# *Análisis de datos léxicos por Excel VBA: LETRAS-L*

ver. 2014.2.16

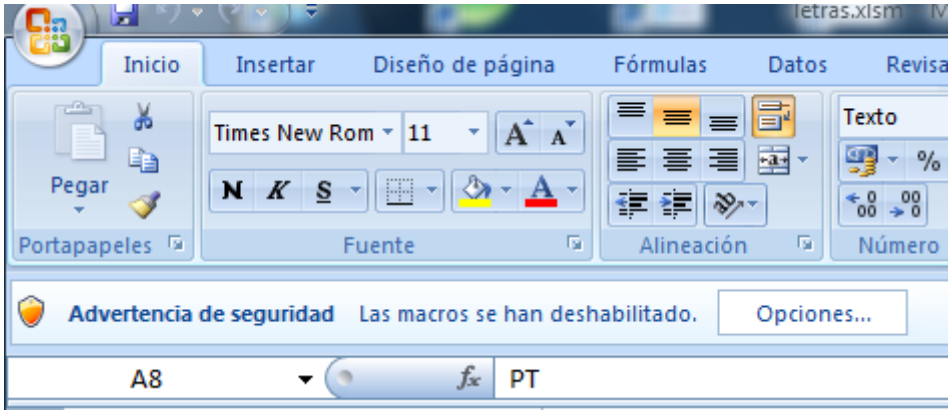
Este documento es un manual de uso para LETRAS.xlsm (en adelante LETRAS), conjunto de programas para el análisis de datos lingüísticos y filológicos. LETRAS está en desarrollo continuo, de modo que este mismo documento también cambia continuamente sin previo aviso. Para el detalle de modificaciones, véase la primera hoja (L) de LETRAS. Le rogamos al usuario que al notar inconvenientes o funciones mejorables, nos los comunique a través del correo electrónico puesto en la etiqueta [Top] (Portada) de la interfaz de LETRAS. Le agradecemos su colaboración.

http://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/

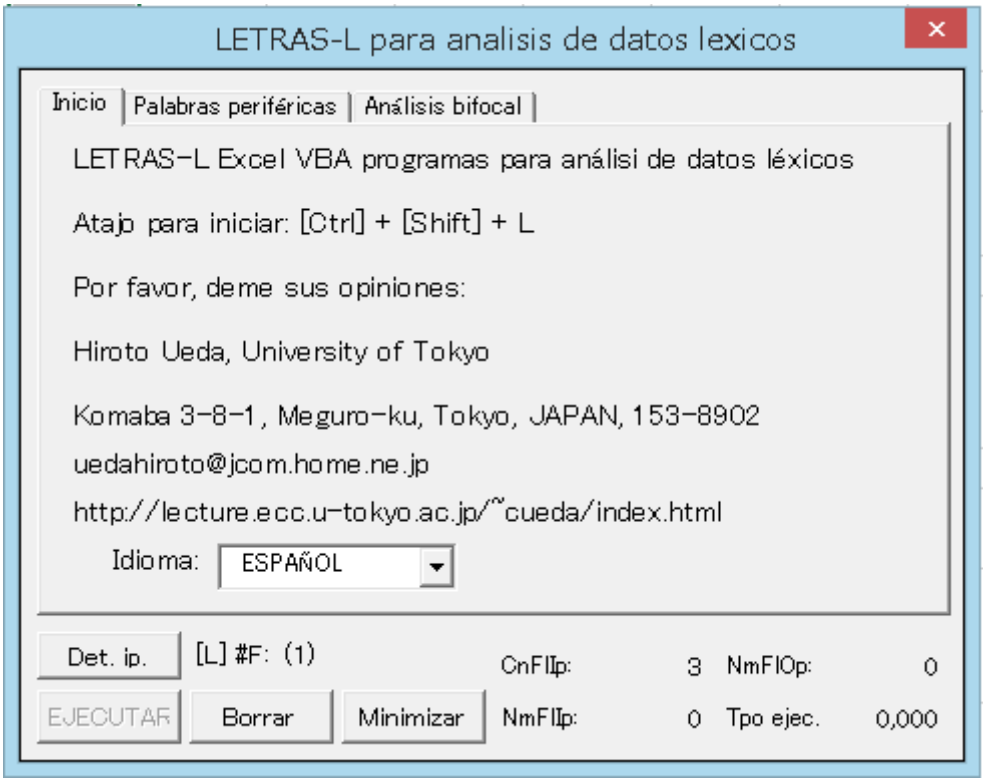
Hiroto Ueda, Ana Isabel García, 2013

# 1. Inicio

Active LETRAS y habilite el programa de Macros contestando positivamente a «Advertencia de seguridad. Opciones…»:




