

# Clarín K-Centre Spain as user-oriented infrastructure

Mikel Iruskieta (UPV/EHU–IXA Group) [mikel.iruskieta@ehu.eus](mailto:mikel.iruskieta@ehu.eus)

Núria Bel (UPF) [nuria.bel@upf.edu](mailto:nuria.bel@upf.edu)



DH@MADRID SUMMER  
SCHOOL 2017

Tecnologías semánticas y herramientas lingüísticas para  
Humanidades Digitales

Madrid, del 3 al 5 de julio de 2017

Disponible online

Sigue el curso de forma presencial u online

UNED

Fundación Uned

LINHD  
LABORATORIO DE INNOVACIÓN  
EN HUMANIDADES DIGITALES

CLARIN

erc

European  
Research  
Council

POSTDATA  
Poetry Standardization  
and Linked Open Data

# CLARIN-K Centre Spain

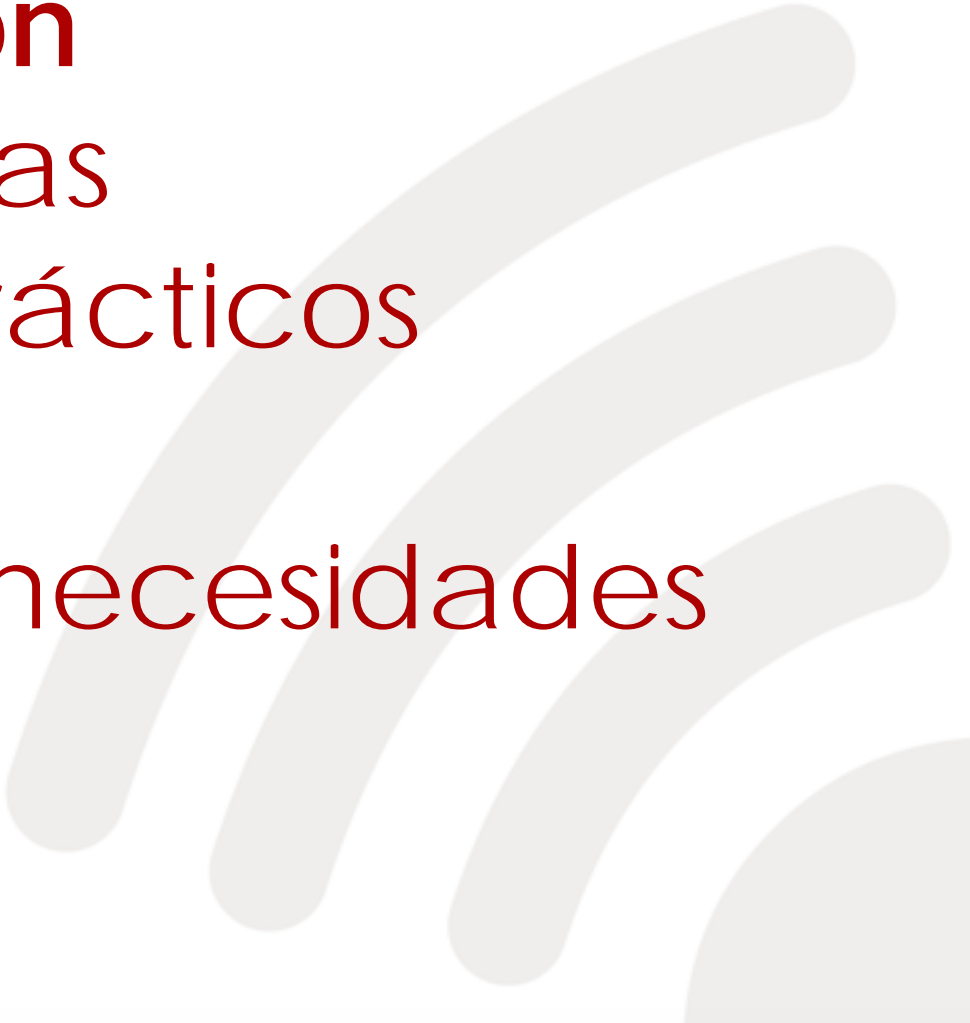
as user-oriented infrastructure

Núria Bel



Mikel Iruskieta



- 1. Introducción**
  2. Herramientas
  3. Ejemplos prácticos
  4. Ejercicios
  5. Análisis de necesidades
- 

# Palabras clave



- Análisis
- CLARIN
- Corpus
- HD
- Herramientas
- Infraestructuras de investigación
- Lingüística computacional
- PLN
- Textos



# Tareas a realizar



- Mi primer “corpus”
- Extraer diccionario
- Observar la concurrencia de palabras en textos
- Mi primer KWIC
- Uso de n-gramas y filtros personalizados
- Analizar necesidades

# Procesamiento del Lenguaje Natural (PLN)



- Objetivo: tareas sencillas como leer, contar, comparar y extraer información de textos digitales



- Pero...
  - ¿y si no tenemos conocimiento técnico?

