DH@Madrid Summer School 2016

Procesamiento del Lenguaje Natural y Poesía

Núria Bel (UPF)
nuria.bel@upf.edu

Madrid, 27 junio al 1 de julio de 2016

PLN y poesía

Núria Bel

upf. Universitat Pompeu Fabra Barcelona

Procesamiento del Lenguaje Natural

Objetivo

Realizar tareas que asociamos a leer, entender y extraer información

¿Qué tareas?

Todas las que podamos resolver como una secuencia de acciones planteadas en términos de "***si ... entonces ...***"

upf. Universitat Pompeu Fabra Barcelona

“*si ... entonces ...*”???

Los programas informáticos son colecciones de instrucciones diseñadas por humanos que se basan en la declaración de condiciones que se han de cumplir para tomar decisiones y manipular el input.

Por ejemplo:

- Corrector ortográfico: ***si*** una palabra no está en la lista de palabras de la lengua, ***entonces*** señala que es un error.

Ejercicio 1 (1/5)

- Mira este programa. https://www.meaningcloud.com/demo
- Busca “sentiment” en las opciones de análisis
- Nos dice si un texto puede ser interpretado como una opinión positiva o negativa.
- ¿cómo crees que puede decirlo?
  - Piensa en el corrector ortográfico y en plantearlo como “***si … entonces …***”
- Prueba estos dos textos. ¿Hace lo que esperas?
  - Este poema es el más bonito que he leído jamás
  - Este poema es el peor que he leído jamás

Ejercicio 1 (2/5)
Prueba los textos … (por separado)

Ejercicio 1 (3/5)

Respuesta: efectivamente, como en el caso del corrector ortográfico, tiene una lista de palabras que considera “positivas” y otras “negativas”.

De este extracto del poema de Ruben Darío, *Canto a la Argentina*, ¿qué palabras pondrías en la lista de palabras positivas y cuáles en la de negativas?

Argentina, el día en que te vistes de gala, en que brillan tus calles y no hay aspectos ni almas tristes en alturas, pampas y valles; el día en que desde tus fuertes, tus cruceros y tus cuarteles salvas lanzas, músicas viertes entre las palmas y laureles, visitada por los príncipes de reinos y tierras lejanas y mensajeros de repúblicas, son las pairias americanas las

Ruben Darío, Canto a la Argentina (1914), Extraído de Proyecto Gutemberg: https://www.gutenberg.org/files/51458/51458-0.txt

Ejercicio 1 (4/5)
Se calcula la ratio positivas/negativas

Argentina, el día en que te vistes de gala, en que brillan tus calles y no hay aspectos ni almas tristes en alturas, pampas y valles; el día en que desde tus fuertes, tus cruceros y tus cuarteles salvas lanzas, músicas viertes entre las palmas y laureles, visitada por los príncipes de reinos y tierras lejanas y mensajeros de

Palabras positivas	Palabras negativas
gala	tristes
brillan	
laureles	
júbilo	

Ruben Darío, Canto a la Argentina (1914), Extraído de Proyecto Gutemberg: https://www.gutenberg.org/files/51458/51458-0.txt

Ejercicio 1 (5/5)
¿Preguntas?

¿Y las otras palabras?

Si no están en las listas, asignadas a una información **explícita**, el programa no las puede reconocer ni tenerlas en cuenta.

¿Más preguntas?

upf. Universitat Pompeu Fabra Barcelona

Procesamiento del Lenguaje Natural

Procesar: **reconocer** palabras y asignarles una representación (información) que permita poder tomar decisiones.

Representación: **añadir** (substituir) a la cadena de caracteres que forma una palabra (o secuencia de palabras) información explícita de sus características para una tarea determinada

¿**Qué** información explícita? Depende de la tarea ..

PLN, tareas más conocidas

- Análisis de opinión
- Corrección gramatical
- Traducción automática
- Búsqueda y recuperación de información
- Extracción de información
- Resumen automático
- Respuesta a preguntas y asistentes virtuales
- Análisis lingüístico

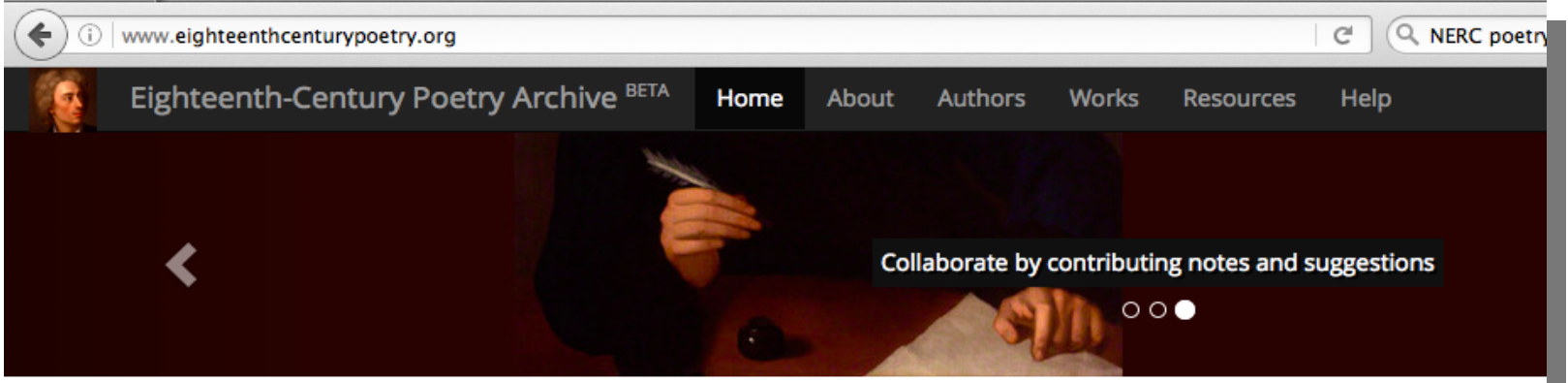
¿¿¿Y para poesía???

¿Extracción de información?

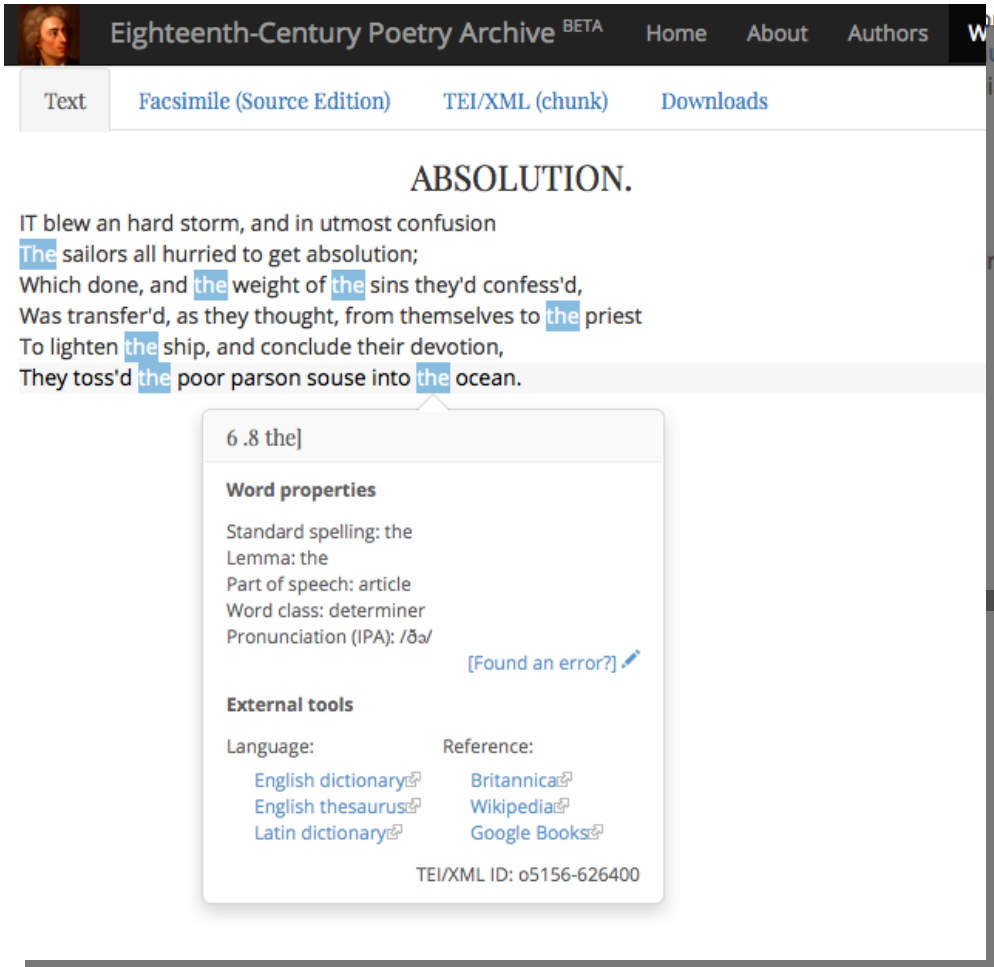
¿Análisis de opinión?

¿Análisis lingüístico?

pero hay que saber qué puede hacer y qué no ...



Eighteenth-Century Poetry Archive

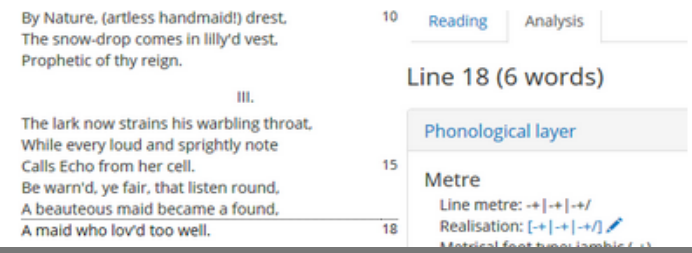


rative digital archive and research project devoted to the poetry of the long e
uthors. ECPA builds on the electronic texts created by the Text Creation Partn
ian Libraries, University of Oxford (editor of the Thomas Gray Archive). See

re.

Analyse

Enjoy close reading individual poems with the help of a range of tools.



upf. Universitat Pompeu Fabra Barcelona

Las palabras y la tecnología ...

