# El uso de *expresiones regulares* en la detección de errores escritos: implicaciones para el diseño de un corrector gramatical.[1]

Dr. Rubén Chacón Beltrán

UNED. Departamento de Filologías Extranjeras y sus Lingüísticas
Senda del Rey, 7. 28040 Madrid
rchacon@flog.uned.es

**Resumen**

El uso de *expresiones regulares* constituye una forma sofisticada de buscar y, en su caso, reemplazar elementos en un texto. Entre sus posibles aplicaciones estaría la identificación de errores escritos cometidos por hispanohablantes que aprenden inglés, y la asociación de estos errores a explicaciones metalingüísticas que permitan al aprendiz de lenguas modular su expresión escrita y autocorregirse. En esta línea se está trabajando en la UNED para incrementar la fiabilidad de un corrector gramatical pedagógico diseñado para hispanohablantes adultos que aprenden inglés y que ha sido desarrollado recientemente en esta universidad.

Mediante el uso de *expresiones regulares* se espera poder incrementar la eficacia del corrector gramatical dado que una misma *expresión regular* puede detectar varias secuencias o errores.

La posibilidad de disponer de un corrector gramatical pedagógico tiene importantes implicaciones en el ámbito de la enseñanza de lenguas dado que permitirá aprender de forma autónoma; ahorrar tiempo al profesor; ayudar al alumno a asumir responsabilidad sobre su propio aprendizaje; e, incorporar el uso de las nuevas tecnologías en una actividad de producción libre escrita; entre otras ventajas.

En este artículo se hace una introducción al diseño de *expresiones regulares* y se ilustra con ejemplos para mostrar el potencial de esta herramienta de programación. Asimismo se describen algunas de las limitaciones que puede tener el uso de esta sintaxis y de ahí la necesidad de utilizar un sistema híbrido para la elaboración de un corrector gramatical fiable. Seguidamente se describen algunas de las implicaciones pedagógicas que puede tener el uso de un corrector gramatical para la enseñanza de lenguas.

**Palabras clave:** metalenguajes computacionales, aprendizaje del inglés, corrector gramatical, *expresiones regulares*.

**Abstract**

Regular Expressions are a sophisticated way of searching for elements in a text. They can therefore be used to identify mistakes in written English made by students of English as a Foreign Language (EFL). These mistakes can then be linked to metalinguistic feedback that allows language learners to change their own output and so self-correct their compositions. At present a team of linguists at the UNED is working to improve a pedagogic grammar checker specially designed for such students. By means of Regular Expressions we expect to significantly increase the reliability of this grammar checker.

The development of a pedagogic grammar checker brings with it significant implications for foreign language learning and teaching since it can propitiate learner autonomy, helping students assume responsibility for their own learning, and can save valuable teacher time.

**Key words:** computational metalanguages, English language learning, grammar checker, regular expressions.

**Resumen**

[1] La investigación descrita en este artículo se ha llevado a cabo con la financiación del Ministerio de Educación y Ciencia mediante el proyecto I+D con referencia HUM2006-08469/FILO.

L'ús d'expressions regulars és una forma sofisticada de buscar elements en un text. Per aquesta raó, les expressions regulars es poden utilitzar per identificar errors escrits produïts per parlants d'espanyol que aprenen anglès. Aquests errors estan associats a explicacions metalingüístiques que permeten als estudiants de llengües canviar la seva expressió escrita i autocorregir-se. Actualment, un grup de lingüistes de la UNED està treballant per incrementar la fiabilitat d'un corrector gramatical pedagògic dissenyat per parlants adults d'espanyol que aprenen anglès. Mitjançant la utilització d'expressions regulars esperem incrementar de manera significativa l'eficàcia d'aquests corrector gramatical ja que una mateixa expressió regular pot detectar vàries seqüències o errors. El desenvolupament d'un corrector gramatical pedagògic suposa implicacions molt importants per l'aprenentatge i l'ensenyament d'una llengua estrangera ja que pot propiciar l'autonomia de l'estudiant, pot ajudar als alumnes a assumir responsabilitat sobre el seu aprenentatge i pot estalviar temps al professor.