Al pulsar [Ctrl] + [Shift] + L, aparecerá la interfaz de la figura siguiente. Esta figura muestra la etiqueta [Top] (Portada) arriba a la izquierda. Si el usuario desea analizar su Libro (Book) de Excel, después de abrir el Libro, al activar Macro con el atajo [Ctrl] + [Shift] + L, puede analizar el mismo Libro.



En la barra de título, está puesto el nombre del programa: «LETRAS for textual data analysis», seguido de los botones de minimización (_), maximización (☐) y finalización (X). Para maximizar la interfaz minimizada o iniciarla de nuevo, utilice las teclas de atajo: [Ctrl] + [Shift] + L.

La primera hoja [L] tiene en la columna Están asignados la lengua, el sistema, el color de la lengua en la columna [A]. Cambie la asignación de idioma en [A6]=2, [A8]=CM, [A10]=BL y seleccione el color de fondo en [A12]:

| English | Español | 日本語 |
|---|---|---|
| **LETRAS for textual data analysis**<br>**ver. 2013.10.5** | **LETRAS para análisis de datos textuales**<br>« | **LETRAS: テキストデータ分析用プログラム集**<br>« |
| | | |
| **Select laguaguage in the cell [A6]: English=1; Spanish=2; Japanese=3, and restart LETRAS.** | **Seleccione el idioma en la celda [A6]: inglés = 1; español = 2; japonés = 3, y reinicie LETRAS.** | **言語を選択してください。英語=1; スペイン語=2; 日本語=3 をセル[A6]に書き込み再度LETRASを起動してください。** |
| 2 | « | « |
| **Select decimal separator in the cell [A8]: PT (point) or CM (comma), and restart LETRAS.** | **Seleccione el separador decimal en la celda [A8]: PT (punto) o CM (coma), y reinicie LETRAS.** | **小数点を選択してください。(点)=PTまたはCM(コンマ)をセル[A8]に書き込み、再度LETRASを起動してください。** |
| CM | « | « |
| **Select thousands separator in the cell [A10]: PT (point), CM (comma) or BL (blank), and restart LETRAS.** | **Seleccione el separador de miles en la celda [A10]: PT (punto), CM (coma) o BL (blanco), y reinicie LETRAS.** | **千位点を選択してください。PT(点)、CM(コンマ)またはBL(ブランク)をセル[A10]に書き込み、LETRASを再起動してください。** |
| BL | « | « |
| Select background color in the cell [A12]. | Seleccione el color de fondo en la celda [A12]. | 背景色を[A12]に指定してください。 |
| Background color<br>Color de fondo<br>背景色 | « | « |

Al iniciar de nuevo LETRAS la interfaz se cambia en el idioma español, con el separador decimal en coma ( , ).

En esta interfaz se encuentran varias pestañas: [Portada], [Preparar], [Uni/Sep], [Ordenar], [Unifocal], [Bifocal] y [Etiqueta], las cuales iremos explicando en este documento.

En la parte inferior de la interfaz, se encuentran varios botones de comandos, etiquetas, una casilla con una lista y una casilla de verificación:

La figura siguiente demuestra la parte inicial de una muestra de datos textuales, que está en la hoja Tx1 de LETRAS.

| Texto | Título:1 | Título:2 |
|---|---|---|
| A la recepción de un hotel madrileño llega un profesor extranjero para participar como conferenciante en un seminario sobre Nutrición organizado por una universidad de verano con sede en El Escorial. El profesor hablará con el conserje, pidiéndole información sobre los servicios del hotel, así como sobre posibles visitas turísticas por la regió n. | [A] Hotel | (a) Madrid |
| – ¡Buenos días! Desearía una habitación individual para estar tres noches. ¿Qué precio tiene? | [A] Hotel | (a) Madrid |

Se trata de unos ejemplos de conversaciones traducidas al español de cinco lugares distintos: Madrid, Sevilla, Ciudad de México, Lima y Buenos Aires. Los datos están organizados en dos partes de filas: laprimera fila de títulos y las restantes de los datos, divididos según la clasificación de t´titulos. La constitución es libre solo con una condición: los datos textuales siempre en la primera columna [A].

