Analizador lingüístico común
con reglas gramáticales y diccionario, preparados por el usuario

– Una aplicación al análisis tipológico del léxico español –

Hiroto Ueda (Universidad de Tokio)

0. Introducción

(1) forma textual

(2) información gramatical de la forma textual

(3) lema

(4) información gramatical del lema

Cervantes, *Don Quijote de la Mancha*

agilización del procesamiento

lexicoestadística española

1. Elaboración del Analizador

1.1. Entrada y salida

Visual Basic for Applicacion (VBA) de Microsoft incorporado en Excel

Don Quijote de la Mancha, Miguel de Cervantes	**Parte**	**Cap.**	**Orac.**
<Que trata de la condición y ejercicio del famoso hidalgo don Quijote de la Mancha>	I	1	1
En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor.	I	1	2
Una olla de algo más vaca que carnero, salpicón las más noches, duelos y quebrantos los sábados, lentejas los viernes, algún palomino de añadidura los domingos, consumían las tres partes de su hacienda.	I	1	3

Cuadro 1.1. a. Texto objeto de etiquetación

14,522 líneas

Forma Ocurrida (F.O.)	**Etiqueta de F.O.**	**Forma Representante (F.R.)**	**Etiqueta de F.R.**	**Parte**	**Cap.**	**Orac.**
<	B	<	B	I	1	1
que	R	que	R	I	1	1
trata	Vip3j	tratar	Vn	I	1	1
de	P	de	P	I	1	1
la	T	la	T	I	1	1
condición	Sfs	condición	Sfs	I	1	1
y	C	y	C	I	1	1
ejercicio	Sms	ejercicio	Sms	I	1	1
de	P	de	P	I	1	1

el	T	el	T	I	1	1
famoso	Ams	famoso	Ams	I	1	1
hidalgo	Sms	hidalgo	Sms	I	1	1
don	Sms	don	Sms	I	1	1
Quijote	X	quijote	X	I	1	1
de	P	de	P	I	1	1
la	T	la	T	I	1	1
Mancha	X	mancha	X	I	1	1
>	B	>	B	I	1	1
en	P	en	P	I	1	2
un	T	un	T	I	1	2

Cuadro 1.1. b. Salida de etiquetación

1. 2. Diccionario

Verbo (V), Sustantivo (S), Adjetivo (A), Adverbio (D), Preposición (P), Conjunción (C), Relativo (R), Determinante (T), Pronombre (N), Numeral (Z), Interjección (I) y Signos de puntuación (B).

Indicativo (Vi), Subjuntivo (Vs), Conjetural (Vc), Imperativo en segunda persona (Vj), Infinitivo (Vn), Participio (Vp) y Gerundio (Vg).

Persona y Número: Primera persona singular (1), Segunda persona singular (2), Tercera persona singular (3), Primera persona plural (4), Segunda persona plural (5) y Tercera persona plural (6).

Vocablo	**Etiqueta**
¡	B
¿	B
a	P
abacá	Sms
abacería	Sfs
abacero	Sms
ábaco	Sms
abad	Sms

Cuadro 1.2. Diccionario

1. 3. Preedición

Input de preedición	**Output de preedición**	**F. R.**	**Ct.**	**Ejemplos**
al	a el	a	P	al
del	de el	de	P	del
conmigo	con mí	con	P	conmigo
{me|te|nos|os|le|lo|la|les|los|las|se}	{me|te|nos|os|le|lo|la|les|los|las|se}		Vn	amarte comerlo recibirlos
ár{me|te|nos|os|se}lo	ar {me|te|nos|os|se} lo	ar	Vn	dármelo dártelo

Cuadro 1. 3. Tabla de Preedición

1. 4. Edición

F. O.	E.F. O.	Input	Output	Ejemplo
{c|qu}es	{z|c}	([AS])(.)s	[.$1$2p.]	voces audaces fraques
{o|as|a|amos|áis|adeis|an}	ar	Vn	[.Vip{1|2|3j|4|5|5|6}.]	amo amas ama amamos amáis aman
{o|es|e|emos|éis|edes|en}	er	Vn	[.Vip{1|2|3j|4|5|5|6}.]	como comes come comemos coméis comen
{o|es|e|imos|ís|ides|en}	ir	Vn	[.Vip{1|2|3j|4|5|5|6}.]	vivo vives vive vivimos vivís viven vivides

