IX Congreso Internacional de la Asociación Asiática de Hispanistas

Universidad de Chulalongkorn Bangkok, Tailandia, del 22 al 24 de enero de 2016

Tratamiento lingüístico y matemático de textos digitales españoles

Presentación del Programa LEXIS-web

Hiroto Ueda

Universidad de Tokio

1. Introducción

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/

Google: Hiroto Ueda

Google: Hiroto Ueda Letras

2. Interfaces de input y output

Fig. 1. Input de LEXIS-web

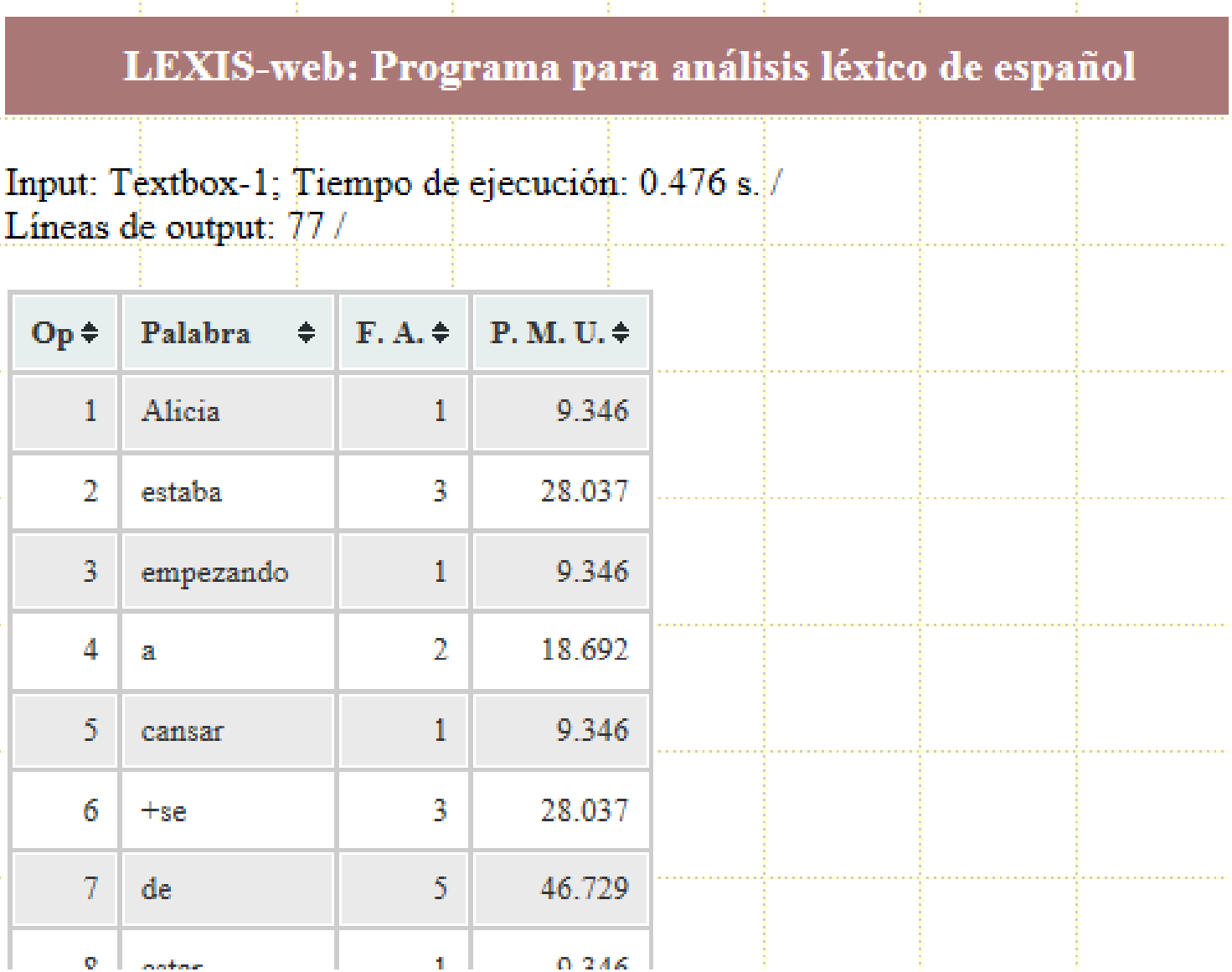
LEXIS-web: Program for lexical analysis of Spanish