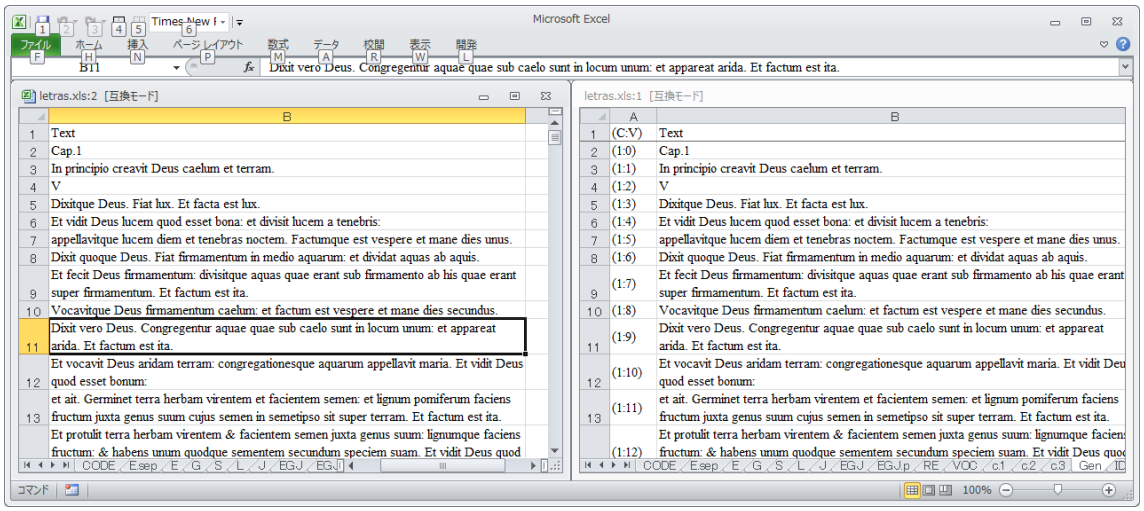
En la parte inferior de la interfaz se encuentra los siguientes botones, casillas, etc.:

[Renovar] Renueva el contenido de la casilla [Ip sheet] en el estado de la hoja actual. Cuando desee analizar los datos de la hoja que está viendo, pulse el botón [RENEW]. Lo puede hacer directamente pulsando la parte correspondiente de la casilla de la lista.

[Borrar] Borra sin confirmación la hoja seleccionada en Excel. Se pueden seleccionar las etiquetas de múltiples hojas (cliqueando con [Shift] o [Ctrl]) y borrarlas. Con el botón [DELETE], se pueden borrar las hojas innecesarias. Se puede seleccionar las hojas no en la casilla de lista de LETRAS, sino en las etiquetas inferiores de Excel. Se borra sin confirmación, de modo que tenga cuidado de no perder las hojas necesarias.

[Ejecutar] Se ejecuta. La primera ejecución puede durar un poco. Puede realizar la ejecución en la etiqueta fuera del inicio. Si se desea detener la ejecución, pulse la tecla [Esc] del teclado. Cuando se trata de un gran número de datos, se renueva el número de fila de input cada 1.000 filas.

[Ventana] Se abre y se cierra alternativamente una ventana más. La hoja de input se visualiza en la ventana izquierda y la hoja de output, en la derecha. Al pulsar el botón de [Ventana], la pantalla muestra dos ventanas horizontales, dos ventanas verticales o una sola, de manera alterna. La opción de dos ventanas es para ver la hoja de input en la ventana izquierda o en la parte superior, y la de output en la derecha o en la parte inferior.




[Fin] Se finaliza. Al pulsar el botón [END] se finaliza el programa de LETRAS. Se pude iniciar de nuevo con las teclas de atajo: [Ctrl] + [Shift] + L.

[Cnt.Fl.Ip]: (Cuenta de filas de input) Se representa la suma de líneas de input.