- Empezamos por “palabras en una lista”
- Hay que hacer un programa para reconocer palabras en un texto
  - ***si*** QUÉ ***entonces*** “es una palabra”

- Hay que saber un poco más ...
- Haremos un pequeño paréntesis:

  Formato y codificación de textos

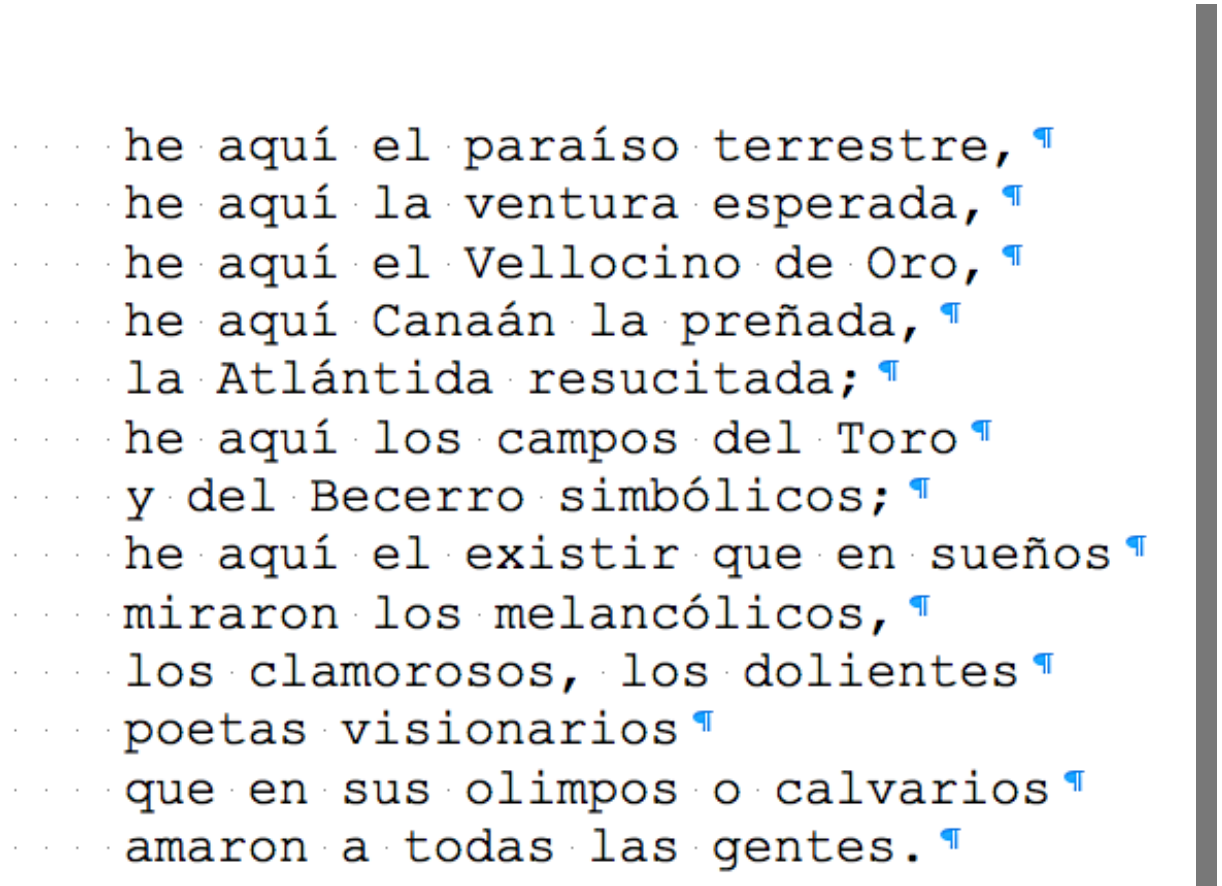
upf. Universitat Pompeu Fabra Barcelona

Formato y codificación de textos (1/11)
Palabras son...

- Una palabra es una secuencia de caracteres
- Un caracter es un código
- Los espacios en blanco también son código!
- Una palabra será una secuencia de códigos de caracteres entre dos códigos de espacio en blanco.

Formato y codificación de textos (2/11)

Código para todo ...
para espacio, tabulador, final de línea ...



Formato y codificación de textos (3/11)

Un caracter también es código: una secuencia de bi

0110 0001 a

0100 0001 A

- Para buscar “triste” en el texto:

  “La nena triste quiere un caramelo”

- Se comparará cada palabra del texto para decidir si es o no es lo que busca
- Se comparará cada caracter con el criterio de búsqueda
- Se comparará cada bit ...

  “ triste” - “La ”

  “c” – “L”

  “01100011” – “01001100”

  “0” AND “0” = 1

  “1” AND “1” = 1

  “1” AND “0” = 0

Formato y codificación de textos (4/11)
Codificación de caracteres

- ✓ "Codificación de caracteres" se refiere a las "correspondencias entre cada caracter ("a") y una secuencia única de código binario (*01100001*)"
- ✓ Hay diferentes listas de correspondencias, (todas estándares):
  - ASCII (para el alfabeto del inglés)
  - ISO-8859 (para el de las lenguas europeas)
  - UTF8 – UNICODE (para todas las del mundo)

http://www.unicode.org

Formato y codificación de textos (5/11)

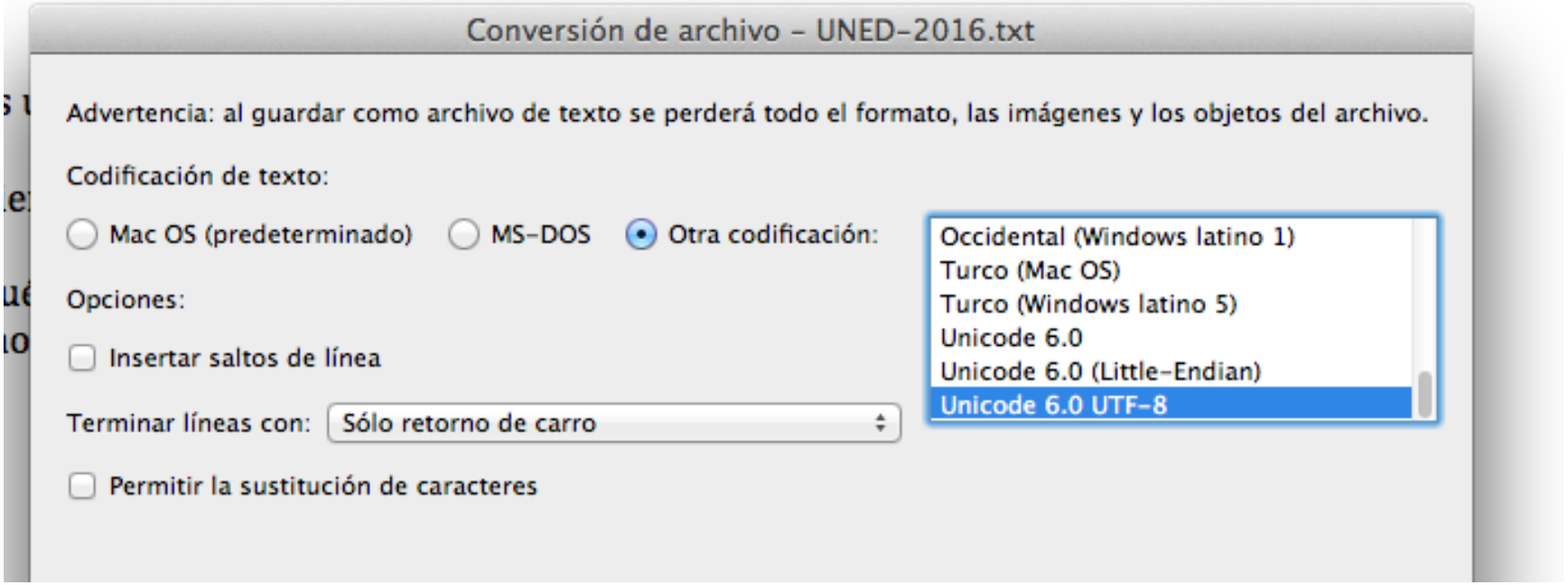
Una aplicación o programa que no tiene la misma correspondencia que la de la codificación del texto que queremos analizar ... no lo leerá bien!

he aquV≠ el paraV≠so terrestre,
he aquV≠ la ventura esperada,
he aquV≠ el Vellocino de Oro,
he aquV≠ CanaV°n la preV±ada,
la AtlV°ntida resucitada;
he aquV≠ los campos del Toro
y del Becerro simbV≥licos;
he aquV≠ el existir que en sueV±os
miraron los melancV≥licos,
los clamorosos, los dolientes
poetas visionarios
que en sus olimpos o calvarios
amaron a todas las gentes.

Muchas aplicaciones para el inglés no pueden ni leer los textos en otras lenguas que necesitan otros caracteres!

Formato y codificación de textos (6/11)

Lo mejor es guardar documento en “formato texto” “.txt” y “codificación de texto”: UNICODE UTF8 ….



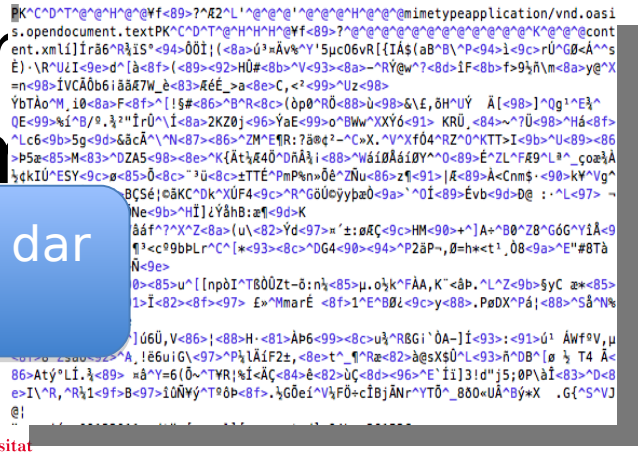
Esto es lo que sale en mi ordenador al guardar un documento en formato “.txt”. Puede salir de otra manera en un ordenador windows o linux.

Formato y codificación de textos (7/11)

¿Por qué “formato de texto”?

- El texto ha de ser **procesable** y no contener código (para la impresora o pantalla)
- “Texto sin formato” es sin ese código
- PDF no es un formato pr
  (está pensado para la in

Se pueden utilizar conversores, pero.... suelen dar problemas.



Formato y codificación de textos (8/11)

¿Dónde encontrar textos en formato texto? ¡un problema!

Aquí http://www.gutenberg.org/

Project Gutenberg

Online Book Catalog

Download | Bibrec

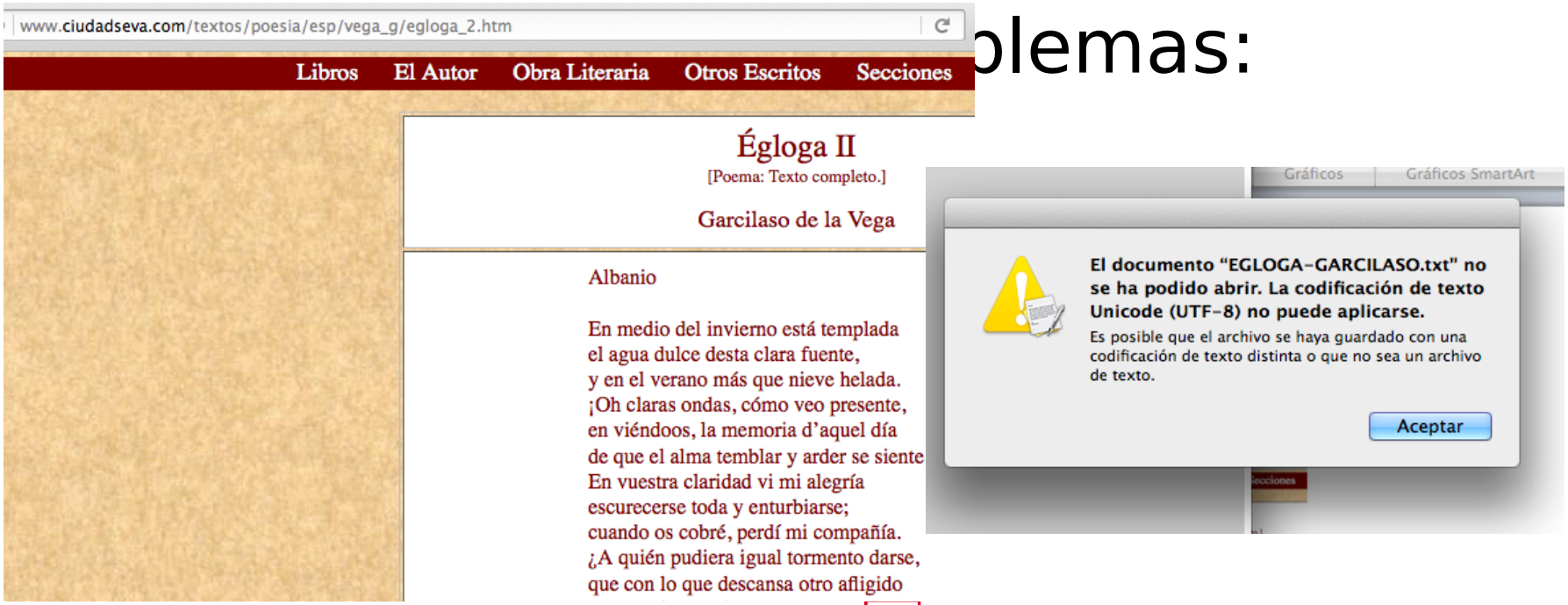
Download This eBook

	Format
	Read this book online: HTML
	EPUB (with images)
	EPUB (no images)
	Kindle (with images)
	Kindle (no images)
	Plain Text UTF-8
	More Files...