- Procesar: reconocer unidades (palabras o *tokens*) y asignarles etiquetas (una representación o información)
- Representar: añadir (y/o sustituir) información explícita de unidades (*tokens*)
  - Permite realizar tareas
- ¿Qué información explícita? Depende de la tarea...

# PLN tareas

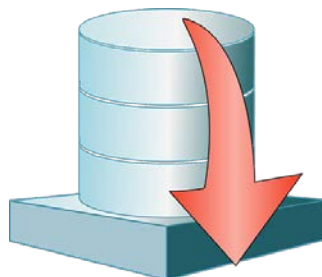
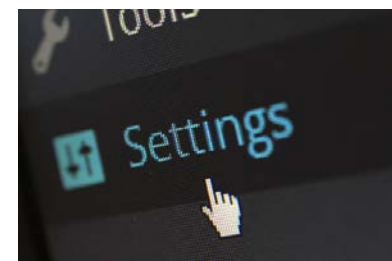
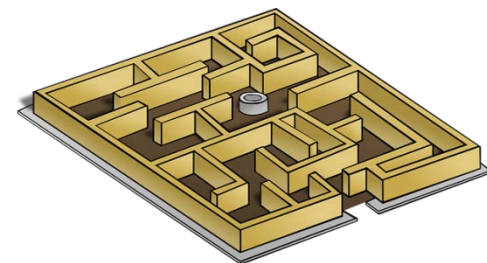


- Extracción de información
- Análisis lingüístico
- Corrección gramatical
- Traducción automática
- Resumen automático
- Simplificación de textos
- Análisis de opinión
- Respuesta a preguntas y asistentes virtuales



# ¿Qué va a cambiar?

- Investigación en Humanidades digitales
  - BUSCAR y BUSCAR
    - INSTALAR
- Cambio de paradigma
  - PEDIR
    - OFRECER



# CLARIN-ERIC (2012)



- **Common Language Resources and Technology Infrastructure: CLARIN** (Krauwert et al., 2014)

- <https://www.clarin.eu/>



- Infraestructura de investigación europea para las Humanidades y Ciencias Sociales que ofrece todos los servicios en una única página



# ¿Qué ofrece CLARIN?



- Datos lingüísticos digitales
  - textos escritos u orales, multimodales
- Herramientas avanzadas para
  - describir, analizar y comparar textos
- CLARIN *language resource switchboard*:
  - <http://weblicht.sfs.uni-tuebingen.de/clrs/#>
    - [Noticia](#)
- ¿Quién y cómo se realiza ese servicio?
  - Los Centros-K de conocimiento

# Participantes y observadores



- Participantes:
  - Austria, Bulgaria, Czech Republic, Denmark, Dutch Language Union, Estonia, Finland, Germany, Greece, Hungary, Italy, Latvia, Lithuania, The Netherlands, Norway, Poland, Portugal, Slovenia, Sweden
- Observadores
  - France, UK, USA (Carnegie Mellon University)
- Más información sobre [CLARIN centres](#)

# Centros-K de conocimiento



K-Centres	Type
<a href="#">Spanish CLARIN K-Centre</a>	Language centre: Spanish, Basque, Catalan and Galician
<a href="#">CLARIN K-Centre for Treebanking</a>	Treebanks
<a href="#">Phonogrammarchiv CLARIN K-Centre</a>	Audio-visual fieldwork
<a href="#">CLARIN K-Centre for Speech Analysis</a>	Speech analysis
<a href="#">CLARIN K-Centre DANSK</a>	Language centre: Danish
<a href="#">The CLARIN K-Centre for Language Learning</a>	Language learning and language disabilities
<a href="#">CLARIN K-Centre for Languages of Sweden</a>	Language centre: Swedish language, minority languages in Sweden
<a href="#">The CLARIN K-Centre of Lund University Humanities Lab</a>	Multimodal and sensor-based methods

