de palabras solamente un pequeño número serán palabras reales presentes en el diccionario del idioma. En el ejemplo anterior, para la palabra *"vidro"* de longitud 5, se obtuvieron 295 transformaciones y sólo 3 palabras válidas.

Tabla 2. Frecuencias de letras en orden descendente (extraídas de [4]).

Letra	Frec.	Letra	Frec.	Letra	Frec.
a	631	l	193	h	31
r	611	d	191	j	27
e	584	u	180	z	25
o	425	m	174	q	18
i	379	p	173	y	15
n	324	g	78	ñ	11
c	309	b	68	x	11
t	282	v	52	k	0
s	239	f	46	w	0

Por consiguiente, el proceso de generar todas las transformaciones puede ser costoso computacionalmente. No obstante, si se considera la frecuencia de las letras, entonces se podría reducir la cantidad de palabras a evaluar. Por ejemplo, en el caso del español usual mexicano, las letras *'k'* y *'w'* podrían no usarse en las inserciones o sustituciones porque su frecuencia es cero (ver Tabla 2).

3.2. Modelo de lenguaje

En [8] se describe la implementación de un corrector ortográfico cuyo modelo de corrección es el siguiente: dada una palabra errónea, se trata de encontrar la palabra en el diccionario con mayor probabilidad de corregirla. Es decir, dada una palabra w se intenta encontrar la corrección c_i, de entre todas las posibles correcciones, que maximice la probabilidad de corregir a w, esto es:

$$argmax\ P(c_i|w). \tag{1}$$

Que de acuerdo con el *teorema de Bayes*, esto es equivalente a:

$$argmax\ \frac{P(w|c_i)P(c_i)}{P(w)}. \tag{2}$$

Pero, puesto que $P(w)$ es la misma para toda corrección c_i, la ecuación 2 se reduce a:

$$argmax\ P(w|c_i)P(c_i), \tag{3}$$

donde $P(c_i)$ es la probabilidad de que la corrección sugerida c_i ocurra en el idioma utilizado; $P(w|c_i)$ es la probabilidad de que la palabra w haya sido escrita en lugar de c_i; y $argmax$ es la función que determina el valor máximo de la ecuación 3 para encontrar la palabra correcta.

A su vez, la probabilidad $P(c_i)$ se aproxima directamente como la frecuencia de ocurrencia de la palabra c_i en el corpus. Mientras que $P(w|c_i)$ se aproxima como el número de veces que se escribe w en lugar de c_i por uno de los errores ortográficos simples (sustitución, inserción, eliminación o transposición). De modo que, para estimar dicha probabilidad se requiere determinar tanto la frecuencia como el tipo de errores que ocurren en el idioma tratado. Por esta razón, en [8] se optó por definir que las correcciones c_i con distancia de edición uno respecto a w son más probables que las c_i con distancia de edición mayor que uno. Así, de todas las correcciones generadas que aparecen en el diccionario, el sistema elige como correcta aquélla que tenga mayor probabilidad de ocurrencia en el corpus de referencia.

4. Método propuesto

Este trabajo propone una modificación al algoritmo propuesto por Levenshtein [5]. Para ello, se asignan a las operaciones de edición costos basados en la frecuencia de las letras que componen a las palabras. Este esquema de ponderación intenta maximizar la distancia entre w y c_i, el error y la corrección; en particular, cuando esta última es poco probable en el vocabulario del español de México ya que utiliza caracteres menos frecuentes en él.

4.1. Costos de edición

En esta primera propuesta solamente se modificará el costo de las operaciones de inserción y sustitución, de modo que los costos de las operaciones son los siguientes:

- Costo de inserción y sustitución: si el carácter es el más frecuente entonces el costo de la operación es el mínimo, en caso contrario el costo se incrementará.
- Costo de eliminación y transposición: el costo es uno para cualquier carácter.

Con esta modificación se observa que, cuando el costo de edición entre dos palabras es cercana a cero su distancia entre ellas es casi nula produciendo un valor de cercanía alto. Por el contrario, si las palabras no son similares en su grafía, entonces su distancia aumentará.

