Tratamiento lingüístico y matemático de textos digitales españoles. Presentación del Programa LEXIS-web

HIROTO UEDA

Universidad de Tokio

Resumen: Desde 1988 he venido desarrollando un sistema de tratamiento de textos digitales españoles en Excel para corpus lingüísticos en español. En 2013 lo amplié en un paquete de programas en web. Actualmente se encuentran en el sitio de la Universidad de Tokio, donde he reunido materiales preparados por distintos grupos de investigadores incluyendo el nuestro:

LETRAS-web: http://lecture.ecc.u-tokyo.ac.jp/~cueda/letras/
NUMEROS-web: http://lecture.ecc.u-tokyo.ac.jp/~cueda/numeros/

En esta ocasión, voy a explicar sus funciones y presentar algunos trabajos hechos con los materiales y herramientas colocados en estos sitios. Hasta el momento, se han reunido 24 corpus y elaborado 8 programas en LETRAS-web y más de 90 programas en NUMEROS-web.

Al mismo tiempo propondré un trabajo en colaboración en los estudios lingüísticos, filológicos y literarios. También es posible aplicar los programas a la enseñanza de español como lengua extranjera. En cuanto a las aplicaciones a las ciencias sociales, el tratamiento de los datos de prensa encontrados en internet está en marcha.

Palabras clave: Excel; letras-web; números-web; estudios lingüísticos

1. Introducción

Desde 1988 he venido desarrollando unos sistemas de tratamiento de textos digitales españoles en Excel para aplicarlos a los corpus lingüísticos españoles.[1] En 2013 lo amplié en un paquete de programas en web en colaboración con el equipo de la Universidad Autónoma de Madrid dirigido por Antonio Moreno Sandoval. Últimamente los he instalado en el sitio web de la misma universidad y en el de la Universidad de Tokio, donde he reunido materiales preparados por distintos grupos de investigadores españoles y japoneses, incluyendo el nuestro (Ueda, en prensa): LETRAS-web, NUMEROS-web y LEXIS-web.[2] En

[1] http://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/ El último acceso a las direcciones indicadas en este estudio ha sido en [8/1/2016]/

[2] LETRAS-web (Tokio): http://lecture.ecc.u-tokyo.ac.jp/~cueda/letras/
LETRAS-web (Madrid): http://shimoda.lllf.uam.es/letras/
NUMEROS-web (Tokio): http://lecture.ecc.u-tokyo.ac.jp/~cueda/numeros/
NUMEROS-web (Madrid): http://shimoda.lllf.uam.es/numeros/
LEXIS-web (Tokio): http://lecture.ecc.u-tokyo.ac.jp/~cueda/lexis/
LEXIS-web (Madrid): http://shimoda.lllf.uam.es/lexis/

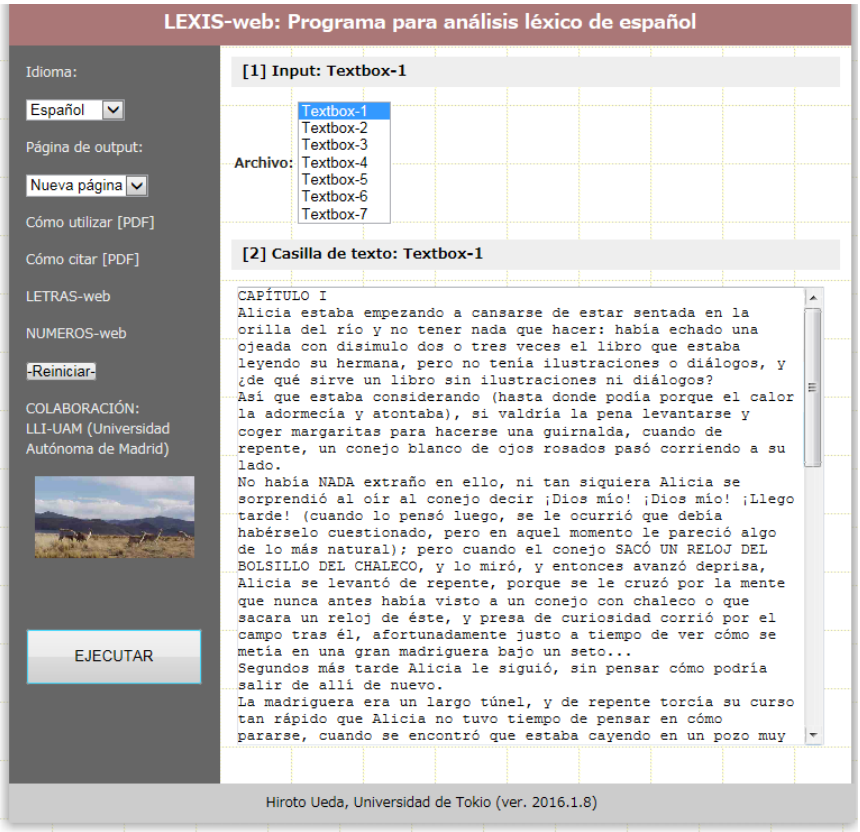
esta ocasión únicamente voy a presentar el último Programa LEXIS-web, con funciones de etiquetador gramatical, debido a la limitación de tiempo en el Congreso y espacio en estas páginas.