**Paraules clau:** metallenguatges computacionals, aprenentatge de l'anglès, corrector gramatical, expressions regulars.

**Tabla de contenidos**



## 1. Las *expresiones regulares* en lingüística.

Una *expresión regular* en programación es una forma de representar una secuencia de caracteres que describe un conjunto de cadenas sin enumerar sus elementos y haciendo uso de la sintaxis propia de los diferentes lenguajes de programación. Por ejemplo, el grupo formado por las cadenas *color* y *colour* se puede describir mediante la *expresión regular* "col(o|ou)r".

El uso de *expresiones regulares* constituye un mecanismo flexible y eficiente para el procesamiento de textos que requiere el conocimiento y destreza en un lenguaje de programación relativamente sencillo que sirve para hacer búsquedas en un texto determinado, o en un corpus lingüístico. Con el uso de algunas herramientas informáticas como motores de búsqueda, las *expresiones regulares* pueden utilizarse para añadir, aislar o quitar textos o datos de forma rápida y ágil.

Las *expresiones regulares* se componen de dos tipos de caracteres, los llamados metacaracteres que son de tipo especial, por ejemplo, * y los llamados literales o caracteres normales de texto. Para ilustrar esta organización podría considerarse que las *expresiones regulares* son una lengua con texto literal, que serían las palabras, y metacaracteres que serían la gramática. Las palabras se combinan con la gramática de acuerdo con un conjunto de reglas para crear una expresión que comunica una idea. (Friedl, 2006: 5).

La alternancia, por ejemplo, consiste en la posibilidad de indicar al programa informático que ha de buscar una secuencia que puede contener uno u otro caracter de forma indistinta y suele utilizarse conjuntamente con otro metacaracter, esto es, el paréntesis, que sirve para limitar el alcance de la alternancia, es decir, dónde comienza y dónde acaba. Así, la *expresión regular* \sb(i|y)c(i|y)cle\W podría detectar las distintas variantes que los alumnos suelen escribir, es decir: *bicicle*, *bicycle*, *bycicle* y *bycycle*, tres de los cuales son errores frecuentes de los aprendices de inglés en un determinado nivel de competencia. En esta secuencia, el metacaracter \s representa un espacio en

blanco y \W representa cualquier carácter que no estaría incluido en una palabra, por ejemplo, un espacio, un signo de puntuación, etc.

## 2. El uso de *expresiones regulares* para detectar errores en inglés.

Las *expresiones regulares* tienen múltiples utilidades no sólo en el ámbito de la lingüística del corpus y de la lingüística computacional sino también en los lenguajes de programación y, de hecho, es ahí donde surgen sus primeras aplicaciones. En este caso, en la confluencia de dos áreas de investigación en auge como son la enseñanza y aprendizaje de lenguas, y el aprendizaje de asistido por ordenador encontramos una nueva aplicación de las *expresiones regulares*.

La detección de errores, y su posterior tratamiento, supone un área de investigación que viene estudiándose desde hace décadas. Su origen se encuentra en el análisis de errores que se inicia a finales de la década de los 60 como una evolución de la lingüística contrastiva. En su versión fuerte, el análisis de errores consideraba que en la influencia interlingüística, en este caso negativa, residía el origen de todos los errores cometidos por un aprendiz de una segunda lengua. En su versión débil, el análisis contrastivo encontraba aquí una serie de errores, que podían tener su origen en la influencia de la lengua materna pero que, ante todo, eran oportunidades para aprender. Es esta última, es decir, la consideración del error como una oportunidad para aprender la que en este momento nos interesa y de ahí que lo que perseguimos como profesores de inglés e investigadores sea que el aprendiz de lenguas pueda: a) detectar el error; b) asumir que se trata de una forma irregular en la lengua objeto de estudio; c) procesar correctamente el *feedback* que se le proporciona para solucionar el problema; y, d) autocorregir el error de forma eficaz.

En este trabajo nos hemos centrado en la expresión escrita y es éste uno de los ámbitos donde este proceso de cuatro fases puede llevarse a cabo de forma más directa por el aprendiz de lenguas, con la ayuda de una herramienta informática que es el producto de años de investigación, el corrector gramatical *e-gramm*.

