Capítulo 1

Recursos de ayuda a la edición
Ortografía, sintaxis y estilo

Xavier Gómez Guinovart

Introducción

La edición de documentos con la ayuda del ordenador es la aplicación de la informática personal más generalizada en nuestros días. Gracias a los programas de procesamiento de textos, el ordenador llega a ser una máquina de escribir con capacidades de edición que contiene herramientas de asistencia a la escritura inusitadas, como la ruptura automática con guión de las palabras a final de línea, el cómputo de palabras o el control de los cambios efectuados en diferentes versiones de un mismo documento. Dedicaremos este capítulo al estudio de las herramientas de verificación automática de la corrección ortográfica, sintáctica y de estilo de los textos. Se trata de un conjunto de herramientas de edición caracterizado por su entronque con el campo del procesamiento del lenguaje natural y por su incidencia en la calidad final de la escritura.

Las finalidades de este capítulo son las siguientes:

1. Delimitar el ámbito y los objetivos de la verificación lingüística automática de la ortografía, la sintaxis y el estilo como herramientas de ayuda en la edición de textos.
2. Analizar las características de los errores de ortografía, los errores gramaticales y los problemas estilísticos propios de la escritura por ordenador.
3. Examinar las técnicas lingüísticas e informáticas utilizadas en la verificación automática de la ortografía, la sintaxis y el estilo en el marco general del procesamiento de textos.

1. La verificación ortográfica

Los verificadores ortográficos son programas informáticos que sirven para revisar la ortografía de un texto. Estos programas suelen realizar dos tareas diferenciadas: por una parte, la identificación de las palabras del texto que suponen algún error de ortografía; por otra, la determinación de la forma correcta de la palabra o, cuando esto no es posible, sugieren la forma correcta. En la primera parte de la sección, examinaremos las características de los errores ortográficos propios de la escritura con ordenador, y en la segunda parte veremos las técnicas utilizadas para su detección y corrección.

1.1. Los errores ortográficos

Los errores de ortografía que las personas cometemos cuando escribimos un documento con ordenador pueden ser motivados por un desconocimiento de la norma o bien por un descuido. En el primer caso, donde la persona no sabe cómo tiene que escribirse correctamente la palabra que contiene el error, se habla de errores de competencia; en el segundo caso, cuando la persona conoce cómo tiene que escribirse una palabra, pero por alguna causa comete un error de escritura, se habla de errores de actuación.

1) Los **errores de competencia** tienen una vertiente individual, ya que no todas las personas disfrutan del mismo nivel de conocimiento de las reglas de ortografía de su lengua. Sin embargo, hay ciertos factores lingüísticos que favorecen su aparición, como el grado de disparidad entre la ortografía y la fonética de una palabra (**ombro* por *hombro*), las discrepancias entre la normativa y el habla (**líbido* por *libido*), la interferencia con otras normativas típica de las situaciones de plurilingüismo (**immenso* por *inmenso*, por interferencia con el catalán *immens*) o la baja frecuencia de uso de una palabra.

2) Los **errores de actuación** de la escritura por ordenador pueden reflejar los aspectos fonológicos de los errores del habla (**desmolarizar* por *desmoralizar*, con intercambio del rasgo fonológico de localización entre las consonantes líquidas), pueden provenir de un descuido visual (**problemente* por *probablemente*, con el segmento *blemente* iniciado tras la letra *o* que no le corresponde) o pueden ser fruto de un desacierto mecanográfico (**escritutra* por *escritura*, a causa de pulsar dos teclas vecinas del teclado al mismo tiempo).

En cuanto a su aspecto formal, se ha encontrado que la gran mayoría de las palabras incorrectas de los textos escritos con ordenador pertenecen a una de estas cuatro categorías:

- palabras con una letra de más (**meditterráneo* por *mediterráneo*),
- palabras con una letra de menos (**exausto* por *exhausto*),
- palabras con una letra en el lugar de otra (**inexcrutable* por *inescrutable*) y
- palabras que contienen una inversión del orden de dos letras adyacentes (**perposición* por *preposición*).