4.2. Composición del diccionario

Como se ha dicho, cuando se construye la lista de sugerencias para un error ortográfico se necesita comprobar que cada posible corrección esté presente en el idioma, razón por la cual es necesaria la composición de un diccionario del español. Así que, para la composición del diccionario se utilizó la lista de todas las formas ortográficas presentes en el Corpus de Referencia del Español Actual (CREA)[4] que cuenta con casi 140 000 documentos; más de 154 millones de

[4] http://corpus.rae.es/lfrecuencias.html

palabras procedentes de textos de todos los países hispánicos y producidos entre 1975 y 2004; más de 700 000 palabras diferentes y más de 100 materias distintas.

El criterio para decidir si una palabra pertenece o no al idioma español fue que se encontrará en la lista de palabras y que su *frecuencia normalizada* en el corpus fuera mayor a un umbral, el cual fue establecido a 0.5 con base en experimentos; si la palabra no cumple con estas condiciones se considera un error ortográfico.

4.3. Algoritmo

El método de corrección involucra todas las técnicas que fueron descritas anteriormente y para ser más específicos se explica a continuación:

1. Separar cada palabra del texto, estas palabras se evaluarán individualmente.
2. Detectar palabras erróneas mediante el uso del diccionario. Para identificar a las palabras que no están escritas correctamente, se compara cada palabra con las existentes en un diccionario.
3. Para cada error generar todas sus posibles transformaciones (con distancia de edición uno), que a su vez se buscan en el diccionario para eliminar todas aquellas que no estén presentes en el mismo.
4. Ordenar la lista de palabras correctas de acuerdo con la distancia que estas correcciones tengan con la palabra errónea y con la probabilidad de ocurrencia en el corpus.
5. Seleccionar la sugerencia con costo más bajo.

5. Experimentos

Con el fin de evaluar la precisión del método propuesto, para detectar y corregir errores, se utilizaron 55 oraciones cada una con un error ortográfico, dando como resultado 45 errores distintos. La fuente de este material fue la Fe de erratas de los libros de educación primaria del ciclo escolar 2013-2014 publicada por la Secretaría de Educación Pública (SEP) [3].

Aunque en realidad fueron 117 los errores ortográficos, gramaticales y semánticos los que se detectaron en los libros (de todos los niveles formativos y en la totalidad de sus asignaturas) sólo se utilizaron los errores ortográficos simples. De ahí que, errores ortográficos como *"**tsetal**"* o *"**Iztacihuátl**"* no se consideraron en las pruebas porque la corrección, *"tzeltal"*, del primer error se obtiene con 2 operaciones (una sustitución y una inserción) y la corrección, *"Iztaccíhuatl"*, para el segundo error requiere de 3 operaciones (una inserción y dos sustituciones). La Tabla 3 muestra algunas de las oraciones con problemas de ortografía utilizadas para probar el método.

Además, se implementaron los siguientes métodos de corrección:

– Método base: usa la distancia de edición clásica donde cada operación de edición tiene un costo uniforme de 1. En este modelo se selecciona la palabra más frecuente en el corpus, como se realizó en [8].

Tabla 3. Ejemplos de errores ortográficos reales.

Dice	**Debe decir**
"Relaciones entre los ***numeros****"*	*"... números"*
*"****Rescribir*** *canciones conservando la rima"*	*"Reescribir ..."*
"A la ***vibora*** *de la mar"*	*"... víbora ..."*
"En esta lección ***alaborarás*** *un ritmo visual"*	*"... elaborarás ..."*
*"****Elije*** *alguno que te haya gustado e imita su postura"*	*"Elige ..."*
*"****Codice*** *florentino, siglo XVI"*	*"Códice ..."*
"Si Juanito rompió el ***vidro*** *a propósito"*	*"... vidrio ..."*
"Juan ***Nepomuseno*** *Almonte"*	*"... Nepomuceno ..."*
*"****Gadalajara****, Jalisco"*	*"Guadalajara ..."*
"Mi único medio de ***trasporte*** *era un burro"*	*"... transporte ..."*

– Método propuesto: usa la modificación a la distancia de edición original para que las operaciones tengan un costo diferente y luego se selecciona la palabra más utilizada en el corpus.