Cuadro 1. 4. Tabla de Edición

1. 5. Posedición

Input	Output	Procesamiento	Ejemplo
(Np)(T)¥n(S)	$2¥n$3	(Np)(T)¥n(S)=>(T)¥n(S)	las heridas la mancha
(Np)(T)¥n(V)	$1¥n$3	(N)(T)¥n(V)=>(N)¥n(V)	lo estudia
(Np)¥n(Np)(T)	$1¥n$2	(Np)¥n(Np)	te lo
(Np)(T)	$2	(N)(T)=>(T)	los estudios lo importante
(Np)¥n(P)(V)	$1¥n$3	(Np)¥n(P)(V)=>(V)	lo como los como
(P)(V)	$1	(P)(V)=>(P)	como participante

Cuadro 1. 5. Tabla de Posedición

Dimensión aleatoria asociativa

Expresiones regulares

Menos de 114 segundos en analizar 442,610 palabras en 14,522 líneas (oraciones)

18,387 casos de no análisis y de análisis equivocados en la totalidad del léxico, 442,610 (4.15%)

2. Aplicación del Analizador

2.1. Categorías gramaticales y frecuencias

Patterson (1975: 73) hace la observación de las frecuencias de cada categoría utilizando los datos estadísticos de Juilland y Chang-Rodríguez (1964), y precisa que el Léxico Funcional representa tan solo un 1,8 % del vocabulario básico, pero contribuye al 52 % del uso total.

Por otra parte el Léxico de Contenido cubre el 97 % de la totalidad de miembros del léxico tratado y, sin embargo, ocupa solamente un 46 % de palabras ocurrentes.

Categoría.	Frecuencia	Miembros	F./M.
Conjunción	37.674	13	2898.0
Preposición	54.278	21	2584.7
Relativo	10.594	5	2118.8
Pronombre personal átono	19.171	13	1474.7
Determinante	45.846	35	1309.9
Pronombre demostrativo	2.363	3	787.7
Pronombre personal tónico	5.438	12	453.2
Pronombre indefinido	1.845	5	369.0
Interrogativo	1.549	8	193.6
Adverbio	15.415	169	91.2

Numeral	1.766	36	49.1
Verbo	70.633	1646	42.9
Adjetivo	24.994	1225	20.4
Sustantivo	65.978	3810	17.3
Interjección	311	22	14.1

Cuadro 2.1. Categoría, Frecuencia (F), Miembros (M), Frecuencia / Miembros (F./M.)

2.2. Rangos de frecuencia

Preposición	Frec.	Log. N.	Rango
de	20231	1.0000	10
a	11330	0.9415	10
en	8000	0.9064	10
con	4292	0.8436	9
por	3849	0.8326	9
como	2211	0.7767	8
para	1385	0.7295	8

sin	1130	0.7090	8
sobre	435	0.6127	7
hasta	378	0.5986	6
entre	361	0.5939	6
desde	164	0.5144	6
según	161	0.5125	6
contra	117	0.4803	5
tras	84	0.4469	5
ante	56	0.4060	5
hacia	56	0.4060	5
salvo	21	0.3071	4
excepto	6	0.1807	2

fasta	6	0. 1807	2
cabe	4	0. 1398	2

Cuadro 2.2. Preposición (Prep.), Frecuencia (Frec.), Logaritmo normalizado (Log.N.), Rango

2.3. Categorías gramaticales y rangos de frecuencia

Categoría / Rango	1	2	3	4	5	6	7	8	9	10	Total
Sustantivo	1.656	973	579	349	171	70	10	2			3.810
Verbo	631	399	271	183	93	41	16	9	2	1	1.646
Adjetivo	562	279	191	122	39	25	5	2			1.225
Adverbio	55	36	20	17	18	11	8	4			169
Interjección	10	7	3	1		1					22
Numeral	7	8	8	8	1	3	1				36
Pronombre demostrativo	1	2			1	1	1				6
Pronombre indefinido	2	2	1		8	3					16

Interrogativo			2	1	2	2	1				8
Pronombre personal tónico		1	1	1	3	2	2	2			12
Preposición		3		1	4	4	1	3	2	3	21
Determinante				4	11	10	5	4	3	2	39
Conjunción		1		1	1	1	4	3		2	13
Pronombre personal átono						3	7		3		13
Relativo						1	3			1	5

Cuadro 2.3. Categorías gramaticales (miembros) y Rangos (1 – 10)

2. 4. Nuestra propuesta

Tipo Léxico / Frecuencia	**Alta Frecuencia** ← → **Baja Frecuencia**	
Léxico Funcional	Vocablo Gramatical	Vocablo Instrumental
Léxico de Contenido	Vocablo Común	Vocablo Específico

Cuadro 2.4. Combinación de Tipos Léxicos y Frecuencias

«Vocablo Gramatical»se caracteriza por su condición fonológica: son vocablos átonos y de longitud reducida, de una o dos sílabas en su mayoría. Se utilizan constantemente en todos los textos en general por su carácter gramatical, lo mismo que las terminaciones verbales de modo, tiempo y persona, las declinaciones de caso latino o las posposiciones japonesas.

