Búsquedas avanzadas con expresiones regulares

Miércoles 24 de agosto de 2016

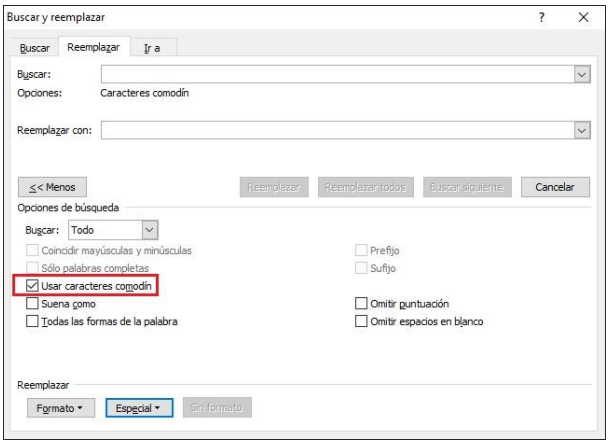
Por el Trad. Públ. Matías E. Desalvo.
Revisión (mayo de 2017): Trad. Públ. Sol Brienza.

Palabras clave: Herramientas para traducción, SDL Trados Studio, Memoq, Microsoft Word

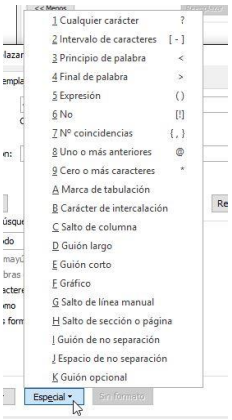
Las expresiones regulares son símbolos que describen un patrón de texto. Se utilizan para realizar búsquedas avanzadas y manipular texto. Según Wikipedia, «Una expresión regular, a menudo llamada también *regex*, es una secuencia de caracteres que forma un patrón de búsqueda, principalmente utilizada para la búsqueda de patrones de cadenas de caracteres u operaciones de sustituciones». Básicamente, permiten buscar secuencias de números, de letras o de números y letras sin la necesidad de escribir las secuencias exactas.

Con pequeñas diferencias en cada caso, se pueden usar en la mayoría de las herramientas de memoria de traducción, como SDL Trados Studio o memoQ, en herramientas de control de calidad, como ApSIC Xbench, y en procesadores de texto, como MS Word. En este caso, y a modo de introducción al maravilloso mundo de las expresiones regulares, vamos a ver algunos ejemplos en MS Word. De acá se puede descargar un archivo de práctica.

En el procesador de texto de Microsoft, las expresiones regulares o caracteres comodín se habilitan desde **Usar caracteres comodín**, de la ficha **Reemplazar** del cuadro de diálogo **Buscar y reemplazar**, como se muestra en la imagen. Aparece cuando presionamos la combinación de teclas Ctrl+L o haciendo clic en **Inicio** > **Edición** > **Reemplazar**, en la cinta de opciones.



Para ver con qué opciones contamos, hacemos clic en el botón **Más** y luego en **Especial** de la parte inferior del cuadro de diálogo.



Ahora veamos paso a paso un ejemplo práctico y útil en el cual describiremos y usaremos algunos de los comodines más útiles de MS Word. Supongamos que tenemos que corregir la puntuación de las cifras de un documento extenso, que contiene muchos números y tablas. El cliente dejó todos los números con puntuación en inglés. Usó coma tanto para las unidades como para las decenas y las centenas de millar. Además, usó punto para los decimales. Nosotros tenemos que usar espacio de no separación para las decenas y centenas, y nada para las unidades de millar. Asimismo, corregiremos los decimales y usaremos la coma propia del español.

Tabla 4

Tipo	África	Asia Pacífico	Europa	América Latina	América del Norte	Desconocido	TOTAL
2009	30	450	619	306	800	5	**2,210**
2013	457	1,000	1,000	800	1,401	13	**4,671**
2014	769	1,443	1,000	800	1,085	22	**5,119**
TOTAL	**1,256**	**102,894**	**2,619**	**521,906**	**3,285**	**40**	**12,000**

Tabla 5

Tipo	Porcentaje de dominios	Muestra inicial	Submuestra analizada	Porcentaje de submuestra
2009	2.4%	4,682	2,210	18.4%
2013	41.9%	82,297	4,671	38.9%
2014	55.7%	109,283	5,119	42.7%
TOTAL	100.0%	196,262	12,000	100.0%