Las *expresiones regulares* nos permiten identificar errores en un corpus lingüístico compuesto por las redacciones de aprendices de inglés como lengua extranjera. Se ha recopilado un corpus de redacciones de aprendices que no han sido corregidas anteriormente y que contienen los errores tal y como los escribieron los aprendices. Así, las *expresiones regulares* nos permiten encontrar los errores cometidos por los aprendices de forma automatizada con lo que se pueden realizar búsquedas para analizar la frecuencia relativa de un determinado error o posibles variantes del mismo, que podrán ser tratadas conjuntamente desde un punto de vista pedagógico.

Una herramienta informática como *e-gramm* permitirá al aprendiz de lenguas escribir su redacción en un procesador de textos y luego recibir indicaciones de los errores, o posibles errores, que hay en su producción escrita. De este modo, el programa informático realizaría la labor correspondiente al primero de los pasos descritos anteriormente (“detectar el error”) y que sin duda supone uno de los más difíciles de conseguir pues es éste el momento en el que el aprendiz de lenguas necesita que un elemento externo (en este caso es un programa informático pero también podría ser el profesor o un interlocutor) propicie el *noticing* que iniciará la secuencia de etapas necesaria para autocorregir el error. Una vez que el aprendiz se ha percatado de la presencia del error escrito, éste asume que debe corregirlo y ahí interviene el *feedback* pedagógico escrito que recibe también del programa informático *e-gramm*. Se trata de explicaciones sencillas que le permitirán procesar la información recibida y autocorregir su propia producción escrita libre con el consiguiente procesamiento pedagógico que al

haberse realizado de forma autónoma tendrá un efecto especialmente provechoso sobre el alumno que continúa así depurando su interlengua.

Los aprendices de inglés que comparten una misma lengua materna muestran una clara tendencia a cometer los mismos errores y de ahí que la corrección de errores de este tipo resulte tan provechosa. El análisis de los corpus consultados muestra que los errores se repiten una y otra vez y que, a menos que haya una intervención directa sobre ellos, los alumnos continuarán cometiéndolos siempre y cuando no se produzca el correspondiente *noticing* y la adecuada intervención pedagógica.

## 3. Algunos ejemplos de *expresiones regulares*.

A continuación se analizan algunos ejemplos de errores cometidos por aprendices de inglés como lengua extranjera y las *expresiones regulares* que se han desarrollado para detectarlos a través del corrector gramatical *e-gramm*.

### Ejemplo 1:

Forma incorrecta: * *I think that to buy weapons…*

Forma correcta: *I think that buying weapons …*

*Expresión regular* que detecta el error: \sthink\sthat\sto\s

Éste es un error frecuente cometido por hispanohablantes. El verbo *think* seguido de una oración subordinada suele ir acompañado por un verbo en gerundio y no por un verbo en infinitivo con *to*.[2] Es un ejemplo que sólo encuentra una secuencia, es decir, en esta ocasión la *expresión regular* descrita anteriormente detectará únicamente una secuencia en la que aparecerán las palabras indicadas. Sin embargo, se trata de una ejemplo muy útil por tratarse de una confusión muy frecuente, y su inclusión en el corrector gramatical resulta muy pertinente.

### Ejemplo 2:

Forma incorrecta: **I am agree with you.*

Forma correcta: *I agree with you.*

*Expresión regular* que detecta el error: A. \s(am|are|is|)\s+agree\W

B. \w(’m|’re|’s)\s+agree\W

En este ejemplo encontramos dos *expresiones regulares* capaces de localizar seis errores. Se utilizará una u otra en función del contexto y del registro lingüístico que esté usando la persona que comete el error.

Así, esta expresión contenida en el ejemplo número 2.A. podrá detectar tres posibles errores como son: * *I am agree*; **You/they are agree*; **S/he is agree*. En esta

[2] Puede darse algún caso en el que la secuencia *I think that to* sea gramaticalmente correcta aunque es muy remota. De hecho, en un corpus de 100 millones de palabras aparece sólo 17 veces. En el *feedback* que recibe el alumno puede incluirse información sobre excepciones a la regla.