«Vocablo Instrumental», que es Léxico Funcional de Frecuencia relativamente Baja. Son pronombres personales tónicos (*yo, tú, él,* etc.), pronombres demostrativos (*este, ese, aquel*), pronombres indefinidos (*uno, alguno, algo, ninguno, nada, alguien, cualquiera*) e interrogativos (*qué, dónde, cuándo, cómo,* etc.). Son palabras acentuadas a diferencia de los Vocablos Gramaticales, e incluye varias palabras trisilábicas (*nosotros, vosotros, aquellos, cualquiera* etc.)

La categorización bipartita del Léxico Funcional debe ser detallada con ajustes de miembros excepcionales. Por ejemplo acabamos de ver en el Cuadro 2.2. que las preposiciones pertenecen al Léxico Funcional de Alta Frecuencia («Vocablo Gramatical») por excelencia.

Excepciones: *salvo*, *excepto*, *cabe*, «Vocablos Instrumentales»

«Vocablos Gramaticales» que son de uso general

«Vocablos Instrumentales» que se necesitan en ocasiones particulares

«Vocablos Gramaticales» (artículos determinados e indeterminados)

«Vocablos Instrumentales» (posesivos, demostrativos, indefinidos)

Pronombres personales átonos («Vocablos Gramaticales»)

Pronombres personales tónicos («Vocablos Instrumentales»).

«Vocablos Comunes»: (10) *ser;* (9) *haber, decir;* (8) *tener, dar, hacer, estar, ver, poder, responder, querer, saber*

«Vocablos Específicos»: *abollar, abonar, abstener, acaballar, acanalar, acariciar, acarrear, aclamar, acocear, acondicionar, acuciar, acurrucar, acusar, administrar, adular, afinar, aforrar, agujerar, ahuyentar, alancear, alejar, alimentar, amañar, amarrar, apacentar, aparecer, apelar, apellidar, apetecer, aposentar,* etc.

2.5. Gramaticalización y frecuencia

Hopper y Traugott (2003: 127)

«Efecto de reducción» *I'll* (< I will) y *won't* (<will not) en inglés.

«Efecto de conservación» explica las formas irregulares invariables frente a las formas regularizadas analógicas de vocablos con poca frecuencia.

De esta manera la misma alta frecuencia causaría tanto la reducción como la conservación, lo cual puede parecer contradictorio.

Bybee (2003: 621) argumenta que el cambio fonético afecta primero a los ítems de alta frecuencia, mientras que la nivelación analógica afecta primero a los ítems de baja frecuencia, es decir, la conservación se presenta en los ítems de alta frecuencia.

La reducción se produce en los «Vocablos Gramaticales» (Léxico Funcional de Alta Frecuencia), mientras que la conservación es de los «Vocablos Comunes» (Léxico de Contenido de Alta Frecuencia). La reducción de forma es cuestión del Léxico Funcional; y la conservación, del Léxico de Contenido. Suponemos que la forma se reduce por su carácter gramatical y por su consiguiente carácter fonológico (átono) de la palabra en cuestión.

El verbo *haber*, Menéndez Pidal (1968: 303): «[su] frecuente uso como auxiliar le daba carácter de átono».

El verbo *haber* ha reducido su forma por ser un «Vocablo Gramatical», y no precisamente por su uso de alta frecuencia.

La frecuencia no sería la causa del cambio formal, sino más bien el resultado natural de su pertenencia al «Vocablo Gramatical».

Las lenguas dotadas de acento de intensidad suelen distinguir la función gramatical por el rasgo átono

Las lenguas de acentos tonales, para la distinción léxica entre Funcional y de Contenido, no recurren a la presencia y ausencia del acento y, por esta razón, no se presenta la reducción incluso en los vocablos gramaticales.

ika - *nakerebanaranai* (‘hay que ir’; que proviene literalmente de 'no puede ser que no vayamos')
furu-kamoshirenai (‘puede llover’; literalmente, 'no se sabe si llueve').

Las formas irregulares de un «Vocablo Común» se conserva, por ejemplo: *ser, decir, tener, dar, hacer, estar,* etc.
La alta frecuencia misma tampoco causaría la conservación.