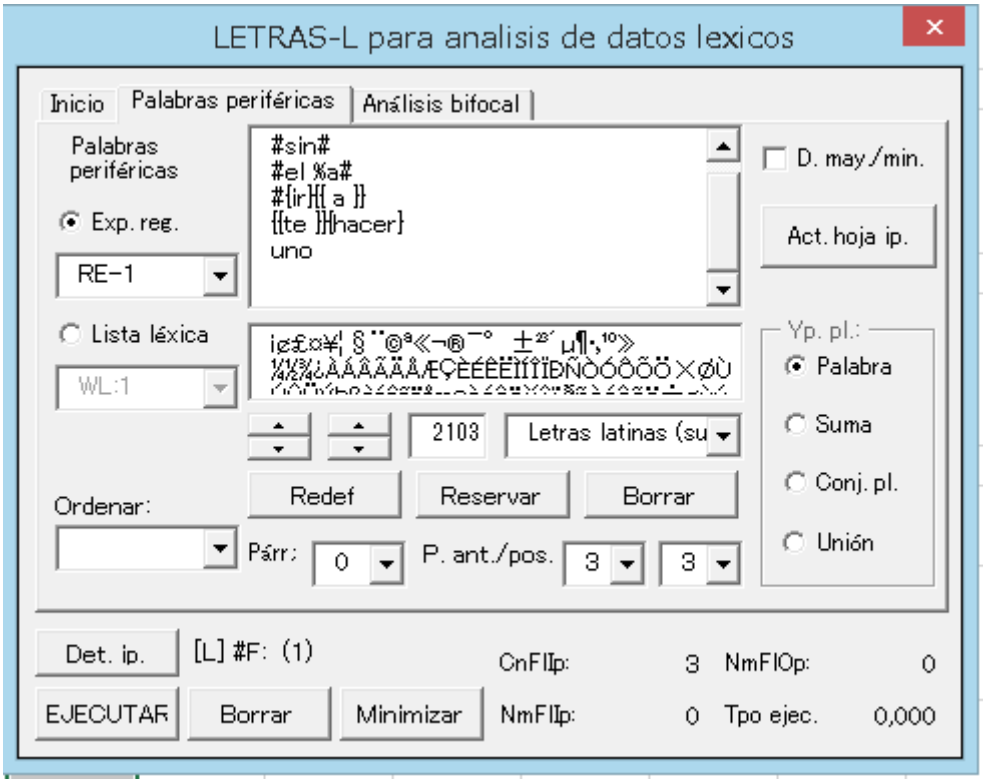
[Nm.Fl.Ip] (Número procesado de filas de input) Se representa en cada momento el número de la línea en proceso.

[Cn.Fl. Op] (Cuenta de filas de output) Al finalizar se representa la suma de las líneas de output.

[Tiempo ej.] (Tiempo de ejecución) Se presenta el tiempo de proceso en milisegundos.

# 2. Palabras periféricas

Con la función de [Seq. (Sequence)] analizamos la secuencia de palabras que se encuentran tanto a la izquierda como a la derecha de la palabra clave, junto con su frecuencia. Se puede seleccionar uno de los resultados: [Form] (Forma), [Sum] (Suma), [Set] (Conjunto), [Union] (Unión), de los cuales solo con [Form] se observa la relación sintagmática, mientras que los otros tres son para análisis paradigmáticos. Se puede seleccionar uno de los dos modos de [Sort] (Ordenación).




## 2.1.1. Forma

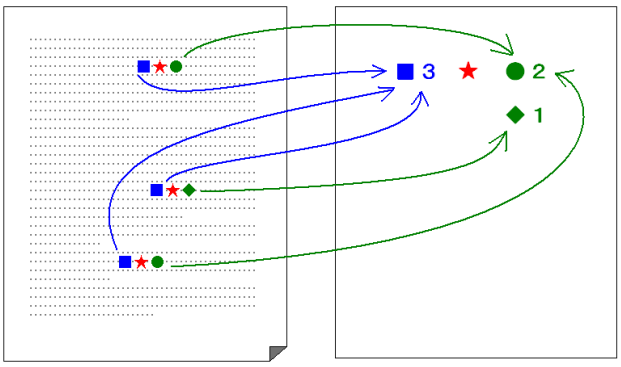
Se realiza la misma operación que [KWIC] (Palabra clave en su contexto), menos la forma de segmentación de los contextos en palabras. Veamos el resultado de «#(por|para)#», donde «#» representa el linde léxico y «(*x*|*y*)», opción de *x* o *y*:

Output [Exp. reg.: #(por|para)#; orden de ocurrencia]

| Pl-3 | Pl-2 | Pl-1 | **Foco** | Pl+1 | Pl+2 | Pl+3 | Título:1 | Título:2 | Fila |
|---|---|---|---|---|---|---|---|---|---|
| un | profesor | extranjero | **para** | participar | como | conferenciante | [A] Hotel | (a) Madrid | 1 |
| sobre | Nutrición | organizado | **por** | una | universidad | de | [A] Hotel | (a) Madrid | 1 |
| posibles | visitas | turísticas | **por** | la | región | . | [A] Hotel | (a) Madrid | 1 |
| una | habitación | individual | **para** | estar | tres | noches | [A] Hotel | (a) Madrid | 2 |
| de | la | habitación | **por** | día | , | desayuno | [A] Hotel | (a) Madrid | 3 |
| descrito | es | perfecta | **para** | mí | . | Me | [A] Hotel | (a) Madrid | 4 |

## 2.1.2. Suma

Al seleccionar [Sum] (Suma), se calculan las frecuencias de las palabras anteriores y las posteriores, inclusive la palabra clave. Se descarta la relación sintagmática, de modo que es para estudiar las frecuencias de las palabras en relación paradigmática. Por esta razón, no reproduce el número identificador de las filas. El resultado será ordenado en orden alfabético, de ocurrencia y de frecuencia.

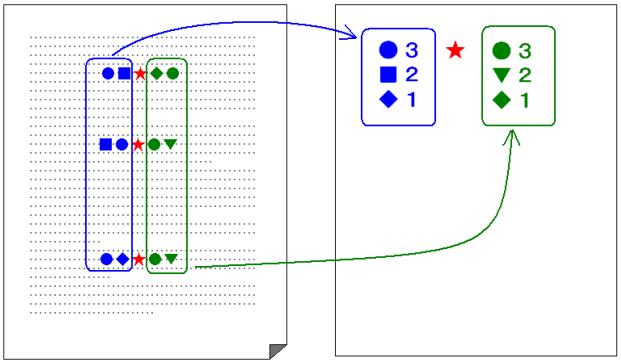



Output [Exp. reg.: #de#; orden de frecuencia]

| Suma | Pl-3 | Suma | Pl-2 | Suma | Pl-1 | Suma | **Foco** | Suma | W+1 | Suma | W+2 | Suma | W+3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 60 | en | 128 | la | 36 | , | 1008 | **de** | 110 | la | 137 | , | 68 | $ |
| 45 | de | 77 | el | 21 | y | | | 29 | las | 124 | . | 57 | . |
| 41 | , | 65 | , | 20 | es | | | 23 | los | 75 | y | 55 | , |
| 41 | y | 42 | un | 13 | centro | | | 22 | todo | 30 | en | 45 | de |
| 32 | $ | 31 | a | 13 | dentro | | | 19 | un | 26 | ... | 42 | y |
| 30 | . | 22 | una | 13 | lo | | | 17 | acuerdo | 24 | ? | 33 | ¿ |

## 2.1.3. Conjunto

Al seleccionar [Set] (Conjunto), se calculan las frecuencias totales del conjunto de las palabras anteriores, el de las posteriores y el de la palabra clave. Esta función se utiliza para ver qué palabra ocurre con frecuencia alrededor de la palabra clave:

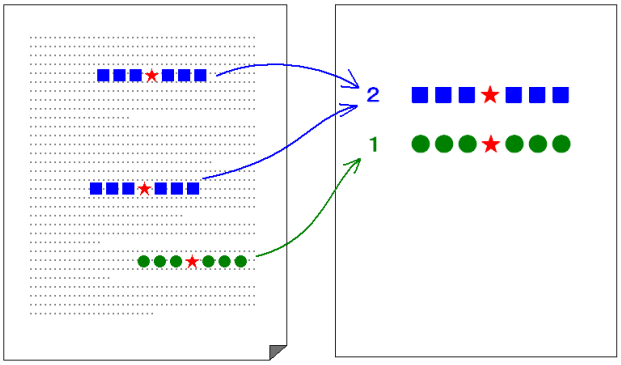



Output [Exp. reg.: #de#; orden de frecuencia]

| Suma | -3 Pl | Suma | **Foco** | Suma | +3 Pl | Suma | Suma: 6 Pl |
|---|---|---|---|---|---|---|---|
| 145 | la | 1008 | **de** | 192 | , | 334 | , |
| 142 | , | | | 181 | . | 284 | la |
| 116 | el | | | 139 | la | 235 | . |
| 83 | y | | | 117 | y | 200 | y |
| 72 | en | | | 69 | $ | 131 | el |
| 60 | a | | | 62 | que | 118 | en |