5.1. Resultados

A continuación se presentan los resultados del algoritmo propuesto en comparación con los correctores ortográficos de Google Docs[5] y Microsoft Word 2016[6]. En la Tabla 4 se resume la evaluación, de la detección y corrección de errores, realizada con dichos correctores y el método propuesto. De los resultados obtenidos se puede observar, por una parte, que los métodos basados en la distancia de Levenshtein detectaron el 86.6% de los errores posibles; notese que ambos obtienen el mismo resultado porque utilizan el mismo diccionario. Por otra parte, el método propuesto corrigió el 71.1% de errores y el método base corrigió el 73.3%, esto significa que los algoritmos son comparables en funcionamiento. Por último, Google Docs identificó sólo el 44.4% de errores y Word 2016 detectó el 77.7%; aunque no se conocen los algoritmos que utilizan estas herramientas, es muy probable que utilicen métodos más complejos, a pesar de ello, sólo corrigieron de forma apropiada el 40% y 66.6% de errores, respectivamente.

Conviene subrayar que los resultados se evaluaron como apropiados cuando la primera sugerencia para el error coincidió exactamente con la corrección establecida. Por lo tanto, los resultados del algoritmo propuesto podrían mejorar si se toman en cuenta todas las sugerencias encontradas para los errores. Por ejemplo, al revisar los experimentos se observó que la lista de sugerencias para el error *"****rescribir****"* fue *"describir"*, *"reescribir"*, *"prescribir"*, *"escribir"* y para *"****vidro****"* fue *"video"*, *"vidrio"*, *"vitro"*, *"cidro"*.

[5] https://docs.google.com
[6] https://www.microsoft.com

En general, se observa que el algoritmo que se propone permite identificar y corregir adecuadamente los errores de tipo ortográfico. Sin embargo, debido a su simplicidad, el corrector falla en detectar el error cuando este produce una palabra válida, distinta de la que el usuario deseaba escribir. Así, por ejemplo, la primer sugerencia para el error *"**elije**"* fue *"elija"*.

Finalmente, de la lista completa de palabras analizadas, que se muestran en la Tabla 5 del Anexo 1, se puede observar que ninguna de las herramientas de corrección identifica a todas las palabras erróneas aunque estas sean comunes, por ejemplo: no detectan a *"**closet**"* ni a *"**arboles**"* por lo que se puede concluir que esas palabras sí se encuentran en el diccionario y además, tienen una frecuencia de ocurrencia alta en el corpus de referencia. Así mismo, se observa que el error *"**físicomotrices**"* fue detectado pero no corregido, esto sucedió porque la palabra generada *"fisicomotrices"* aunque es correcta no se encontró en el diccionario utilizado.

Tabla 4. Resultados de la evaluación de los correctores con los 45 errores.

	Errores			
Método	**Detectados**	**No detectados**	**Corregidos**	**No corregidos**
Base	39	6	33	12
Propuesto	39	6	32	13
Google Docs	20	25	18	27
Word 2016	35	10	30	15

6. Conclusiones

El aporte principal de este trabajo es la utilización de la frecuencia de las letras para desarrollar un método de corrección ortográfica orientado al español de México. Este enfoque requiere poca intervención humana puesto que solo necesita de un corpus para construir el diccionario y el modelo de lenguaje. Por supuesto que, el corrector ortográfico como casi todas las herramientas que se construyen para procesar textos depende, entre otras cosas, del dominio de aplicación y del idioma a tratar.

Indiscutiblemente, este método no utiliza ninguna de las propiedades lingüísticas de la palabra ni el contexto en que ésta se utiliza. Esta característica impide la corrección de ciertos errores, en concreto, cuando la palabra sugerida es correcta pero no se encuentra en el diccionario utilizado; en estos casos la palabra se seguirá considerando errónea. No obstante, los resultados obtenidos en la evaluación son prometedores puesto que se trata de un corrector ortográfico de propósito general.

El trabajo futuro consistirá en elaborar un corpus de errores reales, ya que de él se podrán obtener los casos de error más frecuentes y en consecuencia un

modelo de lenguaje que proporcionará información para estimar las probabilidades de las sugerencias, esto con el fin de mejorar la precisión al elegir la palabra correcta. De igual manera, las cuestiones relacionadas con los errores ortográficos múltiples, gramaticales y de estilo, se tratarán en trabajos posteriores.