El objetivo principal de esta aplicación es ofrecer a los investigadores de lingüística y filología españolas una herramienta para análisis léxicos que, al recibir los datos textuales, devuelve las informaciones gramaticales de cada palabra: palabra separada, informaciones gramaticales (categoría sintáctica; género y número; modo, tiempo y persona), lema (forma representativa) e informaciones de desambiguación (máxima secuencia de tres categorías sintácticas, su frecuencia y probabilidad).

2. Interfaces de input y outpu

La figura siguiente muestra el interfaz inicial de LEXIS-web, donde en la zona lateral izquierda se puede hacer la selección de [Idioma], [Página de output]; y en la zona principal, [1] [Input] de siete textos y la [2] [Casilla de texto], donde se puede escribir directamente o pegar textos españoles copiados:[3]

Figura. 1. Input de LEXIS-web



Al pulsar el botón de [EJECUTAR], que se encuentra en la parte inferior de la zona izquierda, el Programa presenta la página siguiente:

[3] A modo de ejemplo. he utilizado la traducción española de *Alicia en el país de maravillas* (Lewis Carroll), preparada por María Dolores Murillo y Ana Isabel García en nuestro proyecto conjunto de ELE, que he colocado en LETRAS-web (nota 1).

Figura. 2. Output de LEXIS-web

LEXIS-web: Programa para análisis léxico de español

Input: Textbox-1; Tiempo de ejecución: 0.860 s. /
Líneas de output: 817 /

Op	Palabra	Palabra.C.S.	Lema	Lema.C.S.	N.P.	Máx.secuencia	Frec.	Prob.	Ip
1	capítulo	Sus:ms	capítulo	Sus	1: Sus	{Sus}-Num	12	1.00	1
2	i	Num	i	Num	1: Num	Sus-{Num}	12	1.00	1
3	Alicia	Xant	Alicia	Xant	1: Xant	{Xant}	265	1.00	2
4	estaba	ESTAR:IndImp13	estar	ESTAR	1: ESTAR	{ESTAR}-Ger-Prep	8	1.00	2
5	empezando	Ger	empezar	Inf	1: Ger	ESTAR-{Ger}-Prep	8	1.00	2
6	a	Prep	a	Prep	1: Prep	{Prep}-Inf-Clit	143	1.00	2
7	cansar	Inf	cansar	Inf	1: Inf	Prep-{Inf}-Clit	143	1.00	2
8	+se	Clit:3	se	Clit	1: Clit	Prep-Inf-{Clit}	143	1.00	2
9	de	Prep	de	Prep	1: Prep	Inf-Clit-{Prep}	54	1.00	2
10	estar	ESTAR	estar	ESTAR	1: ESTAR	{ESTAR}-PP-Prep	5	1.00	2
11	sentada	PP:fs	sentar	Inf	1: PP	{PP}-Prep-L	30	1.00	2
12	en	Prep	en	Prep	1: Prep	{Prep}-L-Sus	764	0.98	2
13	la	L:fs	el	L	2: L; Clit	Prep-{L}-Sus	764	0.98	2
14	orilla	Sus:fs	orilla	Sus	2: Sus; V	Prep-L-{Sus}	764	0.99	2

donde se encuentran:

Op: Número secuencial de output

Palabra:Palabra separada del texto, por ejemplo *del* se separa en *de +el*.

Palabra C. S.: Categoría sintáctica de la palabra con informaciones gramaticales, por ejemplo, en el caso de *empezando,* «Ger» es abreviación de gerundio.

Lema: Forma representante, por ejemplo el lema de *empezando* es *empezar*, es decir, la forma canónica de una entrada de un diccionario.

Lema C. S.: Categoría sintáctica del lema: *empezar* «Inf», que es infinitivo

N. P.: Número y categorías posibles, por ejemplo *la* presenta dos posibles soluciones, que son «L» (forma femenina singular del artículo *el*) y «Clit» (forma femenina singular del clítico *lo*).

Máx(ima) secuencia: La secuencia de las tres palabras continuas que ha dado la máxima frecuencia. Cuando ninguna secuencia de las tres da una frecuencia positiva, se calcula la de las dos palabras. Cuando tampoco da la frecuencia, se calcula la de una palabra.

Frec.: Frecuencia correspondiente.

Prob.: Probabilidad dentro de las combinaciones calculadas en el mismo contexto.

Ip.: Número secuencial de línea de input.

3. Identificación léxica

He preparado la lista de lema más categoría sintáctica con informaciones adicionales correspondientes de unos 38.000 vocablos.[4] Por ejemplo:

[4] He utilizado los materiales de Ueda y Rubio (2006).

Tabla 1

caja	Sus:fs
cajamarca	Xtop
cajero	Sus:ms
(…)	(…)

Donde los lemas en minúscula se colocan en la columna izquierda y sus correspondientes informaciones en la columna derecha. «Sus: fs» representa 'sustantivo: femenino singular'; «Xtop» es 'topónimo' (nombre propio de lugar); «ms» es 'masculino singular'. La forma minúscula del topónimo «cajamarca» se convierte posteriormente en la mayúscula «Cajamarca». El programa recoge estas correspondencias en una memoria de acceso inmediato.