# Spanish CLARIN-K: 25/02/2015



- Spanish CLARIN-K Centre (Bel et al., 2016)
  - <http://clarin-es.org/>



- Grupos que formamos el centro-K
  - IULA-UPF
  - IXA Group
  - LINHD-UNED
  - TALG



# ¿Qué ofrece Spanish CLARIN-K Centre?



- Herramientas y/o servicios en CLARIN
  - Servicio de asesoría de proyectos
  - Servicio de diseño de proyectos de **análisis de textos**:
    - Recomendación de herramientas
    - Adaptar herramientas a necesidades concretas



# ¿Qué pedimos?



- Una tarea bien definida
  - Un corpus (procesable)

“Deja que tus textos trabajen por ti”, tiene la misión de promover y asesorar el uso de tecnología y herramientas de análisis de textos en la investigación en Humanidades y Ciencias Sociales.

Centro de competencias CLARIN del IULA-UPF



# Corpus



- Un corpus es un conjunto de textos procesables con ciertas características similares.
- Gracias a dichas características es posible observar algún fenómeno lingüístico.

NOTA: Dependerá de la tarea, tamaño y la calidad, que el estudio sea útil, representativo o significativo.



XXVIII Edición UNED Cursos de verano

2017

1. Introducción
- 2. Herramientas**
3. Ejemplos prácticos
4. Ejercicios
5. Análisis de necesidades

# Corpora: para empezar



- Para empezar:
  - COCA (eng)
    - Videotutorial

- CORPES-XXI: Corpus del español
- Linguee corpus multilingüe

Corpus of Contemporary American English

SEARCH FREQUENCY CONTEXT HELP

CLICK FOR MORE CONTEXT

ID	Year	Source	Genre	Word	Context
1	1993	ACAD	AmerEbnickits	however	us with a general picture of departure movements; they do however have some invitation as sources of information. In the 27
2	1991	NEWS	WashPost	however	and "immersed" steel-and-concrete tubes. For the Anacostia, however, bridge would n't work... because it would have to
3	2013	MAG	NewRepublic	however	of his mistress. # Dave did warn about Boxhead, however, # bridge would the size of a Buick Skylark that haunted the
4	1992	ACAD	AcademicQs	however	answer the arguments of their opponents. In this case, however, # secret was sent to the PMLA (printed in 1989
5	2007	MAG	PopScience	however	has a grip on nuclear safety and security; if, however, # topic group in Russia obtained and deployed the polonium-21
6	2002	MAG	USNWR	however	patience in the same breath. There is little ambiguity, however, # about what at stake. Joint Chiefs Chairman Gen. Richard
7	2008	ACAD	Style	however	other hand, I used to think, now, however, # about # I have come to see. Debates over land
8	1990	ACAD	Church&State	however	of what is possible and realizable at any one time, however, # always # assessed from the perspective of the poor and
9	2004	ACAD	SchoolPsych	however	of attention problems has been found to be relatively common, however, # and attention problems have been shown to predict achievement
10	1998	ACAD	WorldAffairs	however	the country's national interests. Despite the problems, however, # and beyond the rhetoric about specific policies, there
11	1997	MAG	SkyTelescope	however	- The spacing between the reducer and chip is, however, # and changing # by even a millimeter degrades images. Also

Corpus del Español del Siglo XXI (CORPES)

Concordancias | Coapariciones | Configuración | Ayuda | Modo de cita | Sugerencias

Lema [ pues ] Forma [ ] Clase de palabra [ Todos ] Grafía original [ ] Subcorpus [ ] Proximidad [ ]

Proximidad Limpiar

Lema [ ] Forma [ ] Clase de palabra [ puntuación ] Distancia [ ] Intervalo [ 1 ] Izquierda [ ] Derecha [ ] Izquierda o derecha [ ]

Subcorpus Limpiar

Título [ ] Autor [ ] Fecha de clasificación [ ]

Origen [ Todos ] América España

Medio [ Todos ] Escrito Oral Tipología [ Todos ] Debate Discurso

Concordancia Estadística Nueva consulta

85 casos en 10 documentos.