Agradecimientos. Agradezco las observaciones y sugerencias de los revisores, que sin duda alguna contribuyeron a mejorar la presentación y calidad de este trabajo.

Referencias

1. Castro, D.: Métodos para la corrección ortográfica automática del español, Tesis de Maestría, Facultad de Matemáticas y Computación, Universidad de Oriente, Santiago de Cuba (2012)
2. Damerau, F.: A technique for computer detection and correction of spelling errors, Communications of the ACM, vol. 7(63), pp. 171–176 (1964)
3. Fe de erratas de los libros de educación primaria del ciclo escolar 2013-2014, Secretaría de Educación Pública (SEP) (2013)
4. Lara, L. F.: Diccionario del Español de México (DEM), El Colegio de México, A.C., Consultado el 4 de Septiembre del 2016 en: http://dem.colmex.mx
5. Levenshtein, V.: Binary codes capable of correcting deletions, insertions, and reversals, Soviet Physics Doklady, vol. 10(8), pp. 707–710 (1966)
6. Naber, D.: A rule-based style and grammar checker (2003)
7. Needleman, S., Wunsch, C.: A general method applicable to the search for similarities in the amino acid sequence of two proteins, Journal of molecular biology (Elsevier), vol. 48 (3), pp. 443–453 (1970)
8. Norvig, P.: How to write a spelling corrector (2007), Consultado el 4 de Septiembre del 2016 en: http://norvig.com/spell-correct.html

Anexo 1: Resultados experimentales en detalle

En este anexo se presentan los resultados que comparan la exactitud del método propuesto con la de otras herramientas de corrección (ver Tabla 5). La primera columna es la palabra errónea, la segunda es la corrección apropiada y de la tercera a la sexta columna están los resultados de la corrección. En dicha tabla, el símbolo '−' corresponde a los errores no detectados, '+' señala los errores detectados y corregidos, '×' señala los errores detectados pero no corregidos y un '∘' indica los errores detectados que no tienen sugerencias, por lo cual, tampoco fueron corregidos.

Tabla 5. Resultados para todas las palabras erróneas con diferentes correctores.

Error	**Corrección**	**Método Base**	**Método Propuesto**	**Google Docs**	**Word 2016**
numeros	números	+	+	−	+
panques	panqués	×	×	−	−
escola	escolar	−	−	−	×
exijen	exigen	+	+	+	+
rescribir	reescribir	×	×	×	−
vibora	víbora	+	+	−	+
prrault	perrault	+	+	+	×
atribuída	atribuida	×	+	−	+
heróicos	heroicos	+	+	−	+
quetzalcoatl	quetzalcóatl	+	+	−	+
nauseas	náuseas	+	+	−	−
boiler	bóiler	×	×	−	−
podium	pódium	+	+	−	+
sonreirle	sonreírle	+	+	−	+
closet	clóset	−	−	−	−
alberges	albergues	+	+	+	−
kenedy	kennedy	+	+	−	+
compañia	compañía	+	+	−	+
ditrosionan	distorsionan	+	+	+	+
ocaciona	ocasiona	+	+	+	+
contrarestan	contrarrestan	+	+	+	+
alaborarás	elaborarás	×	×	−	+
elije	elige	+	×	+	−
leccion	lección	+	+	−	+
iconográgico	iconográfico	+	+	−	+
bibioteca	biblioteca	+	+	+	+
ademas	además	−	−	−	+
alumo	alumno	+	+	−	×
físicomotrices	fisicomotrices	∘	∘	−	×
provacar	provocar	+	+	+	+
mantenela	mantenerla	+	+	+	+
vidro	vidrio	+	×	×	+
codice	códice	+	+	−	+
crea-tividad	creatividad	+	+	+	×
gobieno	gobierno	+	+	+	+
trasporte	transporte	−	−	+	−
sequias	sequías	+	+	×	−
arboles	árboles	−	−	−	−
gadalajara	guadalajara	+	+	+	+
simboligía	simbología	+	+	+	+
ligüísticos	lingüísticos	+	+	−	+
transladan	trasladan	+	+	+	+
terrirorio	territorio	+	+	+	+
cristobal	cristóbal	−	−	−	+
nepomuseno	nepomuceno	+	+	+	+