«Vocablos Comunes», vocablos necesarios en general, independientemente del tipo de textos, cuyas formas irregulares suelen conservarse por ser formas establecidas y no derivadas por las reglas morfológicas propias de los verbos regulares.

En cambio, los «Vocablos Específicos» pueden dejar su forma irregular para confluir con las formas regulares. *crov-* (*creer*), *visc-* (*vivir*), *cox-* (*cocer*), etc. han dejado de ser irregulares en el español moderno.
conducir (*conducí*, *conduciste*...) en el habla popular.

Tanto la conservación de los «Vocablos Comunes» como la nivelación de los «Vocablos Específicos» no se deberían al uso frecuente o poco frecuente de cada categoría, sino la causa de la diferencia la encontramos en el rasgo tipológico de vocablos.

Bybee (2003: 604), para explicar el efecto de conservación que supone que posee la frecuencia, compara las 66 ocurrencias de la forma irregular inglesa *broke* y 5 de la forma regular de *damaged* en un corpus de un millón.

Nuestra hipótesis es que el factor más importante es la diferencia de importancia semántica del Léxico. Los «Vocablos Comunes» son más importantes dentro de la cognición del hablante que los «Vocablos Específicos», que son secundarios. El resultado natural de la diferencia de estos dos tipos de léxico es su frecuencia de uso: cuánto más común e importante, tanto más se utiliza en la vida. Insistimos que la relación de causa – efecto no debe de ser inversa: no se puede pensar que por su uso frecuente un vocablo cobre importancia.

Se aprendan los «Vocablos Comunes» antes que «Vocablos Específicos», que los primeros sean relativamente más frecuentes en general y más ampliamente usados en distintos textos que los últimos, y que las formas de los primeros estén más establecidas e invariables que los otros.