REF.	(Clasificación, país)	CONCORDANCIA	Ordenar por:	Año ascendente	sin criterio
1	2001 Perú	congresal. ¿Cuán importante resulta este tema para los peruanos y las peruanas? Pues mucho, por lo menos aquí en Arequipa y en diferentes puntos de nuestro país.			
2	2001 Perú	Bueno, pues nos alegra de Sí.			
3	2001 Perú	Bueno, pues nos alegra Pues, sí, nos alegra de Sí.			
4	2001 Esp.	Pero porque FIFA prohíbe Sí, pues entonces no sé para qué Joserra			
5	2001 Esp.	Vale, pues ahora te lo pongo más fácil, resulta que al Dépor le ampara la ley y el presidente			
6	2001 Esp.	No está bien. Pues porque por esa norma de tres, también se podían negar a jugar los equipos que			
7	2001 Esp.	Pues . pues no.			

Algunos ya vienen con herramientas para visualizar la información.

Iruskietia & Bel

# Herramientas útiles



- CLARIN:
  - <http://weblicht.sfs.uni-tuebingen.de/clrs/#>
- VOYANT: visualización de datos
  - <http://voyant-tools.org/>
- AntConc (y etc.)
  - <http://www.laurenceanthony.net/software.html>
- MeaningCloud: visualización de datos
  - <https://www.meaningcloud.com/demo#>
- Text simplifier: simplificador de textos
  - <http://able2include.taln.upf.edu/>
- Free summarizer: resumidor automático
  - <http://freesummarizer.com/#summarizecontainer>
- Sintetizador de voz
  - <http://aholab.ehu.es/users/agustin/speechtech4all/tts/index.html>
- Más en <http://tapor.ca/home>



# Freeling (Padró & Stanilovsky, 2012)



- Herramienta multilingüe muy útil y versátil
  - English, Spanish, Portuguese, Italian, French, German, Russian, Catalan, Galician, Croatian, Slovene...
  - Demo: <http://nlp.lsi.upc.edu/freeling/demo/demo.php>

## FreeLing Home Page

*Hooked on a FreeLing*

### 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:

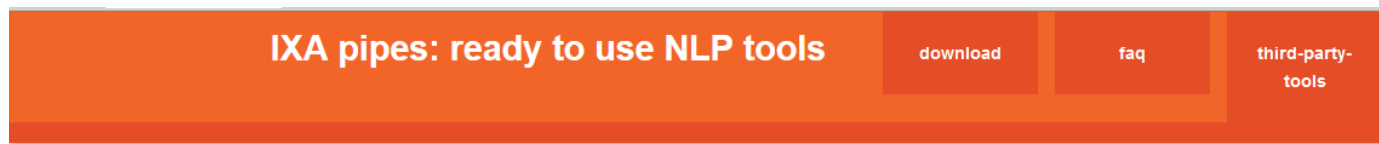
- Lluís Padró and Evgeny Stanilovsky.  
**FreeLing 3.0: Towards Wider Multilinguality**  
Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA.  
Istanbul, Turkey, May, 2012.  
[\[pdf\]](#) [\[bibtex\]](#)

# IXA-pipes (Agerri et al., 2014)



- Herramienta multilingüe adaptable

- Basque, Dutch, English, French, Galician, German, Italian and Spanish



IXA pipes is a modular set of Natural Language Processing tools (or pipes) which provide **easy access to NLP technology for several languages**. It offers robust and efficient linguistic annotation to both researchers and non-NLP experts with the aim of lowering the barriers of using NLP technology either for research purposes or for small industrial developers and SMEs. The *ixa pipes* can be used or exploit its modularity to pick and change different components. The tools are developed by the [IXA NLP Group](#) of the [University of the Basque Country](#).

## NEWS

Release [1.1.1 of Ixa pipes](#) is now available!!

IXA pipes are in [Maven Central](#)!!

## ixa pipes

If you use the *ixa pipes* tools or the models, **please cite this paper**

Rodrigo Agerri, Josu Bermudez and German Rigau (2014): "IXA pipes: Ready to Use Multilingual NLP tools", in: Proceedings of the 9th International Conference on Natural Language Resources and Evaluation Conference (LREC2014), 26-31 May, 2014, Reykjavik, Iceland. [PDF paper](#)

[ixa-pipe-tok](#): Tokenizer and Segmenter for several languages.

[ixa-pipe-pos](#): Statistical POS tagging and Lemmatizer for Basque, Dutch, English,

Freeling y IXA-pipes son herramientas útiles, pero, es verdad, que su manejo puede resultar complejo.