Formato y codificación de textos (9/11)

También el .html (página web) tiene código

Se puede intentar “Guardar página en formato texto” desde el navegador

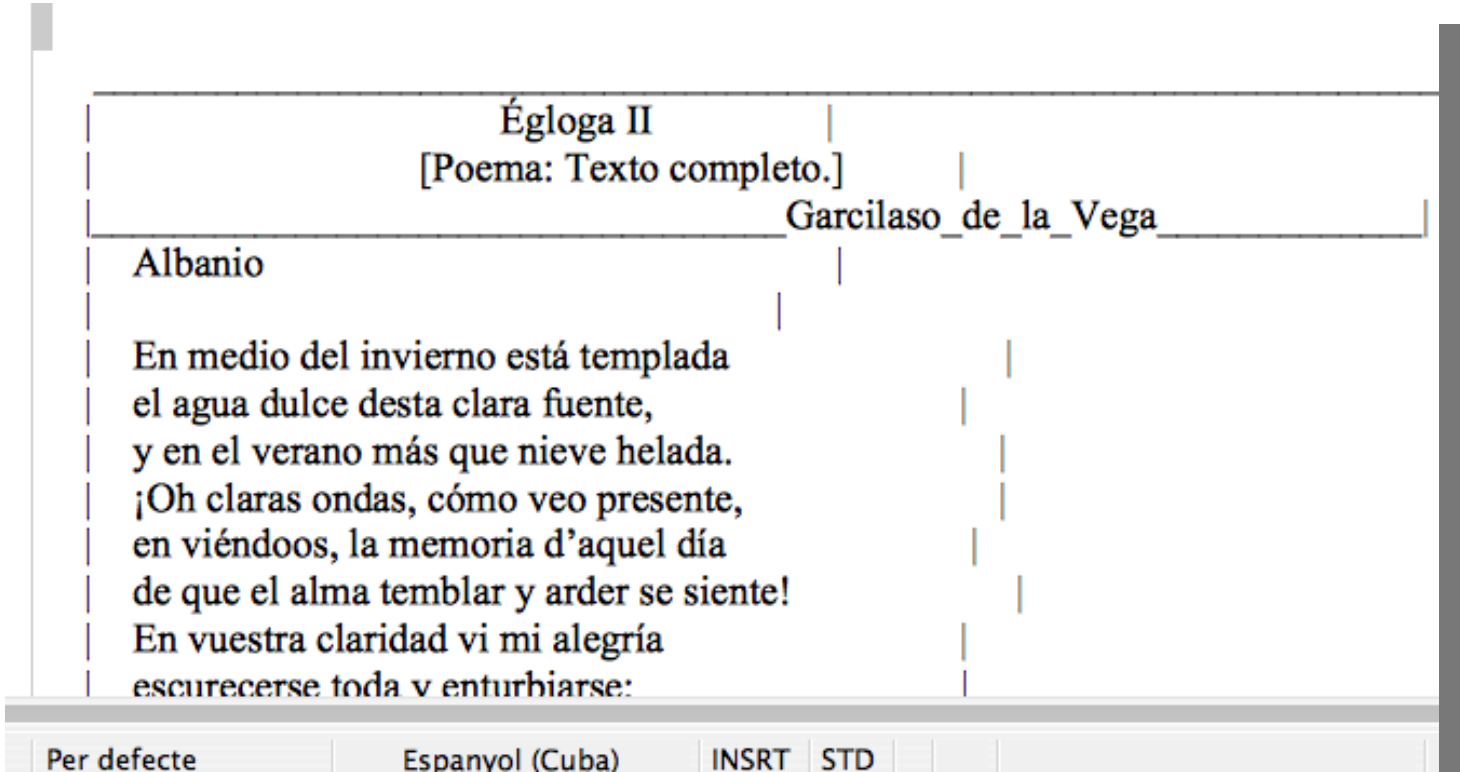


Formato y codificación de textos (10/11)

También .html (página web)

da más problemas:

caracteres que no son del texto, pero que se leen como si lo fueran



Formato y codificación de textos (11/11) Ejercicio 2

- Escojan su obra favorita en castellano y descárguenla
- Pueden buscar en https://www.gutenberg.org/browse/languages/es
- Por favor, ¡que sea larguito!
- Si no está en .txt, codificación utf8, abran el documento en un procesador (word, OpenOffice) y guarden una versión .txt, codificación utf8.

Las palabras y la tecnología

volvamos a lo nuestro ...

“***si ... entonces ...***”???

upf. Universitat Pompeu Fabra Barcelona

El programa añade información en forma de etiquetas

si .... *entonces* ETIQUETA

Si una secuencia de caracteres está entre caracteres de espacio en blanco, o de puntuación (o...), ***entonces*** añadir etiqueta: “es un TOKEN-palabra”.

Si una secuencia de TOKEN-palabra está entre TOKEN-puntuación-inicio-de-frase y TOKEN-puntuación-final-de-frase **entonces** añadir etiqueta: es un TOKEN-oración.

Etiquetas y Lenguajes de Marcas

XML, TEI, por ejemplo pero son etiquetas para PLN

▼ XML output

```
<document>
   <wordcount>70</wordcount>
   <cputime>0.015291</cputime>
   <paragraph>
      <sentence id="1">
         <token form="he" id="t1.1">
```

Si reconocemos palabras ... entonces...

Cuando reconocemos palabras, es posible:

- ✓contarlas
- ✓saber con qué otras palabras aparecen
- ✓identificar qué características lingüísticas tienen:
  - categoría gramatical: Nombre, Verbo, Adjetivo, ..
  - características morfosintácticas: Femenino, Pretérito, ..
  - otras características: nombre propio de persona,

Son programas de **análisis lingüístico**,
la base del Procesamiento del Lenguaje Natural

Programas de análisis lingüístico (1/5)

Análisis lingüístico para saber:

- palabras más frecuentes
- combinaciones (bigramas) más frecuentes de palabras
- combinaciones significativas de palabras: colocaciones

¿cómo sabe el programa qué palabra es un adjetivo?

Si .... ***entonces*** ADJ

Si está en una lista de palabras ADJ (diccionario)

Pero hay algunas complicaciones

Herramientas de análisis lingüístico (2/5)

http://nlp.lsi.upc.edu/freeling/node/1

FreeLing Home Page

Hooked on a FreeLing

Programa de código abierto, descargable, instalable en windows, mac y linux, gratis para investigación.

Main menu

- Home
- Features
- Linguistic Data
- Contributions
- License
- Installing
- Documentation
- Contributing
- Download
- Source code
- References
- Web Links
- Forum & FAQs

References

To cite FreeLing in your academic works, please reference the following papers:

About FreeLing as a whole:

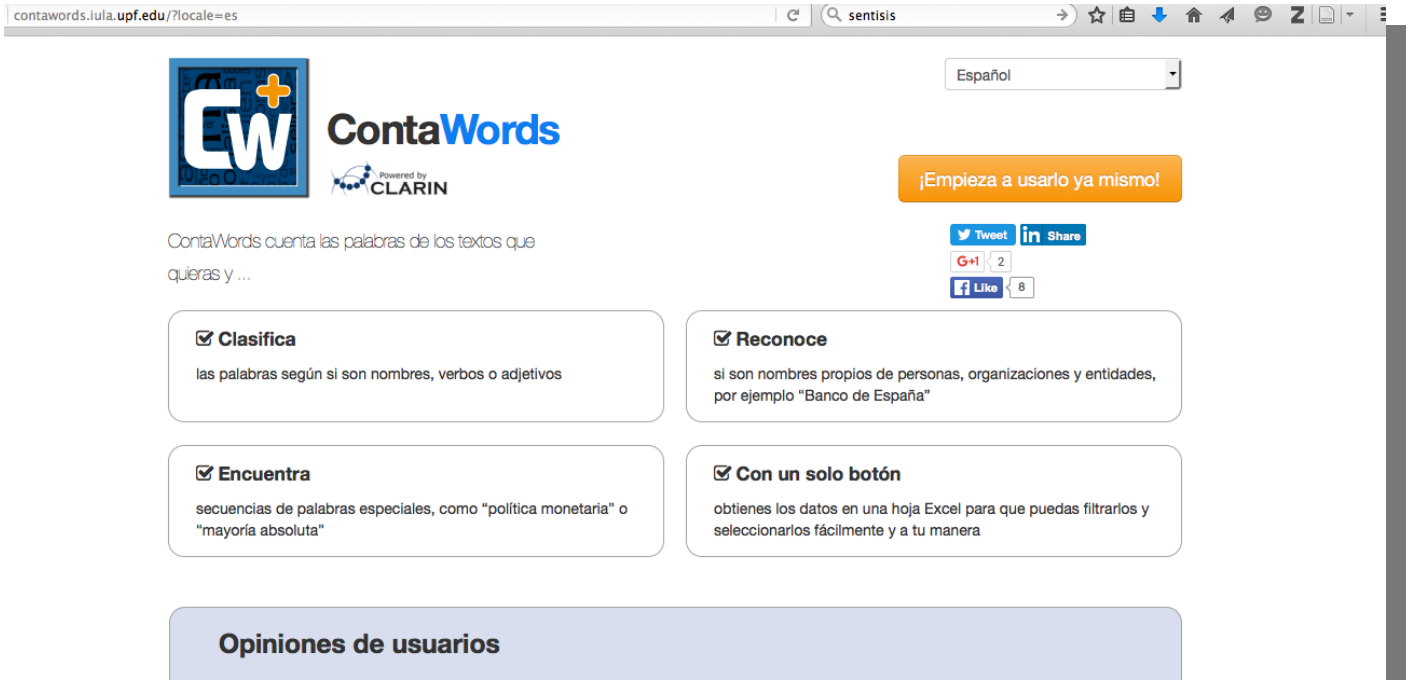
- Lluís Padró and Evgeny Stanilovsky.
  FreeLing 3.0: Towards Wider Multilinguality
  Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA.
  Istanbul, Turkey. May, 2012.
  [pdf] [bibtex]

Herramientas de análisis lingüístico (3/5)

http://contawords.iula.upf.edu

- utiliza Freeling
- es una aplicación wel

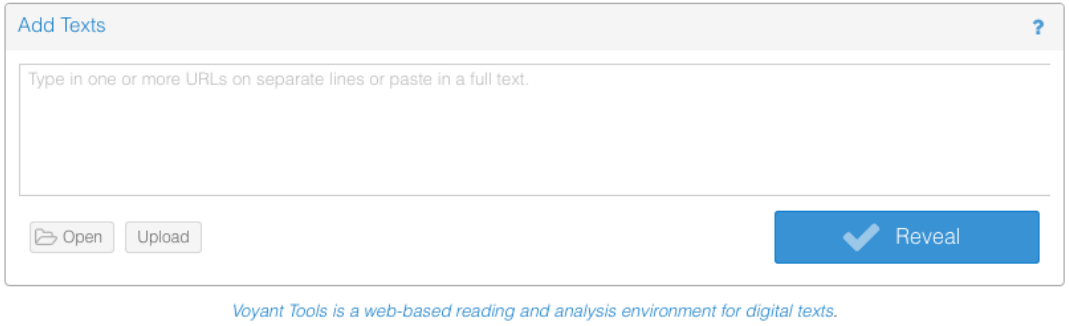
No necesita instalación, gratis para investigación



Herramientas de análisis lingüístico (4/5) http://voyant-tools.org/

- es una aplicación web
- tiene funciones interesantes

VOYANT
see through your text



Herramientas de análisis lingüístico (5/5)

Para obtener:

- ✓ Información cuantitativa
- ✓ Reconocer nombres propios y clasificarlos (lugar, persona, etc.)
- ✓ Anotación morfosintáctica para saber formas de pasado, en primera persona

Se puede hacer a mano, pero depende del texto se puede hacer muy pesado, ¿o no?