## 2.1.4. Unión

Con la función de [Union] (Unión) se calcula la frecuencia de las secuencias continuas de *n* palabras para estudiar las características colindantes de la palabra clave:




Output [Exp. reg.: #de#; orden de frecuencia]

| Sum | -3 Pl | Sum | **Foco** | Sum | +3 Pl |
|---|---|---|---|---|---|
| 9 | en_el_centro | 1008 | **de** | 9 | tema_,_¿ |
| 9 | en_la_sala | | | 9 | vez_en_cuando |
| 8 | $_(_dentro | | | 9 | espera_)_$ |
| 7 | ¿_qué_es | | | 6 | tu_vida_? |
| 7 | oye_,_cambiando | | | 5 | la_ciudad_y |

## ■ Latín tardío y español en origen

Aprendemos «latín clásico» en la universidad con textos de César, Cicerón, etc. Por otra parte, poseemos textos bíblicos traducidos del hebreo al «latín tardío» del siglo V.

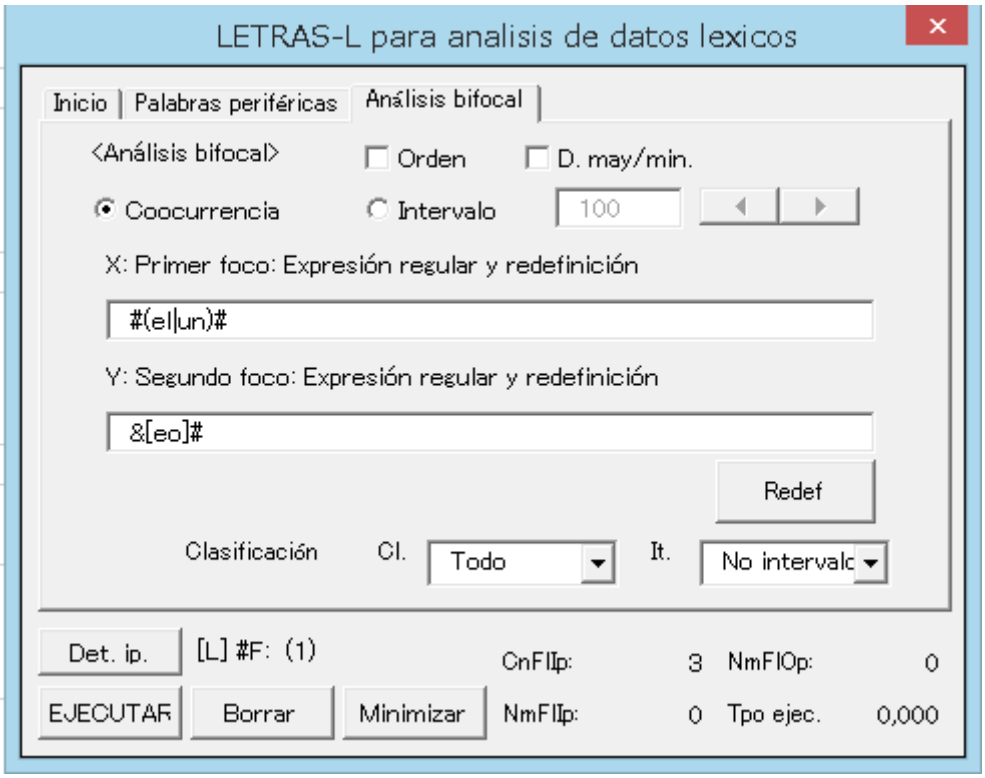
Como la lengua latina presenta unas inflexiones de caso por las cuales se determinan las relaciones gramaticales, la posición del objeto del verbo es libre, pero con tendencia general a situarse delante del verbo, y el verbo, al final de la oración. En el español actual, por otra parte, el objeto del verbo si es sustantivo suele aparecer detrás del verbo.