# Spanish CLARIN-K Centre: herramientas



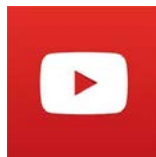
Próximamente también para el Gallego

## ANALHITZA



### ANALHITZA

Laboratorio de Innovación en Humanidades Digitale...  
✓ Harpidetuta 181

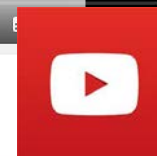


## CONTAWORDS



### Contawords

Laboratorio de Innovación en Humanidades Digitale...  
✓ Harpidetuta 181



# CONTAWORDS powered by IULA-UPF



ContaWords



English



Count!

More info.

FAQ

Credits

Your workspace is empty.



Upload local files

or



Use files from the Internet

Spanish, English, Catalan,  
Portuguese, Italian

Workspace

External files

Tell us the documents language... :)

Run now!



# E1: Mi primer corpus



- Hacer/buscar/guardar más de un texto
  - Formatos PDF o TXT
- Reformatear el PDF
  - De PDF a TXT (UTF8)
  - ¿Hay que “limpiarlo” manualmente?
- De texto a corpus
  - Unir archivos en Windows
    - `for %f in (*.txt) do type "%f" >> corpus.txt`

# ANALHITZA powered by IXA-CLARIN-K



Home

Our products

Videos

Publications

Services



EU EN ES

Basque, Spanish, English



**ANALHITZA**  
Powered by CLARIN

ANALHITZA will help you extracting from text in Basque, Spanish or English, some linguistic information, such as:

- nouns, adjectives, verbs, adverbs...
- person names, location names...
- sequences of two, three and four words
- ... and much more!

The text could be the one that you have in a file, something that you will copy it here, or from a web page, but it should be encoded in UTF8. To use ANALHITZA, enter the text you want to analyze using one of the 3 below options, and then choose the language of your text (Basque, Spanish or English). After waiting a moment, you will get the results on an Excel file. Thus, you will be able to adapt the results to meet your requirements.

Upload file (txt format)

Insert text

Insert url

# Etiquetas EAGLES (y FREELING)



- Codifican la información en una secuencia donde la posición se relaciona con el atributo del que se codifican los valores.
- La primera para la categoría gramatical.
  - Si no valor se pone '0'.
  - Cada forma lingüística tiene posiciones/valores diferentes
- Etiquetas para las lenguas europeas

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

Position		Values
0	A	adjective
0	D	determiner
0	F	Punctuation
0	N	noun
0	P	pronoun
0	R	adverb
0	S	adposition
0	V	verb
0	W	date
0	Z	number
Etiquetas <b>Freeling</b>		

# E2: Extraer diccionario con ANALHITZA



- Extraer los 5 nombres más frecuentes de
  - “One love” Bob Marley (en)
    - <http://www.metrolyrics.com/one-love-lyrics-bob-marley.html>
  - “Me cago en el amor” Tonino Carotone (it/es)
    - <https://www.letras.com/tonino-carotone/6998/>

# R-E2: Extraer palabras con ANALHITZA



- 5 nombres más frecuentes

Marley		Carotone	
9	love	12	amor
7	heart	4	mondo
6	right	4	culpa
5	lord	3	momenti
4	thanks	3	futuro

En la canción de Carotone el sistema ha detectado nombres fuera del diccionario y no ha sabido lematizar:  
Momenti (n, pl) >  
momento (n, sg)

# Experiencias con ANALHITZA



- Comparar riqueza léxica por edades
- Extracción del diccionario por años en una colección de cuentos de Educación Infantil (4-6 años)
- Análisis del lenguaje coeducativo en cuentos modernos
- Elegir cuentos según sus características
- Competiciones: analizar la lengua en textos de educación infantil
- Observar si se conocen las palabras de una lectura
- Buscar características de un corpus para su evaluación

Otras referencias de interés:  
a) Villegas et al. (2012),  
b) Gabrielatos (2005)



XXVIII Edición UNED Cursos de verano 2017

1. Introducción
2. Herramientas
3. **Ejemplos prácticos**
4. Ejercicios
5. Análisis de necesidades




# Artext: editor y ayuda en línea

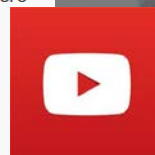


- Revisión y sugerencias (entre otras cosas)
  - de oraciones largas
    - ofrece soluciones posibles
  - de marcadores repetidos
    - propone marcadores similares
  - de concordancia verbal

Ayuda a la redacción de textos en español de ámbitos especializados.  
Editor en línea para añadir fraseología relacionada con el tipo de texto.