ContaWords

http://contawords.iula.upf.edu/more_

Powered by CLARIN

Hojas | Gráficos | Gráfico

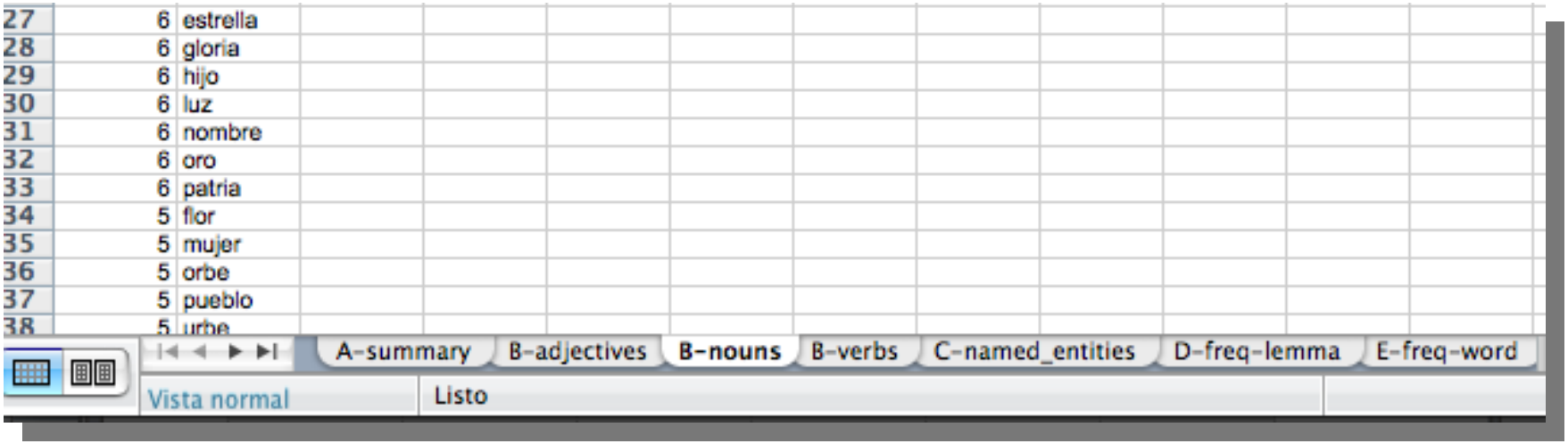
	A	B
9	TABS:	
10		
11	A	INDEX, statistics
12		
13	B	Nouns, adjectives and verbs frequencies
14		
15	C	Named Entities
16		
17	D	Absolute lemmas frequencies (all grammatical categories included)
18		
19	E	Absolute word frequencies by form (all grammatical categories included)
20		
21	F	Absolute frequency of two words sequences (bigrams)
22		
23	G	Absolute frequency of specially selected two word sequences
24		
25		
26	PART OF SPEECH TAGS: => Simple part of speech tags	
27		
28	A	Adjective
29	R	Adverb
30	D	Determiner
31	N	Noun
32	V	Verb
33	P	Pronoun
34	C	Conjunction
35	I	Interjection
36	S	Preposition
37	F	Punctuation mark
38	Z	Numbers
39	W	Dates and Time
40		
41		
42	Complex tags (with information about verbal tense, grammatical person, number, gender, etc.) are ex	
43		
44	English:	https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
45		
46	català:	http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-ca.html

Dado uno o más documentos:

- extrae listas de palabras según su categoría,
- lista las palabras por frecuencia,
- identifica y clasifica nombres propios y
- cuenta las combinaciones de dos palabras

y lo pone todo en una hoja excel para que sea fácil de inspeccionar y manipular

Distingue entre Nombres, Verbos, Adjetivos, y reconoce “Entidades con Nombre”



27	6	estrella
28	6	gloria
29	6	hijo
30	6	luz
31	6	nombre
32	6	oro
33	6	patria
34	5	flor
35	5	mujer
36	5	orbe
37	5	pueblo
38	5	urbe

Entidades con Nombre

"Extracción de
ción

¡Argentina, región de la aurora!

En inglés, Named Entity Recognition (NERC)
Usa información del contexto donde aparece
Comete errores o ¿es el contexto?

¡Argentina! tu ser no abriga

El Plata, padre extraordinario,

A	B	C	D
10	Argentina	Geographic	
6	América	Geographic	
6	Argentina	Organization	
6	Centenario	Others	
4	España	Geographic	
4	Italia	Geographic	
3	Buenos Aires	Geographic	
3	Dios	Geographic	
3	Sol	Person	
2	Dios	Others	
2	Exodos	Person	
2	Libertad	Others	
2	Oid	Person	
2	Plata	Person	
2	Roma	Geographic	
2	Roma	Person	

Entidades con Nombre “Extracción de información”

Comete errores, sí, pero también el error está en la palabra

¡Exodos!
¡Exodos!

Oid

A	B	C	D
10	Argentina	Geographic	
6	América	Geographic	
6	Argentina	Organization	
6	Centenario	Others	
4	España	Geographic	
4	Italia	Geographic	
3	Buenos Aires	Geographic	
3	Dios	Geographic	
3	Sol	Person	
2	Dios	Others	
2	Exodos	Person	
2	Libertad	Others	
2	Oid	Person	
2	Plata	Person	
2	Roma	Geographic	
2	Roma	Person	

Análisis Lingüístic
Nombres (comunes
frecuencia de lema

A	B	C	D
16	argentina		
16	tierra		
14	imagen		
13	fiesta		
13	paz		
11	dios		
11	hombre		
11	sol		
10	himno		
9	libertad		
8	hermano		
8	sangre		
7	centenario		
7	día		
7	grito		
7	mar		
7	nación		
7	pampa		
7	sombra		
7	vida		
7	voz		

Atención!!
Son *lemas*, si queremos *formas* ir a la pestaña **E-freq-word**

Lema: representante del paradigma flexivo

upf. Universitat Pompeu Fabra Barcelona

Análisis Lingüístico
Verbos: frecuencia de lem

A	B	C
40	ser	
24	haber	
16	ir	
12	cantar	
11	dar	
8	estar	
8	pasar	
7	libertar	
7	oír	
7	tener	
7	ver	
6	llegar	
6	mirar	
5	abrir	
5	amar	
5	crear	
5	decir	

Hay listas de palabras con información asignada: lema e información morfosintáctica:

arrebata arrebatar V

arrebatas arrebatar V

Pero hay palabras que tienen más de una etiqueta en la lista:

paso pasar V

paso paso N

El programa elige la que es en contexto:

"El paso de" vs. "Yo paso"

Análisis lingüístico
Etiquetado morfosintáctico
o “PoS tagging”

- Para cada palabra, el programa accede al diccionario y recupera etiquetas asociadas.
- Elige la correcta en contexto (y es el

Ejemplo de etiquetas asociadas a palabras. Ambigüedad categorial

▼ Sentences

Sentence 1

he	aquí	el	paraíso	terrestre	,	he	aquí	la	ventura	esperada	,	he	aquí	el	Vellocino_de_Oro
haber *VAIP1S0* 0.962485	aquí *RG* 1	el *DA0MS0* 1	paraíso *NCMS000* 1	terrestre *AQ0CS00* 1	, *Fc* 1	haber *VAIP1S0* 0.962485	aquí *RG* 1	el *DA0FS0* 0.98926	ventura *NCFS000* 0.638706	esperar *VMP00SF* 1	, *Fc* 1	haber *VAIP1S0* 0.962485	aquí *RG* 1	el *DA0MS0* 1	vellocino_de_oro *NP00000* 1
haber *VMIP1S0* 0.0363423						haber *VMIP1S0* 0.0363423		lo *PP3FSA0* 0.010734	venturo *AQ0FS00* 0.361294			haber *VMIP1S0* 0.0363423			
he *I* 0.00117233						he *I* 0.00117233		la *NCMS000* 6.20105e-06				he *I* 0.00117233			

Análisis lingüístico
El diccionario

User login

Username *

El diccionario es una lista (finita) de palabras con información lingüística que se ha hecho manualmente

Si una palabra no está en la lista, puede suponerse la categoría por el sufijo

in the dictionary.

- The Asturian dictionary contains more than 140,000 forms corresponding to some 40,000 combinations lemma-PoS.
- The Catalan dictionary contains more than 520,000 forms corresponding to 71,000 combinations lemma-PoS.
- The Welsh dictionary contains some 345,000 forms, corresponding to over 8,500 lemma-PoS combinations.
- The German dictionary contains near 395,000 forms, corresponding to almost 130,000 lemma-PoS combinations.
- The English dictionary was automatically extracted from WSJ and other corpuses, with accurate manual post-edition and completion.
  It contains about 68,000 forms corresponding to some 37,000 different
  combinations lemma-PoS.
- The Spanish dictionary contains over 555,000 forms corresponding to more than 76.000 lemma-PoS combinations.
- The French dictionary contains over 350,000 forms corresponding to more than 54.000 lemma-PoS

Diccionarios disponibles en FreeLing
Si la aplicación no tiene el diccionario de una lengua, no puede analizar textos de esa lengua

Análisis lingüístico
Las etiquetas

La información lingüística, las características morfosintácticas de las palabras, es la etiqueta asociada

abalanzara abalanzar **VMIC1S0** | abalanzar **VMIC3S0**
bajo bajar **VMIP1S0** | bajo **AQ0MS0** | bajo **NCMS000** | bajo **SPS00**

La etiqueta intenta ser "comprensible"

Hay estándares de cómo han de ser las etiquetas.