*expresión regular* el metacaracter \s+ representa uno o más caracteres en blanco. Necesitamos la *expresión regular* 2.B. para detectar los casos en los que los alumnos han usado contracciones en su redacción, propio de un registro informal. La diferencia entre la primera *expresión regular* y la segunda es que en la primera de ellas debe ir precedida de un espacio, y por ello comienza con el metacaracter \s, mientras que en el segundo caso va precedido de otros caracteres, de ahí el uso del metacaracter \W. Esta *expresión regular* puede detectar los siguientes errores: **I'm agree*; **You/they're agree*; **S/he's agree*.

**Ejemplo 3:**

Forma incorrecta: * *I think governments shouldn´t allowed…*

Forma correcta: *I think governments shouldn´t allow…*

*Expresión regular* que detecta el error: \sshould|can|would|could\s\wed\s

En inglés los verbos modales deben ir seguidos de la forma verbal en infinitivo y nunca en participio pasado terminado en *–ed.* En este ejemplo, de nuevo, será necesario incluir algunas excepciones en la retroalimentación que recibirá el alumno o usuario de *e-gramm* pues puede haber otros verbos que terminen en *–ed* y no por ello se encuentran en participio. Por ejemplo, el verbo *feed* en la oración: *I would feed this animal*.

**4. El uso de correctores ortográficos y gramaticales en el aula: *e-gramm*.**

En el ámbito de la enseñanza/aprendizaje de lenguas extranjeras, el dominio de las destrezas productivas supone una de las tareas más arduas para el aprendiz de lenguas. Los profesores dedican muchas horas a corregir los mismos errores sin que esto tenga necesariamente un efecto positivo sobre el *output* de los alumnos dado que no se produce un análisis profundo de sus errores. Por este motivo, resultaría de gran utilidad disponer de un corrector gramatical pedagógico que permitiese a los alumnos:

a) identificar errores en su propia producción escrita,
b) obtener *feedback* específico sobre el error cometido,
c) reflexionar y autocorregir los errores,
d) eliminar un porcentaje alto de los errores escritos.

La incorporación de las nuevas tecnologías a la enseñanza y aprendizaje de lenguas ha evolucionado mucho en las últimas décadas debido tanto a los avances logrados por la investigación como a la incorporación de importantes logros tecnológicos en el ámbito de la educación.

En lo que respecta a la producción escrita, se trata de una destreza que requiere mucha atención dado que cada vez con más frecuencia el medio escrito se utiliza para la comunicación interpersonal, lo que queda patente en la irrupción que ha supuesto el uso del correo electrónico para cuestiones laborales y personales.

Dentro del aula, la producción escrita con demasiada frecuencia no recibe la atención que debiera no sólo por el escaso tiempo dedicado al aprendizaje del inglés como lengua extranjera en los entornos educativos oficiales como la enseñanza secundaria, sino porque supone una destreza cuyo perfeccionamiento requiere mucha dedicación, además de la corrección y el *feedback* del profesor como fuente principal de

información para su introspección. La posibilidad de disponer de una herramienta informática como un procesador de textos que ayude a autocorregir la propia producción escrita representa toda una revolución metodológica por varios motivos, algunos de los cuales se detallan a continuación:

a) ahorra tiempo al profesor;
b) permite que el alumno autocorrija su producción escrita. La corrección de la propia producción es la mejor de las posibles dado que incide únicamente sobre los errores que muestra la interlengua del aprendiz;
c) dirige la responsabilidad del profesor al alumno, lo que se encuentra muy acorde con las nuevas tendencias en autonomía del aprendizaje;
d) permite que el profesor dedique el tiempo de clase a la corrección de errores más complejos que no son detectados por el corrector gramatical, o a la enseñanza de estilística y técnicas de escritura.

Hasta el momento, toda la corrección que reciben los alumnos tiene como origen al profesor, u otros alumnos, en el mejor de los casos, lo que supone un gran consumo de tiempo para el profesor que debe corregir decenas de redacciones que luego tendrán un dudoso interés pedagógico para los alumnos pues difícilmente volverán sobre ellas para analizar los errores cometidos y realizar la práctica necesaria para evitarlos.