Elige un género textual

 <b>Administración pública</b>	 <b>Medicina</b>	 <b>Turismo</b>
→ Alegación	→ Artículo de investigación	→ Artículo de divulgación
→ Carta de presentación	→ Artículo de revisión	→ Entrada en blog de viajero
→ Queja	→ Historia clínica	→ Informe
→ Reclamación	→ Resumen de artículo de investigación	→ Normativa
→ Solicitud	→ Trabajo de Fin de Grado (TFG)	→ Plan de negocio





# Voyant Tools (1/7)



Voyant Tools

Cirrus Terms Links Reader TermsBerry Trends Document Terms

Los vestidos nuevos del emperador Cuento  
 Un cuento de Hans Christian Andersen Andersen  
 9.4/10 - 552 votos  
[http://www.andersenstories.com/es/andersen\\_cuentos/los\\_vestidos\\_nuevos\\_del\\_emperador](http://www.andersenstories.com/es/andersen_cuentos/los_vestidos_nuevos_del_emperador)  
 Los vestidos nuevos del emperador  
 Hace de esto muchos años, había un Emperador tan aficionado a los trajes nuevos, que gastaba todas sus rentas en vestir con la máxima elegancia. No se

Segment	a	no	le	había	más
1	50	35	20	15	10
2	75	30	15	15	15
3	55	40	15	15	10
4	75	35	20	10	10
5	75	25	25	10	10
6	75	30	15	10	10
7	70	20	15	20	10
8	75	25	20	10	10
9	65	25	15	15	10
10	60	30	15	15	10

Terms:

Summary Documents Phrases Contexts Bubblelines Correlations

This corpus has 1 document with 29,475 total words and 5,286 unique word forms. Created now.  
 Vocabulary Density: 0.179  
 Average Words Per Sentence: 19.9  
 Most frequent words in the corpus: **a** (703); **no** (318); **le** (180); **había** (141); **más** (133)

items:

Document	Left	Term	Right
1) 1498...	había un Emperador tan aficionado	a	los trajes nuevos, que gastaba
1) 1498...	de paseo por el campo,	a	menos que fuera para lucir
1) 1498...	bulliciosa. Todos los días llegaban	a	ella muchísimos extranjeros, y una
1) 1498...	milagrosa virtud de ser invisibles	a	toda persona que no fuera

703 context expand

# Voyant Tools (2/7)



XXVIII Edición UNED Cursos de verano **2017**

Cirrus

Terms

Links



25 palabras

Terms:

105 palabras



Terms:

Iruskietia & Bel

# E3: Voyant Tools (3/7)



- Descargué los siguientes los siguientes cuentos
  - Personajes femeninos (4):
    - 1) [Sirenita\\_FA](#), 2) [Niña-fosforos\\_FA](#), 3) [Princesa-guisante\\_FA](#) y 4) [Reina-nieves\\_FA](#)
  - Personajes masculinos (4):
    - 1) [Soldadito-plomo\\_MA](#), 2) [Yesquero\\_MA](#), 3) [Emperador-vestido\\_MA](#) y 4) [Sastrecillo\\_MA](#)
- Diga cuales de estas palabras aparecen en más cuentos tradicionales:
  - Soldado (58)
  - Margarita (105)
  - Princesa (54)
  - Príncipe (49)

# R-E3: Voyant Tools (4/7)



- Diga cuales de estas palabras aparecen en más cuentos tradicionales:
  - Soldado (58): 2 textos
  - Margarita (105): 1 texto
  - Princesa (54): 6 textos
  - Príncipe (49): 3 textos

# Voyant Tools: distinctive words (5/7)



- Compare que palabras aparecen en un texto y no en otro

Most frequent words in the corpus: **a** (703); **no** (318); **le** (180); **había** (141); **más** (133)

Distinctive words (compared to the rest of the corpus):

1. emperador-vestidos\_MA: **emperador** (24), **telar** (8), **cargo** (6), **tejedores** (5), **lleva** (5).
2. niña-fosforos\_FA: **fósforos** (6), **frio** (8), **fósforo** (5), **niña** (8), **cerilla** (3).
3. princesa-guisante\_FA: **guisante** (5), **edredones** (2), **colchones** (2), **veinte** (3), **princesa** (11).
4. reina-nieves\_FA: **margarita** (105), **carlos** (51), **comeja** (27), **nieves** (25), **reno** (20).
5. sastrecillo-valiente\_MA: **sastrecito** (41), **gigante** (17), **golpe** (15), **gigantes** (11), **siete** (11).
6. sirenita\_FA: **mar** (51), **sirena** (40), **príncipe** (41), **barco** (17), **sirenita** (14).
7. soldadito-plomo\_MA: **soldado** (21), **plomo** (17), **soldadito** (6), **pierna** (4).
8. yesquero\_MA: **soldado** (37), **yesquero** (18), **plomo** (17), **pierna** (4).