Análisis Lingüístico
Las etiquetas (Tagset): para cada lengua, y en estilos diferentes

- Penn Treebank: 45 tags
- Lancaster BNC/Full (C5/C7): 61/146 tags
- German Stuttgart-Tübingen Tag Set: 50 tags
- Swedish SUC Tag Set: 25 tags
- EAGLES: 254 tags (las que utiliza FreeLing/ContaWords)

Análisis Lingüístico Las etiquetas: Penn Treebank

una convención .. pero cuestionable

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	*+,%, &*
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WPS	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	*#*
PDT	Predeterminer	*all, both*	"	Left quote	(' or ")
POS	Possessive ending	*'s*	"	Right quote	(' or ")
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PP$	Possessive pronoun	*your, one's*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

Análisis Lingüístico
Penn TreeBank Tag Set for Spanish

ACRNM	acronym (ISO, CEI)
ADJ	Adjectives (mayores, mayor)
ADV	Adverbs (muy, demasiado, cómo)
ALFP	Plural letter of the alphabet (As/Aes, bes)
ALFS	Singular letter of the alphabet (A, b)
ART	Articles (un, las, la, unas)
BACKSLASH	backslash (\)
CARD	Cardinals
CC	Coordinating conjunction (y, o)
CCAD	Adversative coordinating conjunction (pero)
CCNEG	Negative coordinating conjunction (ni)
...	
PAL	Portmanteau word formed by a and el
PDEL	Portmanteau word formed by de and el

Análisis lingüístico
Las etiquetas: EAGLES

- Las etiquetas EAGLES codifican la información en una secuencia donde la posición se relaciona con el atributo del que se codifican los valores.
- La primera posición siempre codifica la categoría gramatical.
- Si una forma no tiene un valor para un determinado atributo, se codifica '0'.
- Dependiendo de la categoría gramatical, habrá un número determinado de posiciones/valor
- Está preparado para cubrir las necesida lenguas europeas

paso Ncms

Position	Attribute	Values
0	category	**N**:*noun*
1	type	**C**:*common*; **P**:*proper*
2	case	**N**:*nominative*; **G**:*genitive*; **D**:*dative*; **F**:*accusative*;
3	gen	**F**:*f*; **M**:*m*; **C**:*c*
4	num	**S**:*s*; **P**:*p*; **N**:*n*

Ejercicio 2: EAGLES Verbos

Codifica las etiquetas para las siguientes formas verbales:

cantaría
abriríamos
maté
leyera
correrás
oídas

Antes hemos visto posición inicial = 0, y aquí decimos = 1. No importa el número, ha de ser la primera!

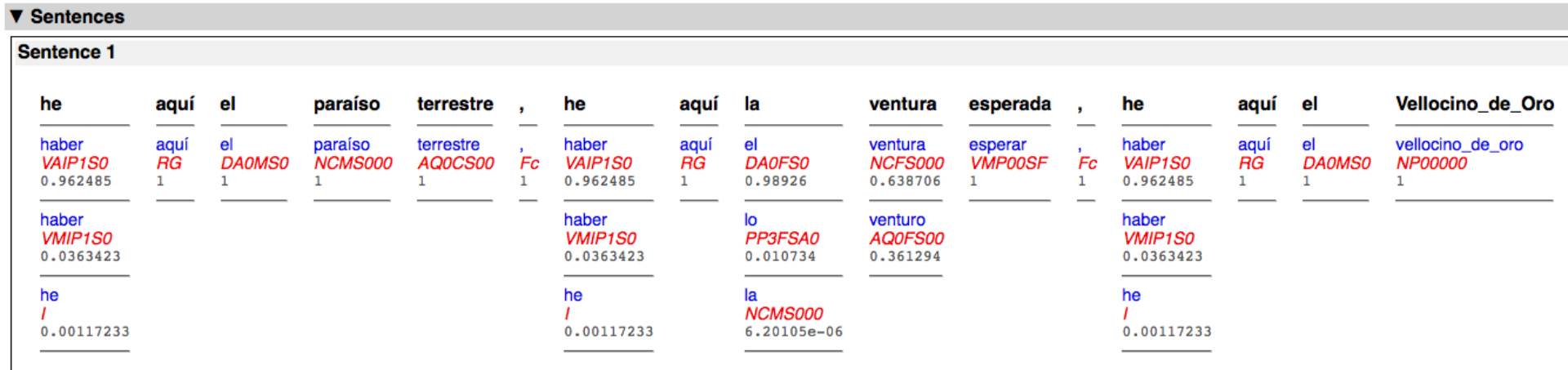
Pos.	Atributo	Valor	Código
1	Categoría	Verbo	V
2	Tipo	Principal	M
		Auxiliar	A
		Semiauxiliar	S
3	Modo	Indicativo	I
		Subjuntivo	S
		Imperativo	M
		Infinitivo	N
		Gerundio	G
		Participio	P
4	Tiempo	Presente	P
		Imperfecto	I
		Futuro	F
		Pasado	S
		Condicional	C
		-	0
5	Persona	Primera	1
		Segunda	2
		Tercera	3
6	Número	Singular	S
		Plural	P
7	Género	Masculino	M
		Femenino	F

Análisis lingüístico
Etiquetado morfosintáctico automático

FreeLing

Se basa en información de contexto ...

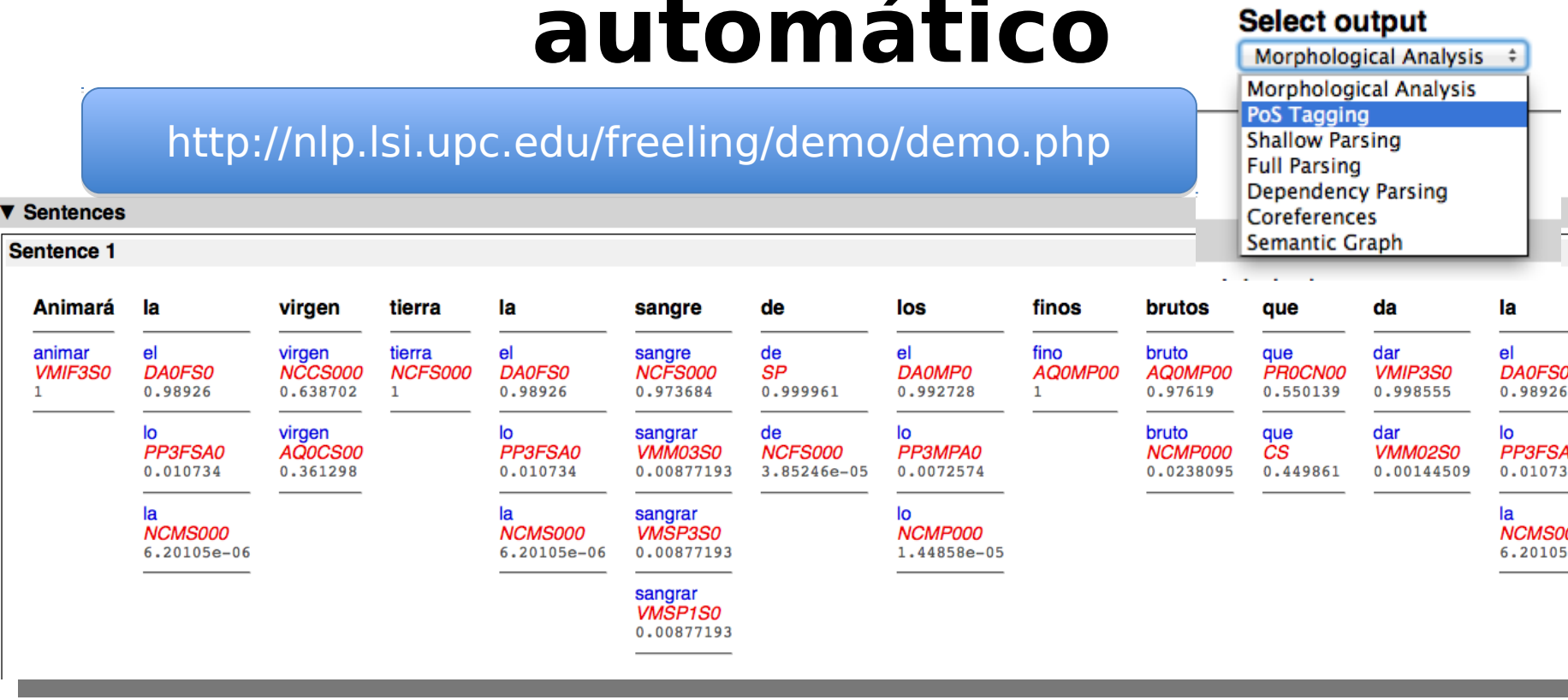
A partir de datos correctos (humanos) aprende la probabilidad de que, dada una etiqueta X, la anterior sea una etiqueta Y de entre las posibles para esa palabra.