De esta manera las informaciones gramaticales vienen en forma abreviada, por ejemplo, «Sus», «Xtop», «fs», etc. Por lo tanto es necesario saber de antemano qué significa cada abreviación, aunque la mayoría de las veces es fácil imaginarse de qué se trata. A continuación, damos una lista completa de la abreviación (Abr.)

Tabla 2:

Abrev.	**Explicación**	**Ejemplo**
Adj	Adjetivo	*alto, interesante*
Adv	Adverbio	*abajo*
Clit	Clítico	*me, te, se, …, lo, le, …*
Comp	Comparativo	*más, menos*
Conj	Conj	*aunque, como, …*
Det.dem	Determinante demostrativo	*este, ese, aquel*
Det.ind	Det.indefinido	*algún*
Det.pos	Det.posesivo	*mi, tu, su, …*
ESTAR	Verbo *estar*	*estar*
Ger	Gerundio	*estando*
HABER	Verbo *haber*	*haber*
Inf	Verbo en infinitivo	*dar*
Int	Interjección	*hola, adiós, …*
L	Artículo definido EL	*el (los, la, las, lo)*
Num	Numeral	*0, 1, 2, …, uno, dos, …, i, ii, …*
Paren	Paréntesis	*( ) < > { } [ ] « »*
PP	Participio pasado	*estado*
Prep	Preposición	*a*
Pro.dem	Pronombre demostrativo	*aquel*
Pro.ind	Pronombre indefinido	*algo*
Pro.pers	Pronombre personal	*él*

Pro.prep	Pronombre prepositivo	*mí, ti, sí*
Punt	Puntuación	*. , : ; - ¿ ? ¡ !*
Q.adj	Interrogativo adjetival	*cuál*
Q.adv	Interrogativo adverbial	*cómo*
Q.pro	Interrogativo pronominal	*cuál*
Rel.adj	Relativo adjetival	*cuanto*
Rel.adv	Relativo adverbial	*cuando*
Rel.pro	Relativo pronominal	*cual*
S|N	Sí o no	*sí, no*
SER	Verbo *ser*	*ser, soy, eres, es, …*
Signo	Signo	*#, $, %, &, +, -, =, *, /, …*
Sus	Sus	*hombre, mujer, animal*
U	Artículo definido UN	*un (una, unos, unas)*
Xant	Xant	*abraham*
Xtop	Xtop	*cajamarca*
Y|O	Y|O	*y (e), o (ó, u)*

Dentro de nuestro Diccionario informático, los miembros de «Sus», «Inf», «Adj», etc. pertenecen al grupo abierto, cuyos miembros son ilimitados, mientras que los de «Clit», «Comp», «Det»,[5] «Num», «Paren», «Prep», «Punt», «Q», «Rel», «S|N», «Signo» al grupo cerrado con miembros limitados. Por otra parte hay 'grupos' unimembres: «ESTAR», «HABER», «SER», «L», «U»; y los bimembres: «S|N», «Y|O». El grupo de artículo definido «L», por ejemplo, es unimembre con la única forma *el*, puesto que otras formas, *los, la, las, lo,* se derivan por las reglas automáticas de inflexión, de que voy a tratar más adelante. Esta sistematización obedece a razones prácticas de caracterización gramatical con métodos automáticos, estadísticos y distribucionales. Por ejemplo, al independizar los tres verbos «ESTAR», «HABER», «SER», verbos altamente gramaticalizados con sintaxis peculiar, se hace la desambiguación más eficaz.[6]

4. Separación de palabras

Utilizando las informaciones léxicas ofrecidas por el Diccionario, el Programa procede a analizar los textos objeto de la etiquetación gramatical. El primer trabajo que hay que hacer para identificar la unidad léxica es separar las formas unidas, por ejemplo: *al, del, verte, pidiéndomelo,* etc. en *a +el, de +el, ver +te, pidiendo + me + lo*. Para llevarlo a cabo he preparado una lista de los posibles cambios necesarios. Reproducimos la parte inicial de la lista de las reglas de separación:

Tabla 3

KEY	ITM
al	a +el/a/Prep

[5] Para la categoría gramatical de Determinante sigo básicamente a Jiménez Julia (2006). Véase Ueda (en prensa).
[6] Desde luego las gramáticas teóricas proponen otros principios y las entradas que hay en los diccionarios normales siguen otras reglas teóricas y/o prácticas.

ándola	ando +la/ar/.
ándolas	ando +las/ar/.
ándole	ando +le/ar/.
ándoles	ando +les/ar/.
ándolo	ando +lo/ar/.
(...)	(...)
rte	r +te/r/Inf

cuya totalidad se almacena en un archivo de acceso inmediato. Cuando el Programa LEXIS encuentra la forma *al* (KEY), en alguna parte del texto, se intenta separarla en *a +el* (la segunda parte de ITM) y la convierte tentativamente en *-a* y consulta el archivo del Diccionario, para encontrar la información gramatical que viene en el Diccionario: «Prep». Si la información gramatical de la forma tentativa coincide con la del Diccionario, la forma tentativa se vuelve la definitiva, es decir: *a* «Prep» +el «L».