Se puede hacer lo mismo dentro de un texto para observar la progresión temática

## E4: Voyant Tools: mi primer KWIC (6/7)



- Realice un KWIC para describir el uso de la puntuación del marcador “pues”
  - Descargue [este corpus](#)
    - Se pueden unir los archivos con:
      - `for %f in (*.txt) do type "%f" >> corpus.txt`
  - Descarte todas las estructuras que tengan un signo de puntuación a la izquierda de “pues”

# R-E4: Voyant Tools: KWIC (resultado) (7/7)



- El KWIC de “pues” + puntuación

Doc.	Left	Term	Right
5	pago del milagroso brebaje. - ¡Sea,	pues	! -dijo la sirena; y la
3	el sentido que encerraba. Contó,	pues	, a la corneja toda su
5	la quilla del navío. Llegó,	pues	, el día en que la
4	qué asustarse con dos. Así,	pues	, el sastrecito se puso en
0	y digno ministro se presentó,	pues	, en la sala ocupada por
3	nada de Carlos. ¿Qué decía,	pues	, la azucena de fuego? - Oye
	aquí quien pueda enfrentársele.		
4	Tomaron,	pues	, la decisión de presentarse al
5	mundo; no volverán a encontrarse	pues	, mientras que yo estoy a
5	humanos de allá arriba. - Así,	pues	, ¿moriré y vagaré por el

Buscar ejemplos o diseñar ejercicios de fenómenos lingüísticos complejos puede ser sencillo con corpus y herramientas adecuadas.

# Consideración



- Todo esto lo hemos hecho manualmente con corpus pequeños
  - Problemas:
    - A veces las cuentas no cuadran, hay que volver a empezar
    - Falta de ejemplos adecuados
    - Representatividad





XXVIII Edición UNED Cursos de verano  
**2017**

1. Introducción
2. Herramientas
3. Ejemplos prácticos
- 4. Ejercicios**
5. Análisis de necesidades

# E5: observar como se adjetivan “mujer” y “hombre” en 4gramas



- ¿Es adecuada el uso de la lengua de los cuentos coeducativos?
  - ¿Cómo se adjetivan las palabras “mujer” y “hombre”?
- Descargue esta [recopilación de cuentos coeducativos](#)
- Cambie el formato de PDF a [TXT](#)
- Analícelo con ANALHITZA y guarde la [hoja de cálculo](#)
- Para crear un autofiltro personalizado siga los siguientes pasos:
  - Inserte una línea para introducir títulos
  - Seleccione todas las columnas y haga clic en la hoja de cálculo “datos>filtro”
  - Haga clic en el triángulo del título y seleccione “filtros de texto > filtro personalizado”
  - El filtro debe contener el lema “mujer” u “hombre”

Se puede hacer directamente en Adobe

# R-E5: ANALITZA: autofiltros



- La búsqueda no es significativa

¡Siga buscando!

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	frec	lema	ca	lema	ca	lema	ca	lema	ca						
498	1	mujer	N	,	O	y	C	relatar	V						
636	1	mujer	N	no	A	est	V	n	N						
1621	1	mujer	N	hacer	V	rar	V	r	V						
2403	1	mujer	N	para	P	su	D	hijo	N						
2582	1	mujer	N	seguir	V	ser	V	uno	Q						
2946	1	mujeres	R	de	P	barbanegra	R	10	O						
3077	1	hombre	N	tan	A	inteligente	G	como	C						
3348	1	mujer	N	de	P	su	D	casa	N						
7047	1	hombre	N	:	O	el	D	mujer	N						
7473	1	mujer	N	ser	V	dulce	G	,	O						
8620	1	mujer	N	y	C	aquel	D	ni	O						
8918	1	mujer	N	como	C	que	D	ni	O						
10513	1	hombre	N	de											

**Autofiltro personalizado**

Mostrar las filas en las cuales:

lema

contiene [mujer]

Y  O

contiene [hombre]

Use ? para representar cualquier carácter individual  
Use \* para representar cualquier serie de caracteres

Aceptar Cancelar

Verificar siempre los resultados.  
En este caso el sistema tiene errores con los acentos.