Primero analicemos lo que vamos a buscar. Sabemos que en inglés las cifras que expresan unidades de millar, decenas de millar, etcétera, tienen, de derecha a izquierda, tres números, una coma y una cantidad variable de números a la izquierda de la coma según corresponda en cada caso. Entonces lo primero que tenemos que hacer es buscar cifras que tengan una coma y luego tres números cualesquiera. Por el momento, no nos importa si tiene uno, dos, tres o más números a la izquierda de la coma. Para buscar intervalos de caracteres, sean números o

letras, en el lenguaje de las expresiones regulares, se utilizan corchetes para encerrar el intervalo y un guion para separar el primero del último carácter del intervalo. En otras palabras, si queremos buscar cualquier número del cero al nueve, usaremos la expresión regular «[0-9]» (o [a-z] si deseamos buscar letras). Si seguimos esta lógica, para buscar tres números cualesquiera, usaremos la expresión «[0-9][0-9][0-9]» sin espacios entre los intervalos. De esta manera, cubriremos cualquier combinación de números del 000 al 999. Ahora bien, solo resta anteponer la coma al intervalo para obtener los resultados deseados, o sea, «,[0-9][0-9][0-9]».

Tabla 4

Tipo	África	Asia Pacífico	Europa	América Latina	América del Norte	Desconocido	TOTAL
2009	30	450	619	306	800	5	2,210
2013	457	1,000	1,000	800	1,401	13	4,671
2014	769	1,443	1,000	800	1,085	22	5,119
TOTAL	1,256	102,894	2,619	521,906	3,285	40	12,000

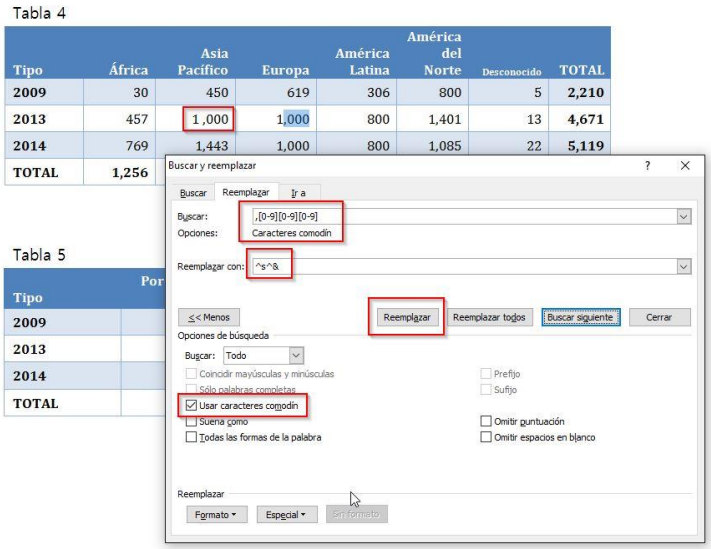
Tabla 5

Tipo
2009
2013
2014

Recordemos que, para utilizar esta función, es necesario marcar la casilla **Usar caracteres comodín** para activarla y que todo aquello que busquemos se debe escribir en el campo **Buscar**. Podemos utilizar el botón **Buscar siguiente** para ir pasando de resultado en resultado y ver que funciona tanto para millares como para decenas y centenas de millar.

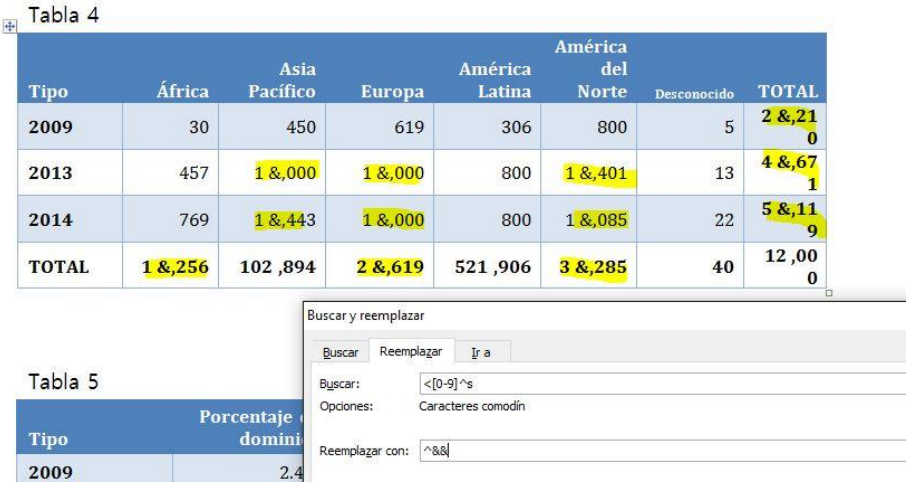
El paso que sigue es realizar las acciones necesarias para reemplazar esa secuencia de caracteres e introducir o eliminar el espacio de no separación, según corresponda, en lugar de la coma incorrecta. Por lo tanto, ahora vamos a colocar el cursor en el campo **Reemplazar con**. En el menú desplegable que aparece al presionar el botón **Especial**, vemos que está la opción **Texto buscado**. Esta permite guardar en la memoria el texto literal que se encontró por medio de la expresión regular. La seleccionamos. Se representa mediante la cadena «^&» que aparece. Luego, colocamos el cursor delante de la cadena de texto buscado «^&» y colocamos un espacio de no separación desde el mismo menú desplegable. Veremos la cadena «^s^&», que significa *espacio de no separación + texto buscado*. O sea, si Word encontró el número «1,000» con los caracteres comodín, reemplazará la cadena «,000» por *[espacio de no separación],000*. Estaremos introduciendo un error, pero es necesario para poder diferenciar las unidades de las decenas y las centenas de millar más adelante.