Hemos calculado las formas acusativas, *aquam*, *arcam*, *terram*, con las palabras inmediatamente anteriores a estas tres formas:

| SUM | W-1 | SUM | aquam<br>arcam<br>terram |
|---|---|---|---|
| 42 | super | 121 | terram |
| 38 | in | 12 | aquam |
| 8 | omnem | 11 | arcam |
| 7 | et | | |
| 6 | ad | | |
| 4 | universam | | |
| 3 | tibi | | |
| 3 | dabo | | |
| 3 | hauriendam | | |
| 2 | replete | | |
| 2 | operaretur | | |
| 1 | aridam | | |
| 1 | illuminent | | |
| 1 | deus | | |
| 1 | ingredieris | | |
| 1 | elevaverunt | | |
| 1 | aquae | | |
| 1 | dissipans | | |
| 1 | inter | | |
| 1 | exercere | | |
| 1 | abram | | |
| 1 | perambula | | |
| 1 | te | | |
| 1 | hydriam | | |
| 1 | hauriam | | |
| 1 | hauriret | | |
| 1 | hausit | | |
| 1 | reppererunt | | |
| 1 | invenimus | | |
| 1 | possideas | | |
| 1 | isaac. | | |
| 1 | habebat | | |
| 1 | sunt | | |
| 1 | attulit | | |
| 1 | eis | | |
| 1 | fames | | |
| 1 | praeter | | |

Es interesante observar la posición del objeto directo en acusativo dentro del texto del *Génesis* en latín tardío, casi sin excepción, los verbos aparecen delante de estas tres formas, lo mismo que en español actual. De esta manera, a pesar de que no poseemos textos de la lengua hablada de aquel entonces, con el texto escrito en latín tardío podemos imaginarnos de qué forma se estaba gestando un «proto-español» en aquella época.

# 3. Análisis bifocal




Con las funciones incluidas en la pestaña [Bifocal] (Análisis bifocal), intentamos averiguar el grado de relación entre dos elementos dentro del mismo contexto. Se calculan distintos tipos de ocurrencia y sus coeficientes.

## 3.1. Coocurrencia

Al seleccionar la opción de [Cooccurrencia] se calcula la frecuencia de dos elementos que ocurren en la misma celda de textos. Se ofrecen la frecuencia del primer elemento en [Frecuencia de X], la del segundo elemento en [Frecuencia de Y], la de coocurrencia en [a (++) Cooccur.], la de la frecuencia exclusiva de X en [b (+-) Exclusive (X)], la de la frecuencia exclusiva de Y en [c (-+) Exclusive (Y)], el total de las palabras en [Total (N)], la cifra de «Información mutua» en [Mutual information], el coeficiente de Dice en [Dice 2á(2a+b+c)], el de Jaccard en [Jaccard á(a+b+c)], el de Ochiai en [Ochiai á√[(a+b)(a+c)]], y finalmente el de Ueda (2013) en [Ueda (2a-b-c)/(2a+b+c)].

| Análisis bifocal: Coocurrencia | Total |
|---|---|
| Frecuencia de X: #(por\|para)# | 521 |
| Frecuencia de Y: &[aei]r# | 882 |
| a (X:+ / Y:+): #(por\|para)#.*?&[aei]r# | 395 |
| b (X:+ / Y:-): #(por\|para)# | 126 |
| c (X:- / Y:+): &[aei]r# | 487 |
| Total (N) | 26,581 |
| Información mutua | 4.5140 |
| Dice 2a/(2a+b+c) | 0.5631 |
| Jaccard a/(a+b+c) | 0.3919 |
| Ochiai a/√[(a+b)(a+c)] | 0.5827 |
| Ueda (2a-b-c)/(2a+b+c) | 0.1262 |

También es posible clasificar la frecuencia por hojas con la selección de ⌈Hoja⌉ (Hoja) en ⌈Classificatión⌉:

| Análisis bifocal: Coocurrencia | Tx1 | Tx2 |
|---|---|---|
| Frecuencia de X: #(por\|para)# | 239 | 282 |
| Frecuencia de Y: &[aei]r# | 428 | 454 |
| a (X:+ / Y:+): #(por\|para)#.*?&[aei]r# | 189 | 206 |
| b (X:+ / Y:-): #(por\|para)# | 50 | 76 |
| c (X:- / Y:+): &[aei]r# | 239 | 248 |
| Total (N) | 12,315 | 14,266 |
| Información mutua | 4.5080 | 4.5207 |
| Dice 2a/(2a+b+c) | 0.5667 | 0.5598 |
| Jaccard a/(a+b+c) | 0.3954 | 0.3887 |
| Ochiai a/√[(a+b)(a+c)] | 0.5909 | 0.5757 |
| Ueda (2a-b-c)/(2a+b+c) | 0.1334 | 0.1196 |