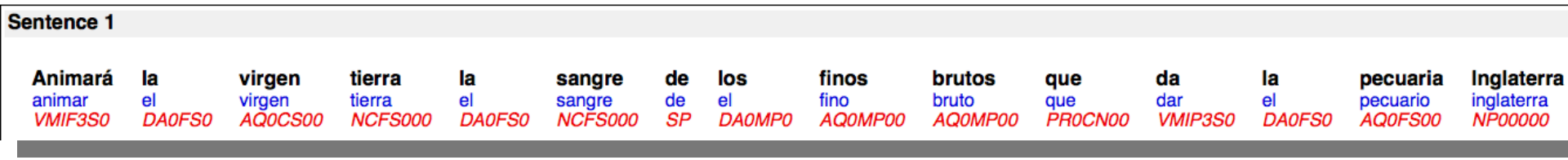


Análisis lingüístico
Etiquetado morfosintáctico automático

FreeLing

http://nlp.lsi.upc.edu/freeling/demo/demo.php

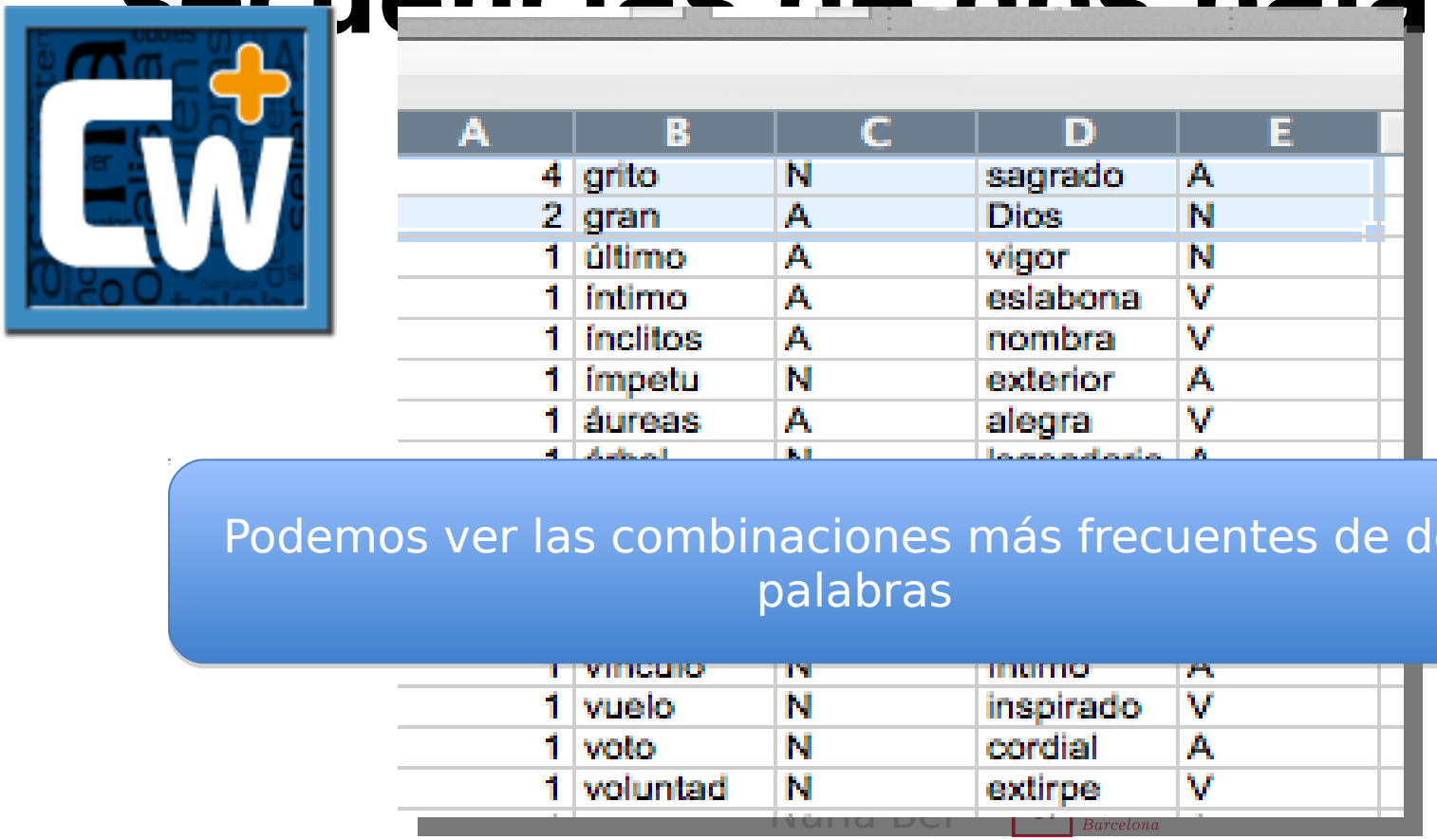




ContaWords

http://contawords.iula.upf.edu/more_info

Bigramas y su frecuencia: secuencias de dos palabras



A	B	C	D	E
4	grito	N	sagrado	A
2	gran	A	Dios	N
1	último	A	vigor	N
1	íntimo	A	eslabona	V
1	ínclitos	A	nombra	V
1	ímpetu	N	exterior	A
1	áureas	A	alegra	V
1	vuelo	N	inspirado	V
1	voto	N	cordial	A
1	voluntad	N	extirpe	V

Podemos ver las combinaciones más frecuentes de dos palabras

ContaWords

http://contawords.iula.upf.edu/more_info

Bigramas

Podemos ver la combinación para una determinada palabra

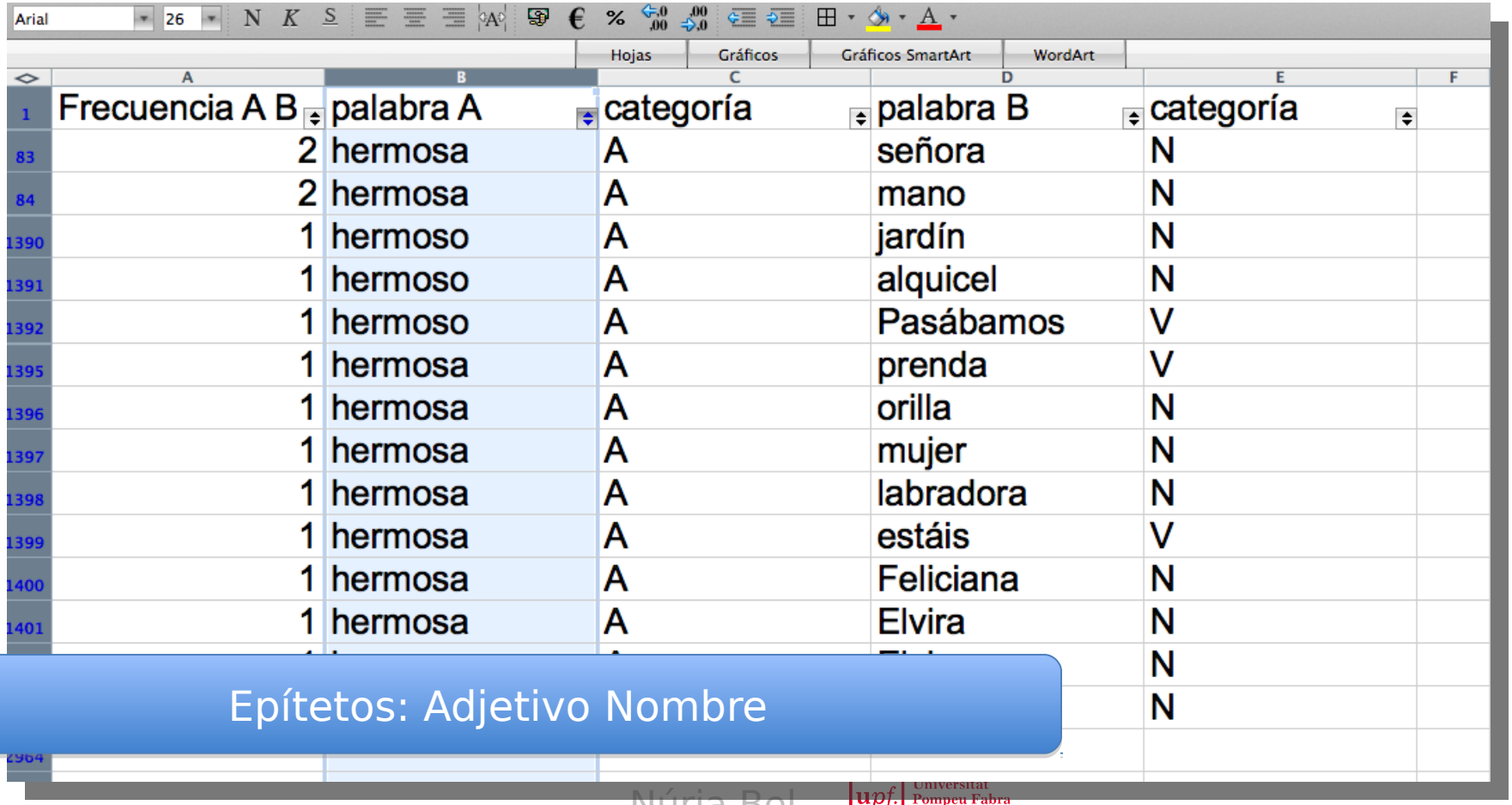
1	tonos	N	distintos	A
1	tigres	N	marciales	A
1	tierras	N	lejanas	A
1	tierra	N	sagrado	A
1	tierra	N	libre	A
1	tierra	N	labrada	V
1	tierra	N	feraz	A
1	tierra	N	austral	A
1	tierra	N	abierta	V
1	testa	N	imperiosa	A
1	tesoro	N	año	N
1	terruño	N	convierte	V

ContaWords

Con Excel (o similar) se pueden hacer “Filtros” para seleccionar vistas de los datos: Para ver las combinaciones de N-ADJ más frecuentes

- insertar fila (que sea la primera!)
- buscar la opción “Filtro” en el Menú Datos y marcar “Autofiltro”
- seleccionar: las combinaciones N-Adj o Adj-N o todas las palabras que combinan con … hermosa?

ContaWords
Filtros hoja Excel
para inspeccionar datos

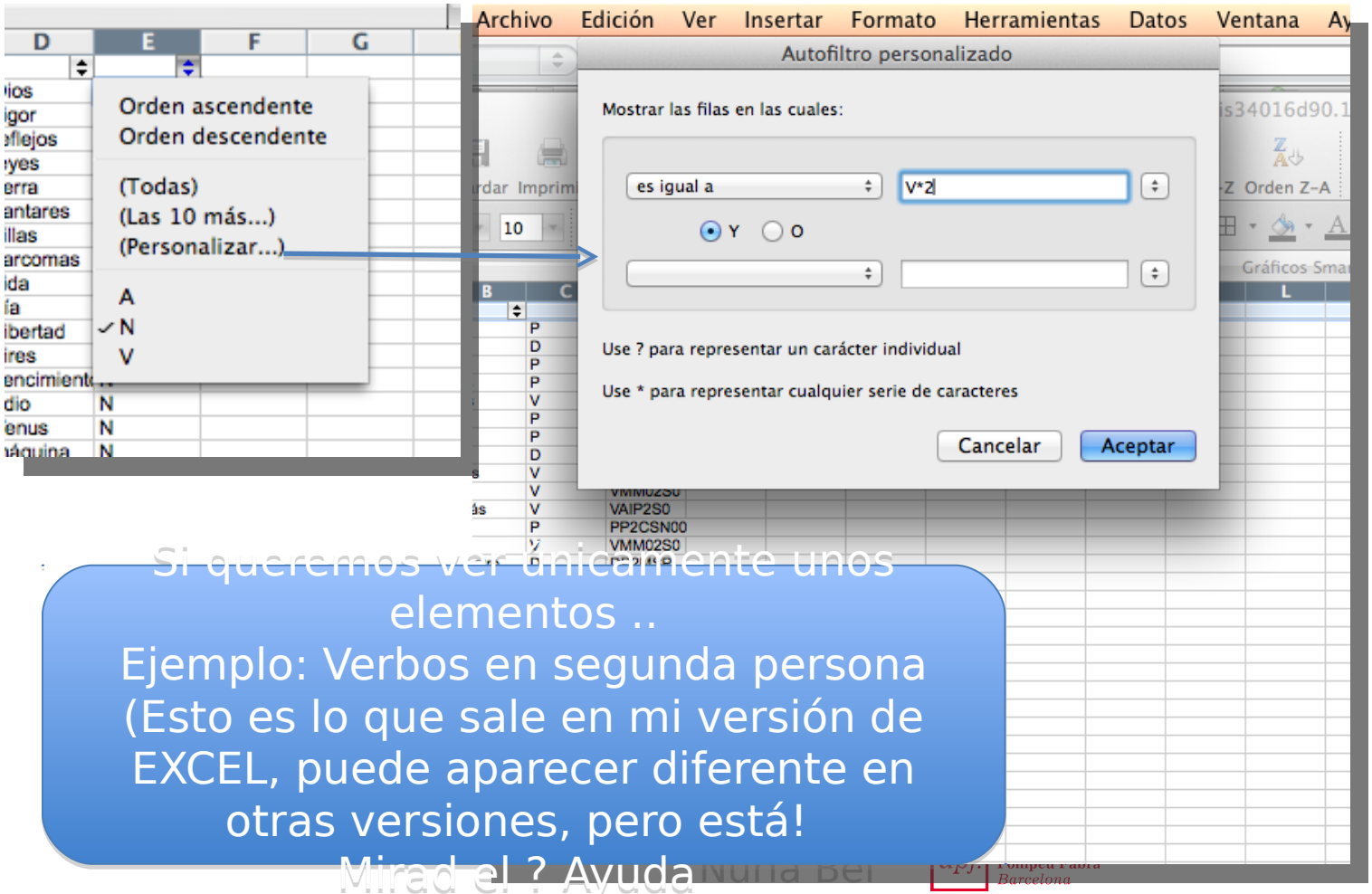


Hojas | Gráficos | Gráficos SmartArt | WordArt

	A	B	C	D	E	F
1	Frecuencia A B	palabra A	categoría	palabra B	categoría	
83	2	hermosa	A	señora	N	
84	2	hermosa	A	mano	N	
1390	1	hermoso	A	jardín	N	
1391	1	hermoso	A	alquicel	N	
1392	1	hermoso	A	Pasábamos	V	
1395	1	hermosa	A	prenda	V	
1396	1	hermosa	A	orilla	N	
1397	1	hermosa	A	mujer	N	
1398	1	hermosa	A	labradora	N	
1399	1	hermosa	A	estáis	V	
1400	1	hermosa	A	Feliciana	N	
1401	1	hermosa	A	Elvira	N	
					N	
					N	
2964						

Epítetos: Adjetivo Nombre

ContaWords Autofiltro para hacer búsquedas



Si queremos ver únicamente unos elementos ..
Ejemplo: Verbos en segunda persona
(Esto es lo que sale en mi versión de EXCEL, puede aparecer diferente en otras versiones, pero está!
Mirad el ? Ayuda

Ejercicio 4 (1/2)

- Utiliza ContaWords en: http://contawords.iula.upf.edu
- Sigue las instrucciones para subir el archivo que has guardado antes como .txt, utf8 y seleccionar la lengua.
- Encuentra y acciona el comando “Ejecutar”.
- Descarga el resultado cuando aparezca en pantalla. Puede tardar un poco.
- Cambia el nombre al documento Excel cuando lo guardes en tu ordenador.