Para realizar cada reemplazo, hacemos clic en el botón **Reemplazar**. Opcionalmente, podemos utilizar (con mucho cuidado) **Reemplazar todos** para hacer modificaciones en todo el documento en un solo clic.

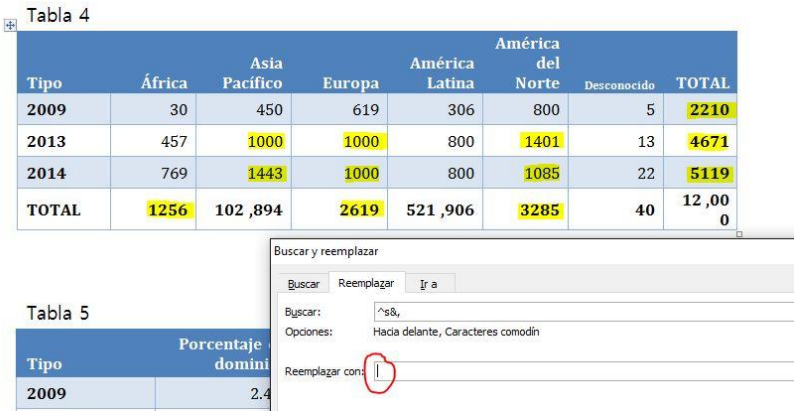


Debido a que tenemos que aplicar dos tipos de puntuación diferentes (espacio de no separación para decenas/centenas y nada para unidades de millar), vamos a diferenciar estos grupos de cifras. Vamos a buscar «<[0-9]^s», donde «<» significa que buscaremos al principio de la palabra/el número y no al final; como sabemos, [0-9] es cualquier número; y «^s» es un espacio de no separación. O sea, con esa cadena, buscaremos cualquier número que aparezca al principio de la cifra y que esté seguido por un espacio de no separación. Esto permite solamente buscar unidades de millar y no decenas o centenas.

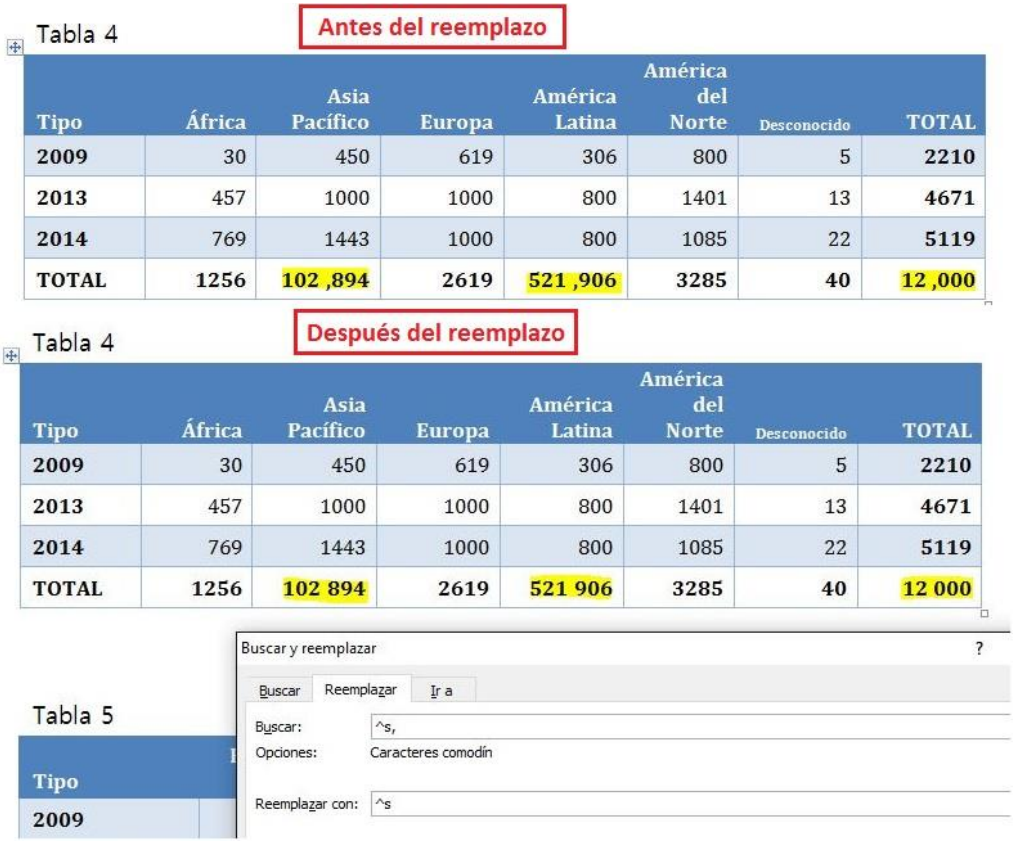
Reemplazamos esa cadena por «^&&», o sea, en el reemplazo se repetirá el texto buscado y se le agregará la letra et. De esta manera, identificamos todas las cifras con unidad de millar. Tendremos resultados del tipo «1 &,000», «5 &,119», «3 &,285», etcétera.