Iruskietta & Bel

# ¡Ahora os toca!



- ¿Cómo extender la búsqueda para tener más datos?
  - ¿Quizá con personajes? ¿Oficios?
- ¿Qué otra pregunta puede ser interesante con este corpus? ¿Y en otro corpus?
  - Puede valer para observar como utiliza un determinado autor una palabra clave
- ¿Qué más?



XXVIII Edición UNED Cursos de verano  
**2017**

1. Introducción
2. Herramientas
3. Ejemplos prácticos
4. Ejercicios
5. **Análisis de necesidades**

# Análisis de las necesidades (1/2)



- ¿Qué es lo que habéis hecho o queréis hacer?
  - ¿Si se hubiera hecho con un corpus o con alguna herramienta automática tendría algún valor añadido?

# Análisis de las necesidades (2/2)



- ¿Sabrías diseñar la herramienta que necesitáis?
  - ¿Cuál es la herramienta que necesitáis?
  - ¿Tenéis algún corpus para hacer algún estudio y su posterior evaluación?

# Invitación de colaboración



- El análisis depende de la disponibilidad de “recursos lingüísticos”, listas de palabras y textos con información explícita.
- Tod@s podemos contribuir.
- ¿Tienes textos?
  - Escríbenos a [clarinkcenter@gmail.com](mailto:clarinkcenter@gmail.com)

Encuesta anónima



# Bibliografía

# Bibliografía (1/2)



- Agerri, R. Bermudez, J. Rigau, G. 2014. "IXA pipeline: Efficient and Ready to Use Multilingual NLP tools", in: Proceedings of the 9th LREC, Reykjavik, Iceland. [PDF paper](#)
- Bel, N, González-Blanco, E. Iruskieta, M. 2016. "CLARIN Centro-K-español." *Procesamiento del Lenguaje Natural* 57 (2016): 151-154.
- Bel, N. 2016. **Taller PLN y sus aplicaciones en poesía.** Curso de Verano LINHD-UNED. <https://www.youtube.com/watch?v=XkqT1MJTosA>
- da Cunha, I. Montané, M. Amor; Hysa, L. 2017. "The arText prototype: An automatic system for writing specialized texts". En [15th Conference of the European Chapter of the Association for Computational Linguistics \(EACL 2017\). Demonstrations Session](#). Association for Computational Linguistics. Valencia (España).

# Bibliografía (2/2)



- Gabrielatos, C. 2005. [Corpora and Language Teaching: Just a Fling or Wedding Bells?.](#) *TESL-EJ* 8.4.
- Padró, L. Stanilovsky, E. 2012. **FreeLing 3.0: Towards Wider Multilinguality.** Proceedings of the LREC. Istanbul, Turkey.
- Sinclair, S. Rockwell, G. Voyant Tools. Available online: <http://docs.voyant-tools.org/> (accessed on 20 June 2017).
- Otegi, A. Imaz, O. Díaz de Ilarraza, A. Iruskieta, M. Uria, L. 2017. [ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research.](#) *Procesamiento del Lenguaje Natural* 58, pp. 77-84.
- Villegas, M. Bel, N. Gonzalo, C. Moreno, A. Simelio, N. 2012. Using Language Resources in Humanities research. In LREC 2012, pages 3284-3288.

mikel.iruskieta@ehu.eus  
<http://ixa2.si.ehu.es/iruskieta/>

nuria.bel@upf.edu  
[www.upf.edu/web/nuria-bel](http://www.upf.edu/web/nuria-bel)



# Clarín K-Centre Spain as user-oriented infrastructure

Mikel Iruskieta (UPV/EHU–IXA Group) [mikel.iruskieta@ehu.eus](mailto:mikel.iruskieta@ehu.eus)

Núria Bel (UPF) [nuria.bel@upf.edu](mailto:nuria.bel@upf.edu)



DH@MADRID SUMMER  
SCHOOL 2017

Tecnologías semánticas y herramientas lingüísticas para  
Humanidades Digitales

Madrid, del 3 al 5 de julio de 2017

Disponible online

Sigue el curso de forma presencial u online

UNED

Fundación Uned

LINHD  
LABORATORIO DE INNOVACIÓN  
EN HUMANIDADES DIGITALES

CLARIN

erc

European  
Research  
Council

POSTDATA  
Poetry Standardization  
and Linked Open Data