Además, según se ha podido comprobar de manera empírica, se cometen muy pocos errores en la primera letra de una palabra y, en cambio, son bastante frecuentes los errores de repetir una misma letra (**escrritura*) o de reducir dos letras seguidas iguales a sólo una (**inación* por *inacción*).

1.2. Técnicas de verificación ortográfica

Las técnicas de **identificación** automática de las palabras con errores de ortografía de un texto suelen fundamentarse en su comparación con una lista de palabras correctas almacenadas en el ordenador. Esta lista puede concebirse como un diccionario ortográfico normativo de la lengua que incluya todas las formas flexivas de las palabras. El programa de verificación indicará un error cuando una palabra del texto no se encuentre en esta lista. El problema de esta técnica son las dimensiones del diccionario y su influencia negativa en la velocidad de procesamiento de la verificación. Una solución posible es utilizar un diccionario que contenga sólo las raíces de las palabras, complementado con un conjunto de reglas morfológicas de aplicación previa a la verificación. Otra solución consiste en convertir la lista de palabras en una lista de números (técnicamente, en una tabla bidimensional de ceros y unos o *mapa de bits*) mediante la aplicación a cada palabra de una fórmula matemática y, aplicando la misma fórmula a cada palabra del texto analizado, comprobar si están la lista.

Independientemente del método que se haya utilizado, puede darse el caso de que una palabra correcta del texto no esté recogida en el diccionario del ordenador, con lo cual se produce una **falsa alarma** de error durante la verificación. Las falsas alarmas son frecuentes con los nombres propios y con las palabras técnicas o poco habituales. Con el fin de reducir su frecuencia, los programas comerciales de verificación ortográfica permiten la ampliación personalizada de sus diccionarios. Otras

veces, sin embargo, el fracaso en la identificación se produce porque el error ortográfico en una palabra origina otra (diferente de la pretendida) que se encuentra en el diccionario (**vienen* por *tienen*) y, por lo tanto, el programa no detecta ningún error de ortografía. Si la secuencia vulnera las reglas sintácticas de la lengua (**lo vienen* por *lo tienen*), el error podrá ser identificado durante la verificación gramatical (podéis verlo más adelante); en cambio, si no las vulnera (**ellas vienen* por *ellas tienen*), lo más fácil es que pase desapercibido. Para resolver este problema, se aplican técnicas de identificación cuantitativas, basadas en la probabilidad de coaparición de palabras en un *corpus* de la lengua.

En cuanto a la **corrección** de los errores, la técnica clásica para averiguar la forma correcta consiste en invertir los cuatro procesos de error mayoritarios antes mencionados (omisión o inserción de una letra, sustitución de una letra por otra y transposición de dos letras adyacentes), respetando la primera letra (posición donde no suelen producirse errores). Así, cuando el verificador identifica un error, trata de buscar en el diccionario las posibles formas correctas entre aquellas palabras que empiezan por la misma letra y que sólo suponen uno de estos cuatro procesos.

Ejemplo

Las sugerencias de corrección de **catra* incluirían las palabras *catara* (omisión de *a*), *cara* (inserción de *t*), *catre* (sustitución de *e* por *a*) y *carta* (inversión de *t* y *r*). Esta técnica suele ampliarse con la incorporación de sugerencias para los errores que contienen más de uno de estos cuatro procesos (**aietra* por *abierta*, con omisión y transposición) y para los que empiezan por una letra incorrecta (**sbierta*).

Otras técnicas de corrección más sofisticadas no se limitan a estos cuatro procesos, sino que tratan de descubrir las palabras correctas que más se parecen fonéticamente u ortográficamente al error identificado.