Por ahora, no existe un corrector gramatical que sea lo suficientemente fiable ni para aprendices de inglés, ni para hablantes nativos de la lengua que también pueden necesitar correcciones en su producción escrita. Una muestra de ello es el corrector gramatical del procesador de textos de *Microsoft Word* que normalmente no supone una herramienta útil y fiable y, como consecuencia de ello, es una función que se suele tener desactivada. No ocurre lo mismo con el corrector ortográfico de *Microsoft Word* que por el contrario suele ser una herramienta útil y muy utilizada para corregir los errores ortográficos que se producen accidentalmente al escribir a ordenador o a los errores de ortografía de la persona que escribe. No obstante, se trata de una función que, a pesar de ser muy práctica para hablantes nativos de la lengua, no resulta tan interesante para aprendices pues la corrección se hace de modo automático y no existe la posibilidad de aprender del error. Esto se podría conseguir fácilmente "obligando" a la persona que escribe a escribir la forma correcta de la palabra que se ha escrito mal, aunque, sin duda, esto resulta poco práctico para hablantes nativos de la lengua.

Por lo expuesto anteriormente, se justifica la necesidad de un corrector gramatical que ayude a los aprendices de lenguas a mejorar su producción escrita y, sobre todo, que tenga un efecto pedagógico sobre el aprendiz.

## 5. Incorporación de *expresiones regulares* en*e- gramm.*

El corrector gramatical *e-gramm* se encuentra en estos momentos en fase experimental y se ha desarrollado un prototipo con el que se están realizando varios estudios piloto con alumnos reales en centros educativos de enseñanza secundaria. Este prototipo cuenta con el apoyo del Ministerio de Educación y Ciencia a través del proyecto de investigación I + D con referencia HUM2006-08469/FILO

Dado que el programa está basado en un sistema de búsqueda, es decir, contiene una amplia base de datos con los errores y el *feedback* que los alumnos pueden recibir para cada error, cuando el alumno comete ese error la palabra queda resaltada en un color diferente y el alumno recibe los comentarios pertinentes que le puedan ayudar a autocorregirse. El problema técnico que se plantea es que según se va ampliando la base

de datos con los errores posibles, ésta se va ralentizado y pude llegar el caso de que se colapse por la cantidad de entradas. Este problema se soluciona con el uso de *expresiones regulares* pues ante la combinación de errores se reduce sustancialmente el número de entradas, y también la cantidad de comentarios pues varias entradas pueden quedar asociadas a un mismo *feedback*. Este ahorro en términos de espacio mejora considerablemente el funcionamiento del programa.

## 6. Conclusiones e implicaciones pedagógicas.

Los resultados de este proyecto deben servir para redirigir la investigación internacional sobre la corrección de la producción escrita de los aprendices de lenguas, así como hacia la investigación de correctores gramaticales basados en el análisis individual de errores en vez de parsers. Del mismo modo, se puede obtener información muy valiosa relativa al uso de nuevas tecnologías en el aprendizaje autónomo de lenguas, especialmente en el caso de hispanohablantes adultos que aprenden inglés como lengua extranjera. También se podrá obtener información relativa al tipo de *feedback* escrito explícito que puede resultar útil para la modificación del propio *output*.

Asimismo el proyecto en el que nos encontramos trabajando es innovador porque permitirá reconducir la investigación sobre *feedback* automático, como demuestran los resultados de un estudio piloto que ya se ha realizado. *E-gramm* constituye, sin duda, un avance del conocimiento o una innovación de carácter metodológico en el campo de las tecnologías de la información y la comunicación.

Los resultados de este estudio también permitirán discernir implicaciones pedagógicas concretas para la enseñanza de lenguas y especialmente en lo relativo al desarrollo de la destreza de la producción escrita por hispanohablantes adultos que aprenden inglés.

## 7. Referencias bibliográficas.


Chacón Beltrán, R. (2007). "Learner Autonomy and Language Learning: A grammar-checker for EFL Students", in K. Rasulic & I. Trbjojevic (2007). *English Language and Linguistics Studies: Interfaces and Interactions.* University of Belgrade: Serbia. Pp. 267-273.

Friedl, J.E.F. (2006). 3rd. ed. *Mastering Regular Expressions*. O'Reilly: USA.

Lawley, J. (2004). "A Preliminary Report on a New Grammar Checker to help Students of English as a Foreign Language". *Arts and Humanities in Higher Education*, 3/3, 331-42.

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