Este sistema aparenta ser complicado sin necesidad, puesto que la forma *al* no tiene otra interpretación gramatical, dejando al lado el latinismo *et al*. Sin embargo, es necesario para formular las reglas generales que tratan todas las formas unidas posibles, por ejemplo, *cantándola, tomándola, respetándola*, etc. que se convierte en *cantando +la, tomando +la, respetando +la.* Ahora no se trata de una totalidad de la palabra, sino una parte de ella. El programa convierte *-ándola* en *-ando +la*, tentativamente, y para buscar la información gramatical de la forma infinitiva la convierte, al mismo tiempo, en *-ar* y de esta manera *cantándola* llega a ser la forma *cantar*, con la que se hace el cotejo con el Diccionario. Esta vez se hace con el signo de un punto (.), que quiere decir cualquier una letra, de modo que se aplica la regla sin limitación. La ilimitación de condición es debida a que prácticamente no existe ninguna forma que termine en *-ándola*.

El caso de *-rte* es distinto, puesto que puede ser tanto la parte final de la combinación de infinitivo + *te*, v. gr. *enviarte, mandarte,* como simplemente la de un sustantivo, por ejemplo, *arte, parte,* y también de una forma conjugada, *imparte, comparte,* etc. La última regla de separación pone la condición de «Inf» para evitar la división equivocada en *-ar +te, par +te, impar +te, compar +te,* etc.

5. Lematización y asignación gramatical

A las palabras separadas el Programa asigna las categorías sintácticas (Sustantivo «Sus», Adjetivo «Adj», Verbo «V», etc.) con sus informaciones inflexivas (Género, Número, Modo, Tiempo, Persona), con las abreviaciones siguientes (Abrev.):

Tabla 4

Abrev.	**Explicación**	**Ejemplo**
«ms»	masculino singular	*libro*
«mp»	masculino plural	*ambos*
«cs»	común singular	*estudiante*
«V»	Verbo conjugado	*voy, comeremos*
«PP»	Participio pasado	*ido, comido*
«Ger»	Gerundio	*yendo, comiendo*
«Ind»	Indicativo	*sé, sabes*
«Sub»	Subjuntivo	*sepa, sepas*

«Fut»	Futuro	*sabré*
«Cond»	Condicional	*sabría*
«Pres»	Presente	*sé, sepa*
«Imp»	Imperfect	*sabía*
«Pas»	Pasado	*supe, supiera*
«1»	Primera persona	*yo, sé*
«2»	Segunda persona	*tú, sabes*
«3»	Tercera persona	*usted, él, ella, sabe*
«4»	Cuarta persona	*nosotros, nosotras, sabemos*
«5»	Quinta persona	*vosotros, vosotras, sabéis*
«6»	Sexta persona	*ustedes, ellos, ellas, saben*

Para asignar estas informaciones inflexivas a cada palabra en el texto, he preparado una lista de reglas morfológicas de manera siguiente:

Tabla 5

KEY	ITM
a	/(Adj).*/$1:fs#ó(Adj|Det.pos|Pro.ind|Rel|Q).*/$1:fs#ar/(Inf:v|Inf:r)/V:IndPres3 #er/(Inf:v|Inf:r)/V:SubPres13#ir/(Inf:v|Inf:r)/V:SubPres13
á	/Inf/V:Fut3
aba	ar/(Inf|ESTAR)/V:IndImp13
abais	ar/(Inf|ESTAR)/V:IndImp5
ábamos	ar/(Inf|ESTAR)/V:IndImp4
(...)	(...)

La primera regla aplicada a todas las formas que terminan en *-a* es una de las más complejas que hay en la lista, de modo que desglosamos la parte izquierda (ITM) por el separador (#):

(1) /(Adj).*/$1:fs
(2) o/(Adj|Det.pos|Pro.ind|Rel|Q).*/$1:fs
(3) ar/(Inf:v|Inf:r)/V:IndPres3
(4) er/(Inf:v|Inf:r)/V:SubPres13
(5) ir/(Inf:v|Inf:r)/V:SubPres13

Por el separador (/) se divide en tres componentes [1], [2], [3] en [1] / [2] / [3]. El primer componente [1] es la forma que sustituye al objeto, [2] es la condición gramatical que permite la sustitución, y [3] es la asignación flexiva. LEXIS-web aplica la primera regla (1), por ejemplo, a *española;* la *-a* final la sustituye por nulidad en forma de *español;* consulta la información gramatical de *español* que hay en el Diccionario; la coteja con «Adj» de la lista de reglas flexivas. Al comprobar la correspondencia afirmativa, pone tentativamente «fs», cambiando «ms» que viene en el Diccionario. Este cambio se realiza por la Expresión Regular "(Adj).*" por "$1:fs"; es decir, la parte entre paréntesis (Adj.) se reproduce por "$1", y cualquier secuencia ".*" se convierte en ":fs". Por esta expresión regular "Adj:ms" se convierte en "Adj:fs". De esta

manera, se hacen dos operaciones al mismo tiempo: lematización y asignación gramatical (Gómez Díaz 2005).