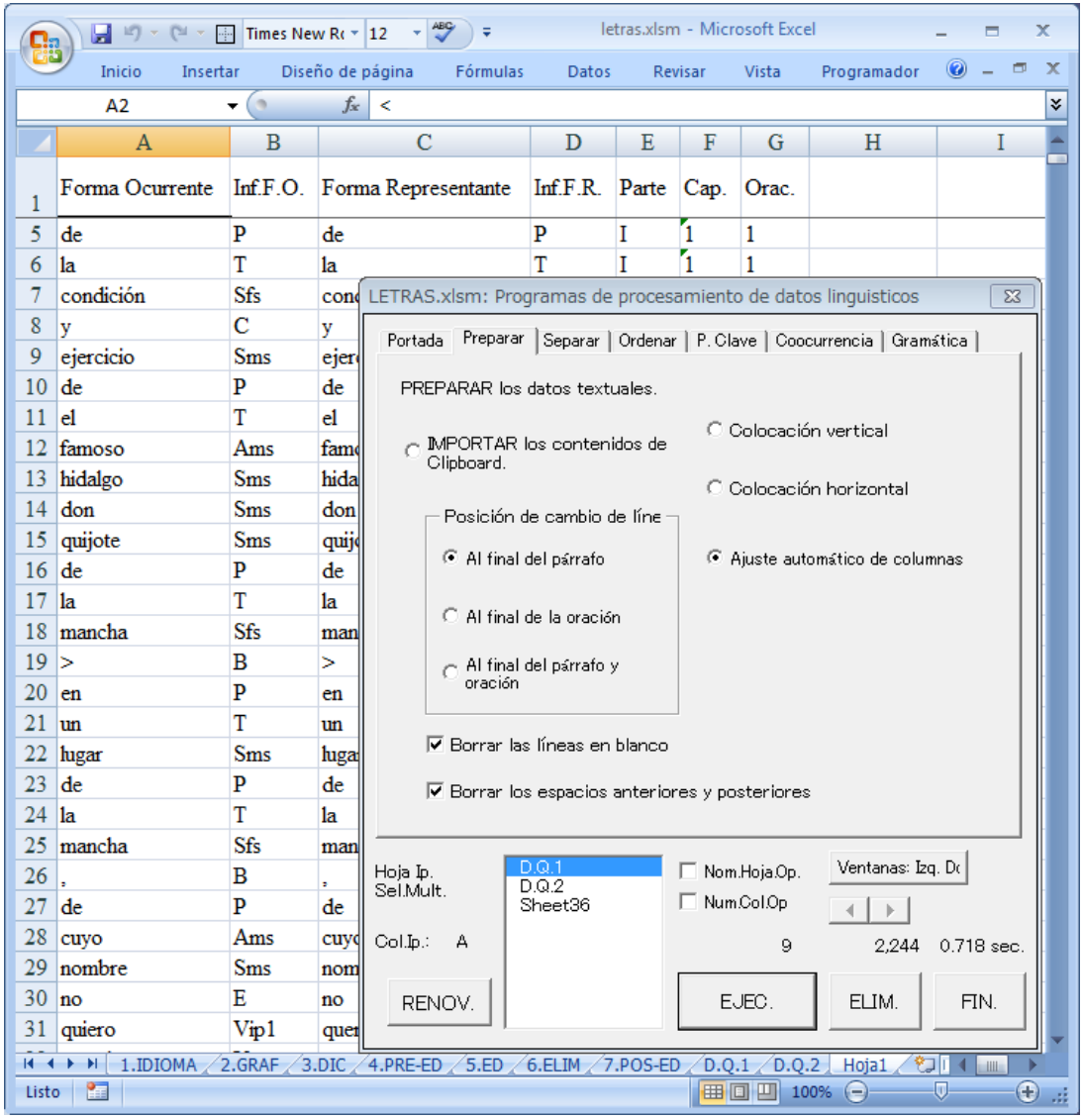
3. Final

Un grado actual de precisión (95.85%) se mejora inmediatamente por la introducción de vocablos y reglas gramaticales como parámetros libres.

Entorno complementario, fácil y libre con un criterio práctico de la velocidad más alta posible.

«Vocablos (V.) Gramaticales», «V. Instrumentales», «V. Usuales» y «V. Informativos»

Suponemos que un vocablo no se reduce en forma contraída ni se conserva en su forma irregular por ser utilizado con frecuencia, sino que más bien todo depende de su carácter léxico-gramatical.


letras.xlsm - Microsoft Excel
Inicio
Insertar
Diseño de página
Fórmulas
Datos
Revisar
Vista
Programador
A2
Forma Ocurrente
Inf.F.O.
Forma Representante
Inf.F.R.
Parte
Cap.
Orac.
LETRAS.xlsm: Programas de procesamiento de datos linguisticos
Portada
Preparar
Separar
Ordenar
P. Clave
Coocurrencia
Gramática
PREPARAR los datos textuales.
IMPORTAR los contenidos de Clipboard.
Colocación vertical
Colocación horizontal
Posición de cambio de líne
Al final del párrafo
Al final de la oración
Al final del párrafo y oración
Ajuste automático de columnas
Borrar las líneas en blanco
Borrar los espacios anteriores y posteriores
Hoja Ip. Sel.Mult.
D.Q.1
D.Q.2
Sheet36
Nom.Hoja.Op.
Num.Col.Op
Ventanas: Izq. D
Col.Ip.: A
9
2,244
0.718 sec.
RENOV.
EJEC.
ELIM.
FIN.
1.IDIOMA
2.GRAF
3.DIC
4.PRE-ED
5.ED
6.ELIM
7.POS-ED
D.Q.1
D.Q.2
Hoja1
Listo
100%

A.L.C.

Referencias Citadas

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

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

Baayen, R. H. (2001) *Word frequency distributions.* Dordrecht: Kluwer Academic Publishers.

Bybee, J. (2003) "Mechanisms of change in grammaticalization: the role of frequency", in Joseph and Janda (eds.), pp. 602-623.

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

Dubois J.; Giacomo, M.; Guespin, L.; Marcellesi, Ch.; Marcellesi J.-B.; Mével, J.-P.

(1979) *Diccionario de Lingüística.* Madrid: Alianza.

Fortson, B. W. (2003) "An approach to semantic change", in Joseph and Janda (eds.), pp. 648-666.

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.

Herdan, G. (1956) *Language as choice and change*. Groningen: Noorhoff.

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

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

Compostela.

Joseph, B. D. and Janda, R. D. (2003) *The Handbook of historical linguistics.* Oxford: Blackwell.

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

Menéndez Pidal, R. (1968) *Manual de gramática histórica española*, 13a ed. Madrid, Espasa-Calpe.

Meyer, Ch. F. (2002) *English corpus linguistics. An introduction*. Cambridge University Press.

Laca, B. (1999) "Presencia y ausencia de determinante", en Bosque, E. y Demonte V. (eds) *Gramática descriptiva de la lengua española*, vol. 1, p. 891-928.

Lieberman, E.; Michel, J-B.; Jackson, J.; Tang, T. and Nowak M. A. (2007)

"Quantifying the evolutionary dynamics of language", *Nature*, vol. 449, p.713-716.

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

Patterson, W. (1975) *The lexical structure of Spanish*. The Hague: Mouton.

Ueda, H. y 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*. p. 919-932, Universidade da Coruña.