Ejemplo

Calculando el número de secuencias de dos o de tres letras (*bigramas* o *trigramas*) que poseen en común. Así, **golzar* (formado por los trigramas [#*go*, *gol*, *olz*, *lza*, *zar*, *ar*#], donde el símbolo # representa un carácter de inicio y final de palabra) tendría tres trigramas en común con *golear* ([#*go*, *gol*, *ole*, *lea*, *ear*, *ar*#] y con *gozar* ([#*go*, *goz*, *oza*, *zar*, *ar*#], pero sólo uno con *glosar* ([#*gl*, *glo*, *los*, *osa*, *sar*, *ar*#]).

Igualmente, puede calcularse la similitud fonética entre dos palabras utilizando un diccionario fonético (una lista de palabras transcritas fonéticamente) y un programa capaz de convertir la palabra incorrecta en su transcripción fonética aproximada, y computando las secuencias de símbolos fonéticos que comparten.

2. La verificación sintáctica

Analizaremos la verificación sintáctica o gramatical en dos partes: en la primera delimitaremos el concepto de gramaticalidad y las características de los errores gramaticales, y en la segunda parte veremos cuáles son las técnicas de verificación gramatical.

2.1. La gramaticalidad y los errores gramaticales

En relación con los errores ortográficos, los errores gramaticales se caracterizan por su mayor grado de indefinición. Siempre puede determinarse si una secuencia de caracteres respeta o no respeta las reglas ortográficas instituidas por la normativa de una lengua. En cambio, no siempre es fácil decidir si una secuencia de palabras ortográficamente correctas contiene un error gramatical o no. Considerad estos ejemplos[1]:

(i)
a) ?La rata que el gato que el perro cazó comió murió
b) *Yo me parece que no tienes razón
c) ?La ventana leía un cuento
d) ?El perro leía un cuento

Estas secuencias ilustran la distinción clásica en teoría lingüística entre **gramaticalidad** y **aceptabilidad**. La aceptabilidad sería un concepto relacionado con el ámbito de la actuación o el uso del lenguaje, mientras que la gramaticalidad estaría vinculada a la competencia o el conocimiento del lenguaje.

Son enunciados aceptables aquellos que no resultan extraños estilísticamente y cuya comprensión no requiere un gran esfuerzo de concentración o de memoria.

Ejemplo

(ib) es una secuencia agramatical (por *A mí me parece que no tienes razón*) pero aceptable, ya que no produce ninguna sorpresa ni dificultad de comprensión. Por el contrario, (ia) sería una secuencia gramatical pero inaceptable, ya que sólo puede entenderse después de una cierta observación y un análisis (observad: *La rata murió* > *La rata [que el gato comió] murió* > *La rata [que el gato [que el perro cazó] comió] murió*). En cuanto a (ic) y a (id), su rareza es sobre todo semántica, ya que el verbo *leer* suele aparecer con un sujeto con el rasgo semántico de humano (*El niño leía un cuento*). La secuen-

1. Como es habitual en los textos de lingüística, indicaremos con un asterisco (*) las secuencias agramaticales, y con signos de interrogación los enunciados con problemas de aceptación.

cia (ic) puede resultar más insólita que (id), porque la desviación semántica con respecto a los humanos sorprende menos con seres animados que con entidades inanimadas. A pesar de esto, notad que tanto (ic) como (id) podrían ser aceptables en un contexto semántico apropiado (en un cuento, por ejemplo).

Desde el **punto de vista formal**, los errores gramaticales propios de la escritura con ordenador suelen producirse por la omisión de una palabra (**Le informó que vendría* por *Le informó de que vendría*), por la adición de una palabra (**Dice de que viene* por *Dice que viene*) o por la sustitución de una palabra por otra (**Gana de diez* por *Gana por diez*).

Finalmente, considerando sus causas, los errores gramaticales, como los ortográficos, pueden ser **errores de competencia** (por desconocimiento de la norma) o **errores de actuación** (lapsus de escritura). Los primeros, en especial los más típicos, aparecen compilados en obras lingüísticas de carácter normativo, como ejemplos negativos de aquello que no se tiene que decir ni escribir, como es el caso de los ejemplos anteriores. Los segundos son mucho más heterogéneos e imprevisibles, y pueden originarse por un error mecánico de tecleado o por falta de concentración en la redacción del texto, como por ejemplo en **Este niña leía un cuento* (por *Esta niña leía un cuento*) o **Lo vienen* (por *Lo tienen*).