El último paso es reemplazar «^s&,» por nada. Es decir, buscar la secuencia *espacio de no separación + letra et + coma* y reemplazarla por nada (colocar el cursor en el campo **Reemplazar con**, pero borrar todo lo que aparezca allí).



Listo. Ya corregimos todas las cifras con unidad de millar. Solo restan aquellas que tengan decena o centena de millar o las cifras millonarias. Este último paso es muy simple. Basta con reemplazar las cadenas de «^s,» (espacio de no separación + coma) por «^s» (espacio de no separación).



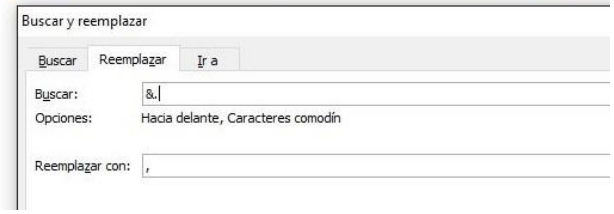
Si presionamos el botón **Mostrar todo** (Ctrl+(), veremos que estas últimas cifras tienen el espacio de no separación correcto y no un espacio común que podría separar la cifra en dos si la cifra estuviera en un párrafo y correspondiera un corte de línea.

Ahora pasamos a la corrección de las fracciones. Sabemos que debemos usar coma en español y no punto. Simplemente reemplazar puntos por comas podría generar un desastre, ya que

también se modificarían comas que aparezcan entre palabras. Por lo tanto, tenemos que volver a marcar esas comas que buscamos de alguna manera especial. Utilizaremos la letra et una vez más. Buscamos la secuencia «.[0-9]» (punto seguido de cualquier número) y la reemplazamos por «&^&» (letra et + texto buscado). Nuevamente veremos cadenas extrañas del tipo «18&.4%», «2&.4%» o «100&.0%», pero ya sabemos qué función cumplen y cómo corregir esto. Reemplazamos «&.» por una coma «,».

Tabla 5

Tipo	Porcentaje de dominios	Muestra inicial	Submuestra analizada	Porcentaje de submuestra
2009	2&.4%	4682	2210	18&.4%
2013	41&.9%	82 297	4671	38&.9%
2014	55&.7%	109 283	5119	42&.7%
TOTAL	100&.0%	196 262	12 000	100&.0%



Misión cumplida. En algunos pasos, corregimos la puntuación de las cifras de un documento extenso, cosa que podría tomarnos todo el día como mínimo si quisiéramos hacerlo de forma manual. Este procedimiento parece complejo, pero no lo es tanto si lo seguimos con cuidado y nos familiarizamos con las expresiones regulares o comodines.

Por último, podemos agregar espacios de no separación entre las cifras y los símbolos de porcentaje. Para ello, buscamos «%» y lo reemplazamos por «^s%». Ahora las tablas quedaron perfectas.

Tabla 4

Tipo	África	Asia Pacífico	Europa	América Latina	América del Norte	Desconocido	TOTAL
2009	30	450	619	306	800	5	2210
2013	457	1000	1000	800	1401	13	4671
2014	769	1443	1000	800	1085	22	5119
TOTAL	1256	102 894	2619	521 906	3285	40	12 000

Tabla 5

Tipo	Porcentaje de dominios	Muestra inicial	Submuestra analizada	Porcentaje de submuestra
2009	2,4 %	4682	2210	18,4 %
2013	41,9 %	82 297	4671	38,9 %
2014	55,7 %	109 283	5119	42,7 %
TOTAL	100,0 %	196 262	12 000	100,0 %

En este artículo, vimos el uso de varios comodines o expresiones regulares en un contexto específico, pero se los puede usar para cumplir con los requisitos que tengamos según el proyecto. Es cuestión de usar la creatividad.

A su vez, vale aclarar que las expresiones regulares son muy parecidas entre programas, pero que pueden variar en cada uno de ellos. Siempre se debe consultar la Ayuda de cada programa para conocer los operadores específicos.