En la segunda regla (2) hay formas optativas (Adj|Det.pos|Pro.ind|Rel|Q), que significa que la regla del cambio de *-a* en *-o* se aplica a una de estas 5 categorías sintácticas. Las reglas (3), (4) y (5) convierten la forma terminada en *-a* en los verbos de *-ar, -er, -ir*, en formas de tercera persona de presente de indicativo («IndPres1»), primera y tercera persona presente de subjuntivo («SubPres13»), respectivamente.

5. Desambiguación

En el Diccionario se presenta con bastante frecuencia una pluralidad de asignación gramatical, por ejemplo, *que* puede ser tanto «Conj», en *Creo que compró ...*, como «Rel.pro» en *el libro que compró....* También en el texto se encuentran multitud de homógrafos flexivos, por ejemplo *pienso* como primera persona de presente de indicativo del verbo *pensar (yo pienso ...)*, y de sustantivo masculino singular *pienso (el pienso)*. Una de las posibles soluciones para desambiguar estos casos es utilizar los contextos inmediatos anterior y posterior. Si viene *que* detrás de un verbo, existe la alta probabilidad de ser conjunción: «Conj». Y la misma *que* detrás de sustantivo es probable que sea relativo pronominal: «Rel.pro».

No obstante la interpretación gramatical no se hace siempre con los contextos simples inmediatos, anterior y/o posterior. Pensemos el caso de, por ejemplo, *pienso que*. Es natural que lo interpretemos como una combinación de «V» + «Conj», pero también existe la posibilidad de tratarse de un «Sus» + «Rel»: *el pienso que compró ayer*. De esta manera conviene ampliar el contexto anterior hasta dos palabras más, es decir, un contexto anterior bimembre: *el pienso*.

Para saber la frecuencia de dos palabras anteriores y posteriores, es necesario analizar unos textos de cierta longitud correctamente anotados. Por ejemplo, del texto "Este es el pienso que compró ayer" se obtiene la secuencia de categorías sintácticas: «Pro.dem» - «SER» - «L» - «Sus» - «Rel.pro» - «V» - «Adv», de la cual el Programa extrae las secuencias trimembres sucesivamente: «Pro.dem» - «SER» - «L», «SER» - «L» - «Sus», «L» - «Sus» - «Rel.pro», «Sus» - «Rel.pro» - «V», y «Rel.pro» - «V» - «Adv»; y cuenta las veces que ocurre cada secuencia trimembre en todo el texto más o menos grande para obtener la lista de frecuencia de la forma siguiente (en orden descendiente de frecuencia):

Tabla 6

Punt-Punt-Punt	1120
Prep-L-Sus	764
Sus-Punt-Punt	732
L-Sus-Punt	543
Punt-Sus-Punt	470
L-Sus-Prep	350
(…)	(…)

Las secuencias trimembres de palabras las aplicamos tanto en la parte anterior, como en la posterior, y también en la central. Nuestra idea es ver las tres secuencias, anterior, central y posterior, para determinar la asignación gramatical de *que*: en secuencia anterior: *el pienso {que};* en secuencia central: *pienso {que} compró;* en secuencia posterior: *{que} compró ayer.* Al considerar la secuencia anterior *el pienso {que},* inicialmente el Programa ofrece la posibilidad múltiple de «L - Sus ~ V - Rel.pro ~ Conj». Si la frecuencia

que figura en la lista anterior de «L - Sus - {Rel.pro}» es más alta que otras posibles combinaciones, «L - Sus - {Conj}», «L -V - {Rel.pro}», «L -V - {Conj}», deberíamos pensar que con más probabilidad se trata «L - Sus - {Rel.pro}».

E incluso podemos ampliar los términos de comparación en la secuencia central *pienso {que} compró*: «Sus ~ V - Rel.por ~ Conj - V», y también en la secuencia posterior *{que} compró ayer*: «Rel.por ~ Conj - V - Adv». Dentro de las 10 posibilidades (4 + 4 + 2 = 10), la que ofrece la mayor frecuencia es: «L - Sus - Rel.pro».

Sin embargo, creo que no se trata solo de buscar la mayor frecuencia dentro de todas las combinaciones posibles. De acuerdo con nuestro sentido lingüístico común, creo que conviene buscar la mayor probabilidad entre los tres pares de cada contexto. Ahora veámosla con frecuencias concretas que hay en la lista anterior de secuencias (las cifras entre paréntesis):

(1) Secuencia anterior: *el pienso {que}*: (58)
«L - Sus - {Rel.pro}» (54)
«L - Sus - {Conj}» (4)
«L - V - {Rel.pro}» (0)
«L - V - {Conj}» (0)
(2) Secuencia central: *pienso {que} compró*: (105)
«Sus - {Rel.pro} - V» (82)
«Sus - {Conj} - V» (23)
(3) Secuencia posterior: *{que} compró ayer*: (16)
«{Rel.pro} - V - Adv» (2)
«{Conj} - V - Adv» (14)

Ahora bien, la ratio de «L - Sus - {Rel.pro}» (54) entre las cuatro posibles combinaciones es 54 / (54 + 4 + 0 + 0) = .931; mientras que la de «Sus - {Rel.pro} - V» es 82 / (82 + 23) = .780, y la de «{Conj} - V - Adv» es 14 / (2 + 14) = .875. Por este cálculo nos inclinamos a pensar que la combinación de «L - Sus - {Rel.pro}» (.931) es más importante que la de «Sus - {Rel.pro} - V» (.781), a pesar de su valor inferior (54) con respecto a «Sus - {Rel.pro} - V» (82).