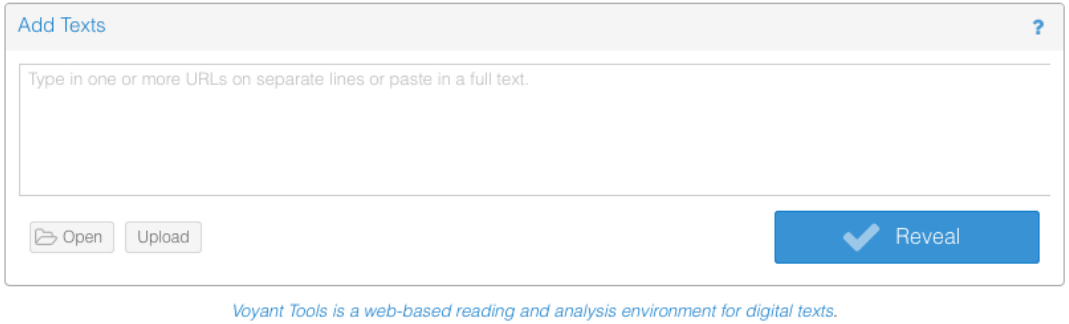
Ejercicio 4 (2/2)

- Inspecciona los resultados del análisis de ContaWords y responde a estas preguntas:
  - En la pestaña de nombres, ¿por qué están todos en singular?
  - Encuentra cuáles son las 10 palabras más frecuentes del texto ¿Tienen alguna característica en común?
- Crea un Filtro en la pestaña “G-freq-bigram” siguiendo las instrucciones de las transparencias del curso.
- Con el filtro creado, selecciona las opciones para que se muestren únicamente las combinaciones de Adjetivo-Nombre, y las de Nombre-Adjetivo.

Herramientas de análisis http://voyant-tools.org/

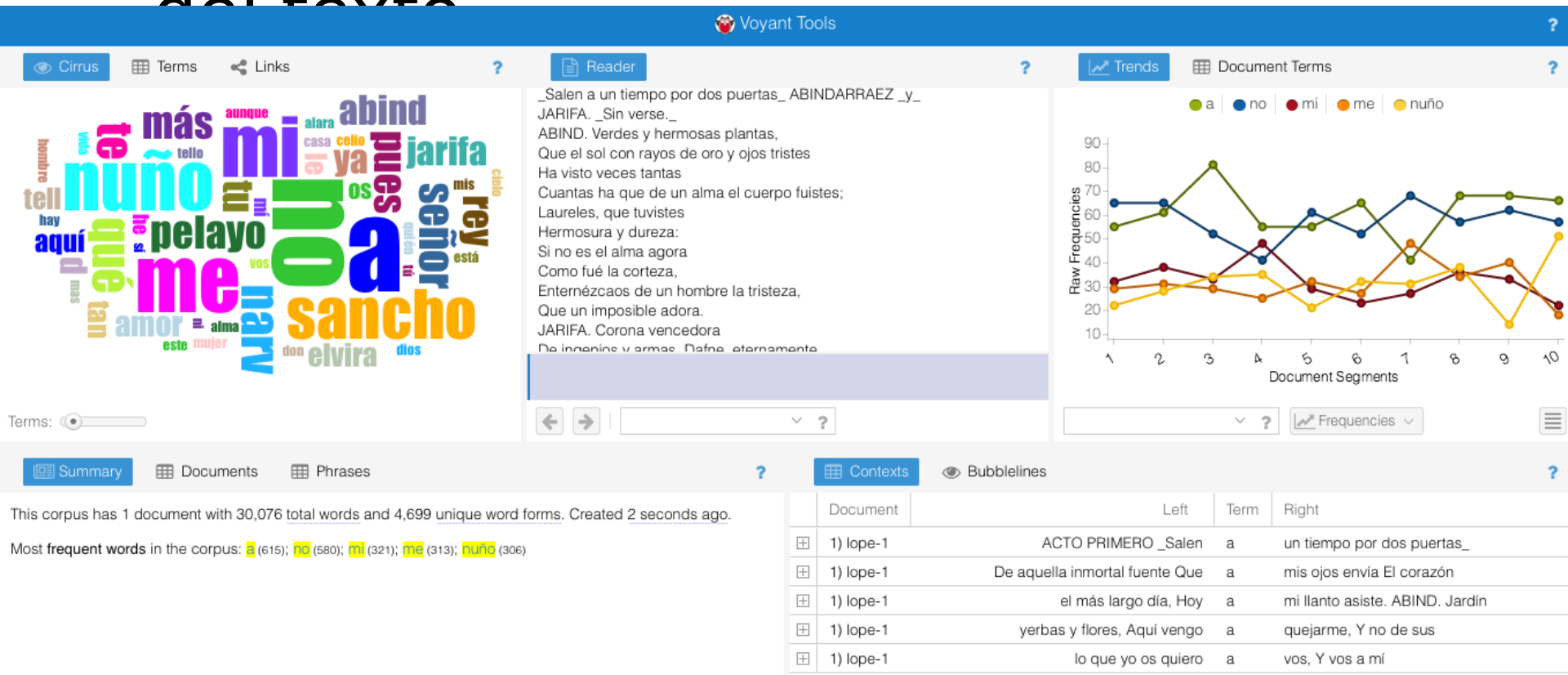
- es una aplicación web
- tiene funciones interesantes

VOYANT
see through your text



Otra herramienta útil
Voyant-tools

- Visualización palabras en nube
- Aparición de palabras en segmentos del texto

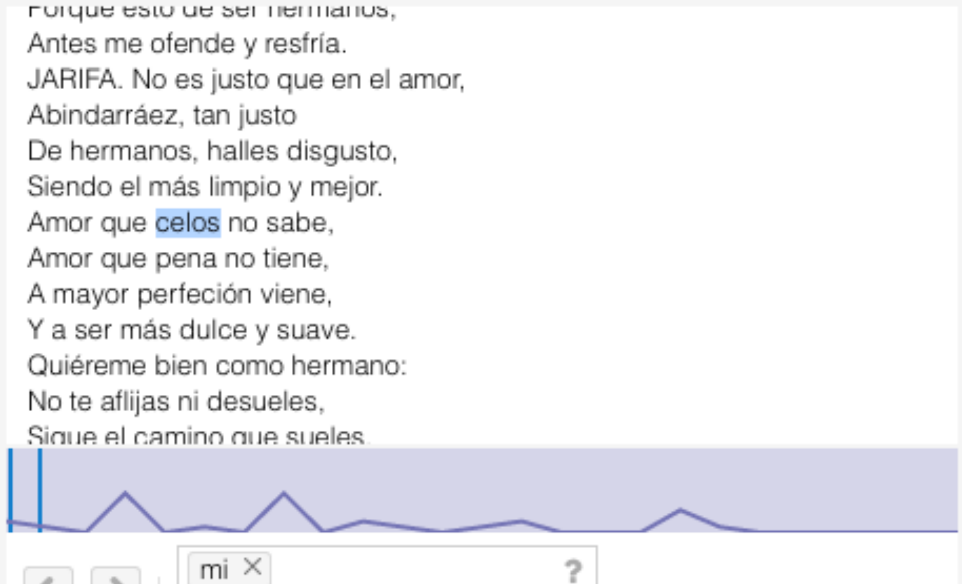


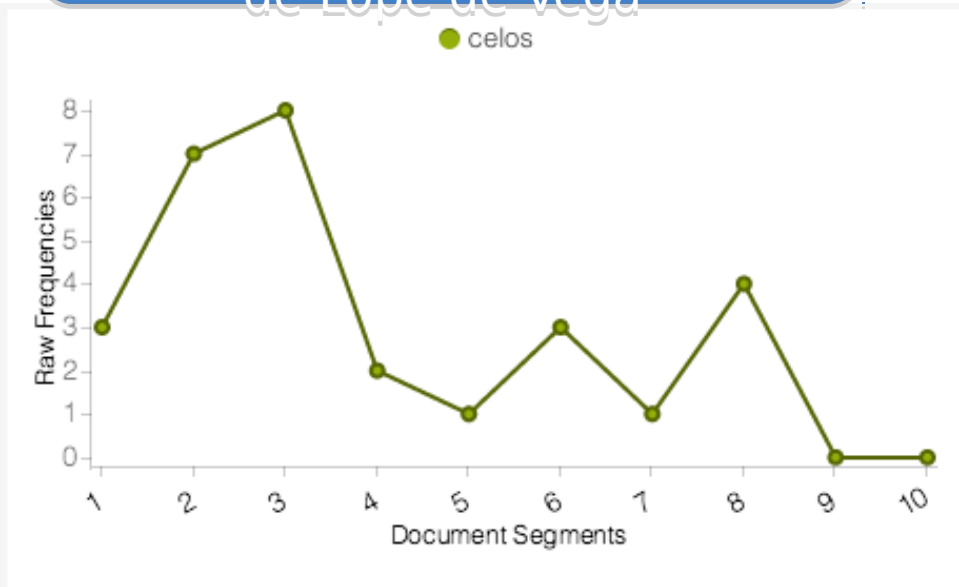
voyant-tools.org

Voyant Tools

Funciones similares, pero algunas cosas nuevas: frecuencias de palabras y por segmentos del texto