La figura siguiente es el resultado de la clasificación por la clave ⌈C⌉:

| Análisis bifocal: Coocurrencia | (a) Madric | (b) Sevilla | (c) Méxic | (d) Lima | (e) B.A. |
|---|---|---|---|---|---|
| Frecuencia de X: #(por\|para)# | 14 | 6 | 14 | 6 | 481 |
| Frecuencia de Y: &[aei]r# | 7 | 2 | 8 | 13 | 852 |
| a (X:+ / Y:+): #(por\|para)#.*?&[aei]r# | 13 | 2 | 12 | 6 | 362 |
| b (X:+ / Y:-): #(por\|para)# | 1 | 4 | 2 | 0 | 119 |
| c (X:- / Y:+): &[aei]r# | -6 | 0 | -4 | 7 | 490 |
| Total (N) | 410 | 255 | 494 | 418 | 25,004 |
| Información mutua | 5.7652 | 5.4094 | 5.7260 | 5.0069 | 4.4651 |
| Dice 2a/(2a+b+c) | 1.2381 | 0.5000 | 1.0909 | 0.6316 | 0.5431 |
| Jaccard a/(a+b+c) | 1.6250 | 0.3333 | 1.2000 | 0.4615 | 0.3728 |
| Ochiai a/√[(a+b)(a+c)] | 1.3132 | 0.5774 | 1.1339 | 0.6794 | 0.5655 |
| Ueda (2a-b-c)/(2a+b+c) | 1.4762 | 0.0000 | 1.1818 | 0.2632 | 0.0863 |

## 3.2. Intervalo

El coeficiente de intervalo es el resultado del cálculo de las frecuencias de los dos elementos con un intervalo de igual o menos de cuantas letras se especifiquen en la casilla [itv]:

| Análisis bifocal: Letras de intervalo100 | (a) Madric | (b) Sevilla | (c) Méxic | (d) Lima | (e) B.A. |
|---|---|---|---|---|---|
| Frecuencia de X: #(por\|para)# | 14 | 6 | 14 | 6 | 481 |
| Frecuencia de Y: &[aei]r# | 7 | 2 | 8 | 13 | 852 |
| a (X:+ / Y:+): #(por\|para)#.{0,100}?&[aei]r# | 13 | 2 | 11 | 6 | 355 |
| b (X:+ / Y:-): #(por\|para)# | 1 | 4 | 3 | 0 | 126 |
| c (X:- / Y:+): &[aei]r# | -6 | 0 | -3 | 7 | 497 |
| Total (N) | 410 | 255 | 494 | 418 | 25,004 |
| Información mutua | 5.7652 | 5.4094 | 5.6004 | 5.0069 | 4.4369 |
| Dice 2a/(2a+b+c) | 1.2381 | 0.5000 | 1.0000 | 0.6316 | 0.5326 |
| Jaccard a/(a+b+c) | 1.6250 | 0.3333 | 1.0000 | 0.4615 | 0.3630 |
| Ochiai a/√[(a+b)(a+c)] | 1.3132 | 0.5774 | 1.0394 | 0.6794 | 0.5545 |
| Ueda (2a-b-c)/(2a+b+c) | 1.4762 | 0.0000 | 1.0000 | 0.2632 | 0.0653 |

## 3.3. Distinción de orden

En la lengua latina el orden de palabras es relativamente libre. El siguiente cuadro muestra la diferencia entre los dos cálculos, uno con distinción de orden, y otro sin ella.

| Coocurrencia | Con dist. de orden | Sin dist. de orden |
|---|---|---|
| X: #domin(us\|i\|um)# | 164 | 164 |
| Y: #de(us\|i\|um)# | 200 | 200 |
| a (++) Cooccur. (XY): | 29 | 33 |

| | | |
|---|---|---|
| #domin(us¦i¦um)#+#de(us¦i¦um)# | | |
| b (+-) Exclusive(X): #domin(us¦i¦um)# | 135 | 131 |
| c (-+) Exclusive(Y): #de(us¦i¦um)# | 171 | 167 |
| Total (N) | 25,569 | 25,569 |
| Mutual information(MI) | 4.4987 | 4.6851 |
| Dice 2á(2a+b+c) | 0.1593 | 0.1813 |
| Jaccard á(a+b+c) | 0.0866 | 0.0997 |
| Ochiai á√⌈(a+b)(a+c)⌉ | 0.1601 | 0.1822 |
| Preference (2a-b-c)/(2a+b+c) | −0.6813 | −0.6373 |