Sin embargo si, por ejemplo, «{Rel.pro} - V - Adv» tuviera una frecuencia mínima de 1, en lugar de la actual 2, la ratio de «{Conj} - V - Adv» sería 14 / (14 + 1) = .933, superior al caso de «L - Sus - {Rel.pro}» (.931). Por su simple mayoría de la ratio (.933 > .931), ¿admitiríamos que la forma *que* es una conjunción «Conj»? Creemos que no, puesto que ahora no se trata solo de la frecuencia absoluta, ni de la relativa, que ambas pueden conducir a la conclusión equivocada. Por esta razón, propongo utilizar un valor absoluto relativizado, que denomino «frecuencia ponderada» (FP), que es la frecuencia absoluta (FA), relativizada por la frecuencia relativa (FR) por medio de la multiplicación (Ueda 2015):

$$FP = FA * FR$$

Por ejemplo, la frecuencia de 14 entre 15 no cobra más importancia que 54 entre 58, a pesar de su mayor valor de FR ([14 / 15] = .933 > [54 / 58] = .931). En cambio la FP de 14 entre 15 es 14 * 14 / 15 = 13.067, mientras que la de 54 entre 58 es 54 * 54 / 58 = 50.276. De esta manera utilizando la FP se invierte

el orden de la importancia (13.067 < 50.276). Defendemos que, para comparar la importancia o el grado de contribución cuantitativa, es conveniente utilizar la FP más que la FA y que la FR.[7]

El programa de LEXIS-web calcula la FP en todas las combinaciones posibles de cada contexto, anterior, central y posterior, para buscar el caso más importante o contribuyente y ofrece la resolución de la mayor probabilidad posible. Cuando no se encuentra la secuencia trimembre, se busca la bimembre. Cuando no se encuentra la secuencia bimembre tampoco, se busca la unimembre.

El programa no tiene en cuenta la secuencia más allá de la Puntuación «Punt». Por ejemplo, al encontrar «V - Punt - L» (*viene. {El}*), salta el cálculo de la frecuencia de la combinación de los tres categorías y llega directamente a la de los dos: «Punt - L» (. *{El}*), puesto que el contexto más allá de la Puntuación no es relevante. El resultado que ofrece el Programa LEXIS es:

Palabra	**Palabra.C.S.**	**Lema**	**Lema.C.S.**	**N.P.**	**M.secuencia**	**Frec.**	**Prob.**
el	L:ms	el	L	1: L	{L}-Sus-Rel.pro	54	0.93
pienso	Sus:ms	pienso	Sus	2: Sus; V	L-{Sus}-Rel.pro	54	0.93
que	Rel.pro	que	Conj#Rel.pro	2:Conj; Rel.pro	Sus-{Rel.pro}-V	82	0.78
compró	V:IndPas3	comprar	Inf	1: V	Sus-Rel.pro-{V}	82	1.00
ayer	Adv	ayer	Adv#Sus	2: Adv; Sus	Rel.pro-V-{Adv}	2	0.50

6. Final

Desde el punto de vista de lingüística teórica, las técnicas estadísticas e informáticas que acabo de explicar pueden parecer no esencial sino más bien trivial. A nadie se le ocurriría la idea de calcular las frecuencias de todas las posibles secuencias de categorías sintácticas para obtener la información gramatical, puesto que el hablante instruido de español conoce su clasificación y estructura gramatical de inmediato, sin depender de estos cálculos costosos y complejos. El ordenador, sin embargo, no conoce la gramática española y procesa el texto sin hacer caso de su semántica. Por ello, el programador humano intenta codificar algunos algoritmos para que la máquina pueda analizar con un acierto cercano al 98%.

Para conseguir la mayor rapidez posible de procesamiento, he utilizado las dimensiones asociativas, es decir, de acceso nominal directo, para las listas de reglas de separación, de diccionario, de asignación - desambiguación y de secuencias trimembres. El mérito del Programa equipado de memorias inmediatas consiste en que lo hace con los datos voluminosos en tiempo breve. En realidad, en la lingüística teórica se supone que las reglas se aplican de manera sucesiva y recursiva. Personalmente, en las versiones anteriores de LEXIS, he adoptado el método de aplicación sucesiva y recursiva de reglas, y resulta que el coste de tiempo ha sido enorme. En cambio, si aplicamos la lista de reglas en archivo de acceso inmediato, la solución

[7] Para el detalle de la frecuencia ponderada, véase la Addenda.