2.2. Técnicas de verificación gramatical

Analizaremos brevemente tres de las técnicas más importantes para el tratamiento automático de la verificación gramatical.

1) La técnica más utilizada tiene un enfoque casuístico y se basa en el **reconocimiento de patrones**. Esto quiere decir que el verificador recorrerá el texto tratando de encontrar las secuencias de palabras que siguen unas determinadas pautas de error preestablecidas. Estas pautas o patrones de error pueden estar definidos en el aspecto meramente gráfico, o utilizar alguna información lingüística accesible para el verificador. Además, los patrones de error pueden ampliarse con la correspondiente sugerencia de corrección.

Ejemplo

Los patrones ampliados de (ii), donde la sugerencia de corrección se indica a la derecha de la flecha, sirven para identificar y corregir cinco categorías de errores gramaticales nada infrecuentes en español: la omisión indebida de la preposición *de* dentro de la locución *a fin de que*, la ausencia de la contracción obligatoria de *a* y *el*, la alteración del orden correcto de los pronombres átonos *se* y *me* (**Me se olvidó* por *Se me olvidó*), la introducción del complemento oracional de *pensar* con la preposición *de* (**Dice de que viene* por *Dice que viene*), y el uso del futuro en lugar del condicional como segunda parte de una oración condicional con subjuntivo imperfecto (**Si vinieras, yo iré* por *Si vinieras, yo iría*).

(ii)
a) a fin que > a fin de que
b) a el > al
c) me se > se me
d) DECIR de que > DECIR que
e) si ... IMPSUBJ ... FUT > si ... IMPSUBJ ... COND

De este modo, utilizando diferentes tipos de abstracciones sobre los datos (como el lema PENSAR para cualquier forma flexiva del verbo *pensar*, o IMPSUBJ para cualquier verbo en imperfecto de subjuntivo) y símbolos (como los puntos suspensivos para cualquier secuencia de palabras), los patrones alcanzan un grado de generalización que les permite detectar y corregir un buen número de secuencias incorrectas. Estos sistemas de corrección presuponen un proceso del texto consistente en su análisis morfológico y lematización.

Para que el verificador sea capaz de detectar un error, éste tiene que corresponder a uno de los patrones previamente especificados; es decir, para que la verificación sea efectiva, es preciso anticipar el tipo de errores que pueden aparecer, y no todos los errores gramaticales son fáciles de prever.

2) Otra técnica de verificación gramatical, todavía en fase de investigación, se basa en los resultados de los programas informáticos de **análisis sintáctico**, también denominados analizadores o *parsers*.

Estos programas permiten determinar automáticamente la estructura sintáctica de un enunciado de acuerdo con las especificaciones de una gramática. Las técnicas de verificación basadas en el análisis sintáctico consiguen ofrecer un análisis sintáctico de los enunciados gramaticalmente anómalos atenuando las reglas gramaticales que la expresión analizada no respeta. Por ejemplo, si se debilita la regla gramatical de concordancia entre determinante y nombre del analizador, el verificador podrá analizar el enunciado **Este niña leía un cuento* y señalar el error de concordancia.

3) Finalmente, la técnica probabilística de identificación de los errores gramaticales, también en fase experimental, parte del análisis estadístico de un *corpus* textual utilizado como modelo del uso lingüístico. En este *corpus*, se establece la categoría morfosintáctica de cada palabra y se determina la probabilidad de aparición contigua de cada combinación posible de dos categorías. La verificación consiste en detectar, en el texto concreto examinado, las palabras contiguas con categorías morfosintácticas de baja **probabilidad de coaparición** en el *corpus* previo tomado como modelo. La aplicación de este método exige que el verificador sea capaz de descubrir la categoría de cada palabra del texto analizado, una tarea que se hace posible gracias a los buenos resultados de los programas *etiquetadores* (véase el glosario).