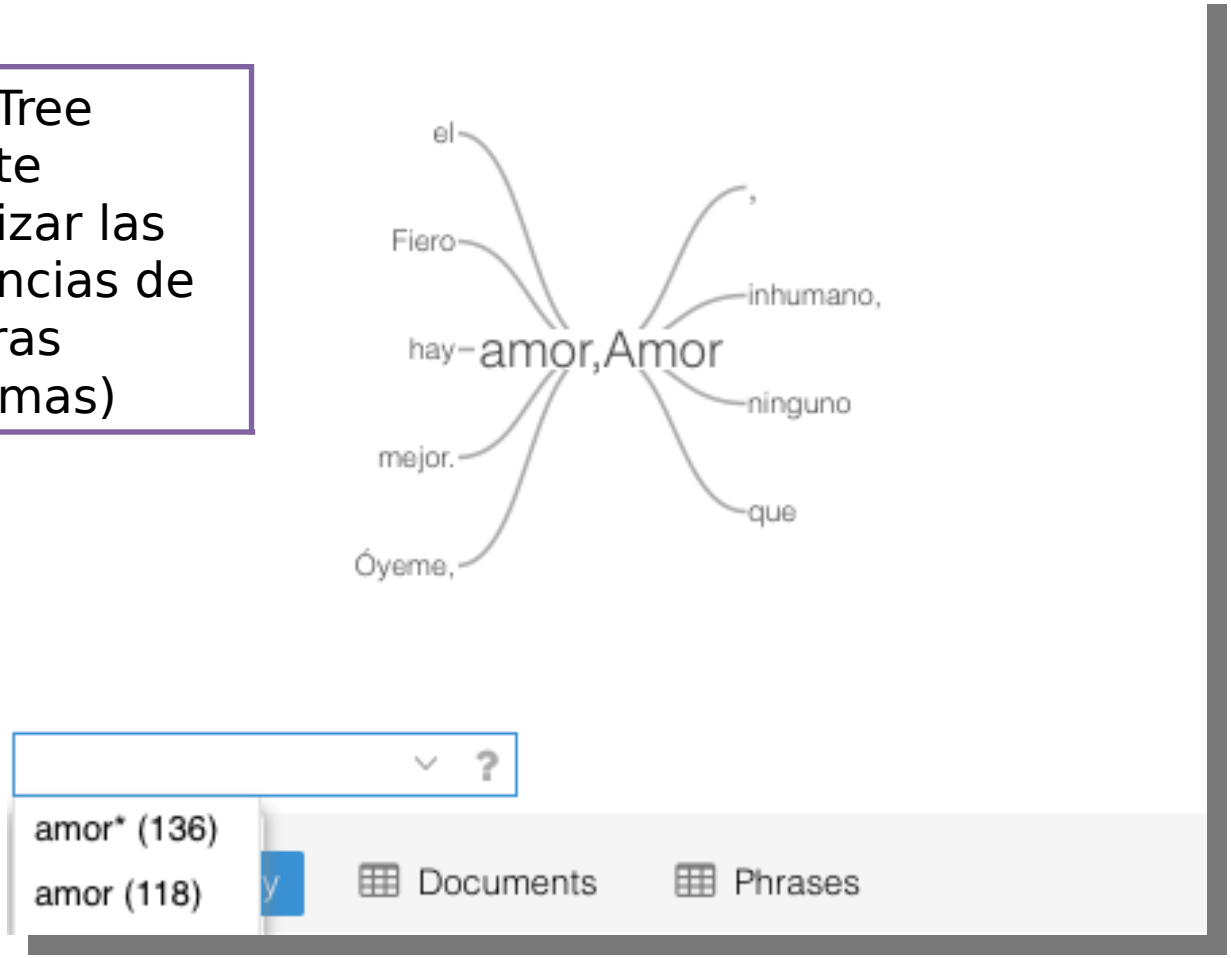
Aparición de la palabra “celos” en
El remedio en la desdicha, de Lope de Vega





voyant-tools.org

Word Tree permite visualizar las secuencias de palabras (bigramas)



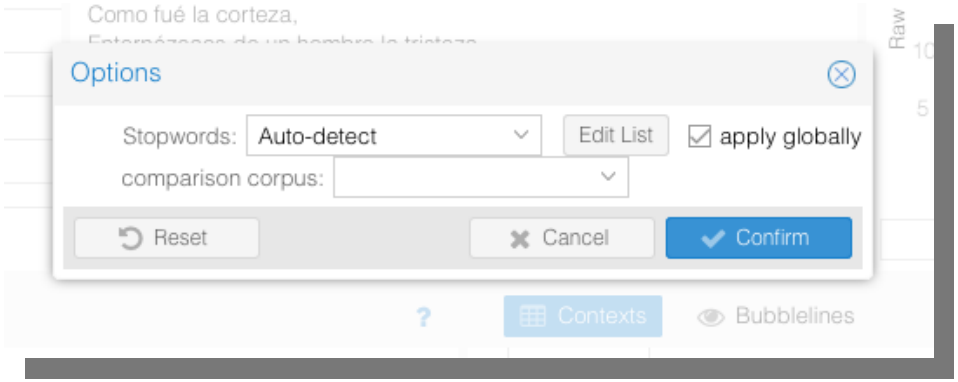
Atención!! palabras más frecuentes

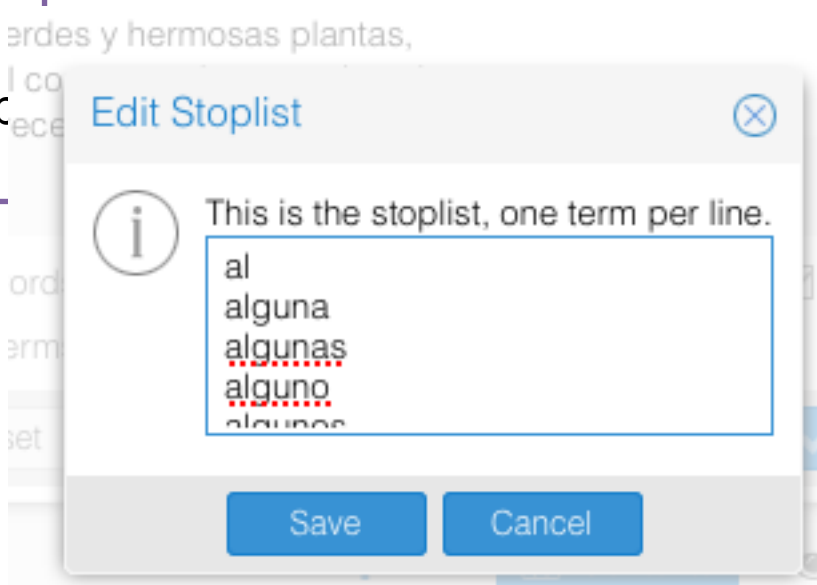
Voyant Tools

This corpus has 1 document with 30,076 total words and 4,699 unique word forms. Created 2 seconds ago.

Most **frequent words** in the corpus: a (615); no (580); mi (321); me (313); nuño (306)

Es posible eliminar palabras poco interesantes, como las preposiciones, las negaciones, pronombres. Se conocen como “palabras vacías” o “stop words”.



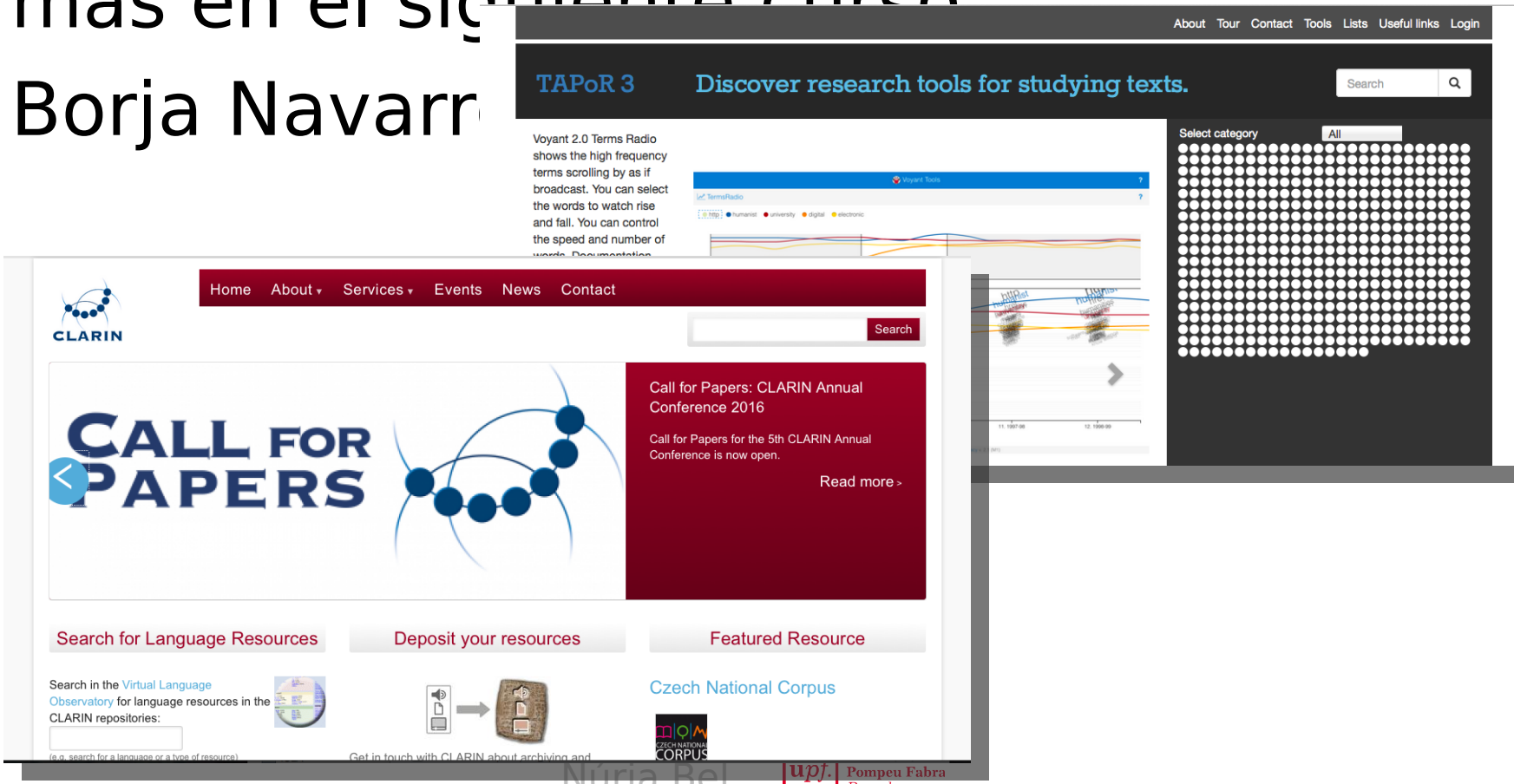


upf. Universitat Pompeu Fabra Barcelona

Más herramientas!!

Descúbrelas! TAPoR3, CLARIN.EU, y más en el siguiente curso:

Borja Navarr



Resumen y conclusiones (1/3)

- Procesamiento del Lenguaje Natural, una tecnología para el análisis de datos textuales, también para la filología (y la poesía).
- En principio, para grandes cantidades de datos, pero ... también para acelerar el análisis de cualquier texto.
- Cualquier texto, pero todavía hay que estar seguros de:
  - codificación de caracteres (utf8)
  - formato del documento (no pdf, mejor .txt)
- Actualmente, la digitalización de textos literarios no prioriza la creación de textos procesables.

Resumen y conclusiones (2/3)

- ✓ Anotación: añadir información con etiquetas específicas a una tarea. Para cada herramienta hay que buscar información sobre las etiquetas.
- ✓ Anotación: hay que verificar que el texto está escrito en la lengua que está prevista por la herramienta: palabras, ortotipografía, versos y mayúsculas, puede dar problemas ...
- ✓ El PLN para poesía podría dar herramientas más adecuadas: se necesitan recursos específicos (diccionarios, analizadores, etc.) para este género literario.

Resumen y conclusiones (3/3)
Dificultades

- ✓ Identificación y localización de herramientas, saber qué hay ...
- ✓ Instalación de herramientas, no apto para ...
- ✓ Comprender su “lenguaje”: etiquetarios, etc.
- ✓ Comprender sus limitaciones: qué lenguas, cobertura histórica, etc.
- ✓ Función de infraestructuras de investigación

Infraestructura Lingüística

Centro de competencias UPF-CLARIN

http://www.clarin-es-lab.org/

El análisis depende de la disponibilidad de “recursos lingüísticos”, listas de palabras y textos con información explícita.
Todos hemos de contribuir.

¿Tienes palabras? ¿Tienes textos?

Escríbeme a nuria.bel@upf.edu

Gracias!