ha sido instantánea. Por ejemplo, estas páginas contienen unas 8.000 unidades léxicas y ahora LEXIS-web las ha analizado en 4 segundos con 325 milisegundos.[8]

La tabla producida de lemas acompañados de informaciones gramaticales y datos cuantivativos es útil para estudios cuantitativos de vocablos.[9] En el tema de gramaticalización, el conocimiento de la frecuencia léxica es fundamental.[10] También las frecuencias de vocablos deben ser consideradas en las obras lexicográficas. Los estudios estadísticos de la lengua dependen de los cálculos anteriormente realizados. Para estudios sociológicos de prensa, por ejemplo, se busca tanto los nombres propios de persona y lugar como los lemas claves de conceptos en cuestión: *protesta, protestó, protestaron*, etc. reunidos en el lema *protestar*. En los trabajos prácticos de redacción se desea un programa de corrección que considere no solamente la forma misma −por ejemplo *él* por sí es una forma correcta−, sino también su frecuencia y, especialmente, su asignación gramatical en su contexto: *él* delante de *pienso* es dudosov

Por consiguiente en la historia de lingüística general la lematización y sus aspectos cuantitativos no han dejado de ser puntos de interés de los investigadores. Recordamos que en la primera mitad del siglo pasado las formas verbales en tiempo y modo en textos españoles fueron recontadas una por una manualmente por el grupo de Bull (1947). Y, 70 años después, en la actualidad contamos con los grandes proyectos de equipos de investigación informática: «TreeTagger»,[11] «FreeLing»,[12] «Grampal»,[13] y «GEDLC».[14]

En cambio, el Programa LEXIS-web es un producto individual, hecho con intereses personales. Lo he elaborado porque lo necesito para tratar los textos españoles a mi manera peculiar. Al mismo tiempo, he pensado que para el uso general una herramienta informática debe ser sencilla de manejo y rápida de ejecución. Por lo tanto me quedan trabajos de mejora de funciones, de ampliaciones de aplicación y, sobre todo, de divulgación tanto en el ámbito particular como en el académico. Para todo esto agradecería que los usuarios me comunicaran sus opiniones y sugerencias.

Agradecimiento

Agradezco de todo corazón la ayuda prestada por Antonio Moreno Sandoval (Universidad Autónoma de Madrid) para terminar este trabajo, tanto en la recogida de informaciones relevantes como en la redacción de estas páginas. Este trabajo ha sido subvencionado por JSPS KAKENHI Grant Number 24520453.

Referencias bibliográficas

Almela, R., Cantos, P., Sánchez, A., Sarmiento, R., Almela, M. (2005). *Frecuencias del español. Diccionario y estudios léxicos y morfológicos.* Madrid : Universitas.

[8] Para el contraste entre el método generativo y el probabilístico, véase Moreno Sandoval (2014), cuya idea expuesta anteriormente en su ponencia en Tokio me ha influenciado para realizar la programación de LEXIS.
[9] Véanse: García Hoz (1953), Juilland and Chang-Rodríguez (1964), Ávila Muñoz (1999), Almena, Cantos, Sánchez, Sarmiento y Almela (2005) y Davies (2006), entre otros. Véase también el estudio de Moreno Sandoval y Guirao Miras (2008).
[10] Veánse Bybee (2003), Lieberman, Michel, Tina Tang y Nowak (2007), Pagel, Atkinson y Meade (2007), Hopper y Traugott (2003), Company (2004). Para el argumento basado en el texto de Don Quijote, véase Ueda (en prensa).
[11] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[12] http://nlp.lsi.upc.edéfreeling/index.php
[13] http://www.lllf.uam.es/ESP/Grampal.html
[14] http://www.gedlc.ulpgc.es/investigacion/desambigua/morfosintactico.htm

Ávila Muñoz, A.M. (1999). *Léxico de frecuencia del español hablado en la ciudad de Málaga.* Málaga: Universidad de Málaga.

Bull, W. E. (1947). "Modern Spanish verb-form frequencies", *Hispania*, 451-466.

Bybee, J. (2003). "Mechanisms of change in grammaticalization: the role of frequency", en Brian D. Joseph and Richard D. (eds.), *The Handbook of historical linguistics.* Oxford: Blackwell, 602-623.

Company Company, C. (2004). "¿Gramaticalización o desgramaticalización? Reanálisis y subjetivización de verbos como marcadores discursivos en la historia del español", *Revista de Filología Española*, 84, 29-66.

Davies, M. (2006). *A frequency dictionary of Spanish. Core vocabulary for learners.* New York: Routledge.

García Hoz, V. (1953). *Vocabulario usual, vocabulario común y vocabulario fundamental.* Madrid: Consejo Superior de Investigaciones Científicas.

Gómez Díaz, R. (2005). *La lematización en español. Una aplicación para la recuperación de información.* Gijón: Trea.

Hopper, P. J., Traugott, E. C. (2003). *Grammaticalization,* 2nd ed. Cambridge: Cambridge University Press.

Jiménez Juliá, T. (2006). *El paradigma determinante en español. Origen nominativo, formación y características. Verba,* anexo 56, Santiago de Compostela: Universidade de Santiago de Compostela .