3. La verificación estilística

A continuación veremos cuáles son las características de los problemas específicos del estilo, así como las técnicas básicas empleadas para su verificación.

3.1. La propiedad estilística y los problemas de estilo

Veamos ahora en qué consisten la propiedad y la impropiedad del estilo y cómo se relacionan estos conceptos con los de aceptabilidad gramatical y variedad estilística o género. Considérense los siguientes ejemplos:

(iii)
a) ?Los antibióticos (sustancias producidas por microorganismos que en bajas concentraciones inhiben o matan a otros microorganismos), usados hoy contra la tuberculosis (auténtica plaga en otro tiempo) y contra muchas otras enfermedades infecciosas, salvan cada año millones de vidas
b) ?La rata que el gato cazó murió
c) ?Yo y él tenemos que hablar
d) ?La rata que el gato que el perro cazó comió murió

Como hemos explicado antes, los cuatro enunciados de (iii) tienen problemas de aceptabilidad, ya que o bien resultan extraños, o bien es preciso un esfuerzo de concentración o de memoria para entenderlos. Por ejemplo, (iiia) y (iiib) presentan unas dificultades de comprensión intrínsecas a raíz de la complejidad de sus estructuras sintácticas. No obstante, su aceptabilidad global puede depender de las condiciones de su situación comunicativa, en particular, del tiempo disponible de procesamiento y del canal de comunicación.

Desde el punto de vista estilístico, la mayoría de los factores que pueden influir en la aceptabilidad de un enunciado es posible explicarlos por las características lingüísticas del tipo de texto (de la variedad estilística, del género) al cual pertenecen. Así, es probable que (iiia) y (iiib) sean comprensibles y relativamente aceptables en aquellas variedades textuales que utilizan el canal escrito y donde se acostumbra a practicar la lectura pausada, como los textos literarios cultos, y que al mismo tiempo sean casi inaceptables en un género escrito de lectura rápida, como los textos informativos periodísticos, o en un género oral, como las arengas políticas espontáneas. Por otra parte, los problemas de aceptabilidad de (iiic) no dependen de factores sintácticos, sino que se relacionan con las convenciones sociales de cortesía que establecen la posición final del pronombre de primera persona respecto de los otros pronombres o nombres con los cuales coaparece. Sin embargo, la aceptabilidad de (iiic), como la de (iiia) y (iiib), puede depender de la

variedad textual donde aparezca. Así, (iiic) podría ser inaceptable en aquellos registros lingüísticos formales que suponen un respeto estricto de las convenciones de urbanidad; por ejemplo, en la defensa oral presentada por un abogado en un juicio. Y por el contrario, el mismo enunciado puede ser relativamente aceptable y menos sorprendente en aquellas variedades estilísticas menos rigurosas con las normas sociales de cortesía lingüística, por ejemplo, en una conversación telefónica íntima.

Hay, pues, una cierta equivalencia funcional entre la aceptabilidad y la propiedad estilística: un rasgo lingüístico –como la inserción de un constituyente extenso y complejo en (iiia), la subordinación interna de relativo en (iiib) o el incumplimiento de una convención de cortesía lingüística en (iiic)– puede considerarse propio o impropio del estilo de un texto, aceptable o inaceptable, según la variedad estilística a que pertenezca el texto. Esto no quiere decir, sin embargo, que no haya enunciados que resulten inaceptables (y, por lo tanto, impropios) en casi cualquier género de textos, como las construcciones similares a (iiid) con más de dos niveles de subordinación interna de relativo.