Input: Textbox-1; Execution time: 0.407 s. /
Output lines: 107 /

Op	Palabra	C.S.(palabra)	Lema	C.S. (lema)	Homógrafos	Total	Prob.	Ip
1	Alicia	Xant	Alicia	Xant	1:Xant	-	-	1
2	estaba	Estar:IndImp13	estar	Estar	1:Estar	-	-	1
3	empezando	Ger	empezar	Inf	1:Ger	-	-	1
4	a	Prep	a	Prep	1:Prep	-	-	1
5	cansar	Inf	cansar	Inf	1:Inf	-	-	1
6	+se	Clit:3	se	Clit	1:Clit	-	-	1
7	de	Prep	de	Prep	1:Prep	-	-	1
8	estar	Estar	estar	Estar	1:Estar	-	-	1
9	*sentada	PP:fs	sentar	Inf	2:Sus/PP	100	0.556	1
10	en	Prep	en	Prep	1:Prep	-	-	1
11	*la	L:fs	el	L	2:L/Clit	180	0.643	1
12	*orilla	Sus:fs	orilla	Sus	2:Sus/V	140	0.933	1

Fig. 2. Output de LEXIS-web

Op	Número secuencial de output
Palabra	Palabra separada del texto, por ejemplo *del* se separa en *de +el*.
C.S.(palabra)	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.
C.S.(lema)	Categoría sintáctica del lema *empezar* «Inf», que es infinitivo
Homógrafos	Número y categorías posibles, por ejemplo *la*: «L» (forma femenina singular del artículo *el*) «Clit» (forma femenina singular del clítico *lo*).
Total	Frecuencia correspondiente.
Prob(abilidad)	Probabilidad dentro de las combinaciones calculadas en el mismo contexto.
Ip.	Número secuencial de línea de input.



LEXIS-web: Programa para análisis léxico de español

Input: Textbox-1; Tiempo de ejecución: 0.476 s. /
Líneas de output: 77 /

Op	Palabra	F. A.	P. M. U.
1	Alicia	1	9.346
2	estaba	3	28.037
3	empezando	1	9.346
4	a	2	18.692
5	cansar	1	9.346
6	+se	3	28.037
7	de	5	46.729
8	estar	1	9.346

Fig. 3. Output de LEXIS-web (2)

3. Identificación léxica

Diccionario:

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

Abreviación (Abr.):

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)*

4. Separación:

KEY	ITM
al	a +el/a/Prep
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

.

5. Lematización y asignación gramatical

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*

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
(...)	(...)

"-a" final de palabra:

(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

6. Desambiguación

sentada, *la*, *orilla*, etc:

Secuencia	Frec.
Adj-Adj	60
Adj-Adv	30
Adj-Clit	60
Adj-Conj	60
(…)	(…)
Y|o-Sus	80
Y|o-U	50
Y|o-V	60
Y|o-Xant	70
Y|o-Xtop	70

... *en la orilla*... Prep - {L/Clit} - Sus/V

(1) Secuencia anterior: *en {la}*

«Prep - {L}» (90)

«Prep - {Clit}» (0)

(2) Secuencia posterior: *{la} orilla*

«{L} - Sus» (90)

«{L} - V» (0)

«{Clit} - Sus» (10)

«{Clit} - V» (90)

=> {L}: 90 + 90 + 0 = 180 / {Clit}: 0 + 10 + 90 = 100

{L} Total: 180; Probabilidad 180 / (180 + 100) = .643

7. Final

Precisión, cantidad, rapidez, comodidad (sencillez)

98% de precisión, 10.000 palabras, menos de 5 segundos, ?

Referencias:

Almela, Ramón / Cantos, Pascual / Sánchez, Aquilino / Sarmiento, Ramón / Almela, Moisés. (2005). *Frecuencias del español. Diccionario y estudios léxicos y morfológicos*. Madrid : Universitas.

Ávila Muñoz, Antonio Manuel. (199). *Léxico de frecuencia del español hablado en la ciudad de Málaga*. Málaga: Universidad de Málaga.

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

Bybee, Joan. (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, Concepción. (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, Mark. (2006). *A frequency dictionary of Spanish. Core vocabulary for learners.* New York: Routledge.

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

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

Hopper, Paul J. / Traugott, Elizabeth Closs. (2003). *Grammaticalization,* 2nd ed. Cambridge: Cambridge University Press.

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

Juilland, Alphonse / Chang-Rodríguez, Eugenio. (1964). *Frequency dictionary of Spanish words.* The Hague: Mouton .

Lieberman, Erez / Michel, Jean-Baptiste / Jackson, Joe; Tang, T. / Nowak Martin A. (2007). "Quantifying the evolutionary dynamics of language", *Nature*, vol. 449, 713-716.

Moreno Sandoval, Antonio (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, Mark. / Atkinson, Quentin D. / Meade Andrew. (2007). "Frequency of word-use predicts rates of lexical evolution throughout Indo-European history", *Nature*, 449, 717-720.

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

Ueda, Hiroto / Perea Maria Pilar. (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, Hiroto / Rubio, Carlos. (2006). *Puerta al español. Nuevo diccionario español-japonés*. Tokio: Kenkyusha.

[Fin]