Juilland, A., Chang-Rodríguez, E. (1964). *Frequency dictionary of Spanish words.* The Hague: Mouton .

Lieberman, E., Michel, J. B., Jackson, J., Tang, T., Nowak Martin A. (2007). "Quantifying the evolutionary dynamics of language", *Nature,* vol. 449, 713-716.

Moreno Sandoval, A. (2014). "Desafíos de y para la lingüística de corpus", *Estudios Lingüísticos Hispánicos*, (Círculo de Estudios Lingüísticos Hispánicos de Tokio) 29, 69-85.

Moreno Sandoval, Antonio / Guirao Miras, José María. (2008). "Frecuencia y distintividad en el uso lingüístico: casos tomados de la lematización verbal de corpus de distintos registros", *Actas del I Congreso Internacional de Lingüística de Corpus (CILC-09)*, Murcia: Universidad de Murcia. 195-210.

Pagel, M., Atkinson, Q. D., Meade A. (2007). "Frequency of word-use predicts rates of lexical evolution throughout Indo-European history", *Nature,* 449, 717-720.

Ueda, H. (2015). "Frecuencia contrastiva, frecuencia ponderada y método de concentración. Aplicación al estudio de las dos formas prepositivas del español medieval «pora» y «para»", *Actas del IX Congreso Internacional de Historia de la Lengua Española (Cádiz, 2012)*, Madrid: Iberoamericana, 1139-1155.

Ueda, H. (en prensa). "Analizador lingüístico común con reglas gramaticales y diccionario, preparados por el usuario: Una aplicación para el análisis tipológico del léxico español".

Ueda, H., Perea M. P. (2010). "Método general de lematización con una gramática mínima y un diccionario óptimo. Aplicación a un corpus dialectal escrito", en Moskowich-Spiegel Fandiño, I; Crespo García, B.; Lareo Martín, I.: Lojo, P. (eds.) *Visualización del lenguaje a través de corpus*. A Coruña: Universidade da Coruña , 919-932, .

Ueda, H., Rubio, C. (2006). *Puerta al español. Nuevo diccionario español-japonés.* Tokio: Kenkyusha.

ADDENDA: Experimento de la frecuencia ponderada

En la sección 5 he propuesto utilizar la FP (frecuencia ponderada), en lugar de la FR (frecuencia relativa) para medir la importancia de las cifras. A continuación voy a presentar el resultado de un pequeño experimento con cifras sencillas:

x	a	y	b	FR(x)=x/a	FR(y)=y/b	FP(x)=x*FR(x)	FP(y)=y*FR(y)	D=FP(y)-FP(x)
1	10	**10**	100	.100	.100	.100	1.000	**.900**
1	10	**9**	100	.100	.090	.100	.810	**.710**
1	10	**8**	100	.100	.080	.100	.640	**.540**
1	10	**7**	100	.100	.070	.100	.490	**.390**
1	10	**6**	100	.100	.060	.100	.360	**.260**
1	10	**5**	100	.100	.050	.100	.250	**.150**
1	10	**4**	100	.100	.040	.100	.160	**.060**
1	10	**3**	100	.100	.030	.100	.090	**-.010**
1	10	**2**	100	.100	.020	.100	.040	**-.060**
1	10	**1**	100	.100	.010	.100	.010	**-.090**
1	10	**0**	100	.100	.000	.100	.000	**-.100**

donde FR (x) = x / a, es decir, en la primera fila, 1 / 10; FR(y) = y / b = 10 / 100. He apuntado que la FR (frecuencia relativa) es engañosa para hacer la comparación, puesto que en este caso resulta igual: F (x) = FR(y) = .100, a pesar de que tenemos la impresión de que 10 entre 100 es más importante que 1 entre 10. Por ejemplo, la contribución de un futbolista que ha metido 10 goles en 100 partidos creemos que es mayor que la del otro que ha metido un gol en 10 partidos. Para medir el grado de importancia he propuesto relativizar la frecuencia absoluta (FA), multiplicada por la frecuencia relativa (FR), para obtener la cifra de la frecuencia ponderada (FP): FP = FA * FR = x * FR.

Ahora bien, buscamos el grado de contribución del futbolista que ha participado solo en 10 partidos, a escala de 100 partidos. Para llegar a la cifra correspondiente, he disminuido uno por uno, de 10 a 0 (y), de los goles de 100 partidos. La última columna (D) representa la diferencia

entre FP(x) y FP(y). Ahí se llega al punto 0, es decir, la no diferencia entre FP(x) y FP(y), en la zona entre 4 y 3 goles (y). Esto significa que la contribución que hace 1 gol en 10 partidos iguala a la de entre 4 y 3 en 100 partidos.

Para obtener la cifra exacta, sin hacer el experimento, vamos a formular una ecuación: FP(x) = FP(y); y de ahí » x * FR(x) = y * FR(y) » x * x / a = y * y / b » x^2 / a = y^2 / b » b x^2 = a y^2 » y^2 = x * b / a » y = x $\sqrt{b/a}$ » y = 1 * $\sqrt{100/10}$ = 3.162.