Por lo tanto, desde el punto de vista lingüístico, la propiedad estilística de un texto se basa en la afinidad de su estilo con la norma estilística del género al cual se adscribe, y la norma estilística de un género puede definirse a partir de las construcciones lingüísticas de uso más frecuente en el género en cuestión. En consecuencia, la posibilidad de determinar el grado de propiedad (o de impropiedad) del estilo de un texto requiere establecer anteriormente cuáles son las variantes estilísticas o los géneros de una lengua, y cuáles son los rasgos lingüísticos que caracterizan a cada una de las variantes postuladas, es decir, cuáles son los rasgos que definen la norma estilística de cada género establecido.

3.2. Técnicas de verificación estilística

De las diferentes técnicas utilizadas para el procesamiento de la impropiedad estilística, veremos dos de las más utilizadas:

1) La primera de estas técnicas se basa en la asignación previa del texto analizado a una determinada variedad estilística. Antes de que el programa pueda realizar la verificación del documento, es preciso que la persona usuaria del sistema especifique a qué categoría pertenece el texto. Así, el programa lleva a cabo la verificación comparando las características del documento analizado con los rasgos lingüísticos que el sistema considera preceptivos para la categoría textual seleccionada.

En general, esta categoría textual puede seleccionarse a partir de un número de **modelos estilísticos** predefinidos por el programa. Cada uno de estos modelos textuales se define mediante un conjunto de rasgos lingüísticos, como el número máximo de palabras por oración, la presencia o ausencia de giros coloquiales, o el número máximo de sintagmas preposicionales consecutivos.

La detección de los sintagmas preposicionales de un texto requiere un preprocesamiento sintáctico. Sobre este tema, podéis ver el apartado 1.6 del capítulo "Técnicas de procesamiento del lenguaje".

A veces, incluso, el verificador permite crear modelos estilístico nuevos, asignando los valores deseados a las características lingüísticas propuestas por el programa.

2) Otra técnica habitual de verificación estilística es la evaluación del **nivel de legibilidad**, que se define como el grado de dificultad de comprensión del sentido de un texto determinado por ciertos factores lingüísticos cuantificables, como la longitud de las oraciones, la longitud de las palabras, la rareza de las palabras (es decir, el porcentaje de palabras del texto no incluidas en una lista de las palabras usuales de la lengua) o la cantidad de preposiciones en una frase. Los estudios de la legibilidad se basan en las regularidades estadísticas que muestran los textos respecto de este tipo de factores. Un de sus objetivos consiste en elaborar fórmulas de legibilidad, contrastadas empíricamente con métodos estadísticos, que sirven para predecir parámetros estimativos objetivos del grado de dificultad de un texto en función de su estilo. La fórmula de legibilidad de más éxito es la propuesta por Rudolph Flesh para el inglés, que determina el índice de legibilidad (IL) de un texto a partir de la media de sílabas por cada 100 palabras (SP) y la media de palabras por oración (PO), según el cálculo $IL = 206{,}835 - (0{,}846 \times SP) - (1{,}015 \times PO)$. Este índice va de 0 a 100 y tiene que interpretarse según esta escala: muy difícil (0-30), difícil (30-50), más bien difícil (50-60), estándar (60-70), más bien fácil (70-80), fácil (80-90) y muy fácil (90-100).

Resumen

La verificación automática de la corrección lingüística constituye hoy por hoy un campo de investigación y desarrollo muy activo dentro del ámbito de la escritura asistida por ordenador y, más genéricamente, del procesamiento del lenguaje natural. En este capítulo hemos tratado de analizar brevemente sus fundamentos lingüísticos e informáticos, presentando al mismo tiempo una perspectiva pedagógica y global de sus límites y aplicaciones.

Se ha visto cómo los sistemas de corrección automática pueden basarse en el simple reconocimiento de formas o bien en la detección de patrones morfológicos, que presuponen un preprocesamiento del texto y permiten definir generalizaciones productivas sobre cuáles son las secuencias incorrectas.

También se han mostrado las dificultades que encuentran estos sistemas para detectar errores que no presentan una anomalía respecto de los patrones de corrección predefinidos, y las principales líneas de investigación que permitirán avanzar en la solución de estas dificultades.