Tecnologías del lenguaje para la enseñanza e investigación en Humanidades Digitales

Mikel Iruskieta and Arantza Diaz de Ilarraza

IXA Group (UPV/EHU)
Universidad del País Vasco
IXA-CLARIN-K CENTRE
La Habana, 26 / Enero / 2019 / Cuba

Resumen

El acercamiento digital al estudio de las humanidades ofrece nuevas oportunidades para la enseñanza, la colaboración, la reutilización de herramientas y la difusión multimodal de estos estudios. Nuevas actividades, objetos de estudio y técnicas de investigación han propiciado nuevas formas para leer, escribir, revisar, buscar, ordenar, describir, enseñar e investigar

Agradecimientos a

- Yudislaidis Exposito por el ofrecimiento de varios textos en formato digital
- Joseba Sarrionandia por sus sugerencias literarias
- Sulema Rodriguez Roche por su atención y toda su ayuda para que en Cuba me sienta como en casa
- Grupo Ixa, sobre todo a **Arantxa Otegi**, Esther Miranda y Kike Fernandez

Por supuesto, todos los errores son nuestros!

¿Qué haremos?

- Entender el contexto de uso de infraestructuras como CLARIN y como podemos ayudar/hacer Humanidades Digitales
- Hacer ejercicios básicos de PLN
  - Descargar un libro digitalizado sin OCR con anotaciones manuales, procesarlo y extraer información lingüística
  - Ver ejemplos de cómo limpiar texto con expresiones regulares
  - Ver algunas ventajas de Python para lingüistas
  - Uso de herramientas de PLN para analizar corpus propio y usarlo en el aula o en la investigación
- ¿Discutir o valorar necesidades? ¿Colaborar?

Contexto de las Humanidades Digitales

- Oportunidades
  - Cambio y conservación, globalización e identidad
  - Entornos digitales
  - Múltiples registros y textos multimodales
  - Aprehensión de la realidad: más datos computables y mejor relacionados
- Retos
  - Brecha digital: proyectos de investigación y uso docente o público
  - Mantenimiento de recursos digitales
  - Mayor necesidad en datos, hardware, conexión

Humanidades Digitales y cómo llegamos a ellas

- Las HD son la aplicación de recursos y métodos digitales a las preguntas humanísticas
  - Pero también un género textual
  - También implica nuevos valores, competencias digitales y modos de trabajar o de diseminar el trabajo
  - Implica un contexto enriquecido: digital vs analógico
- ¿Cómo no hemos acercado a las HD?
  - Desde el Procesamiento del Lenguaje Natural
  - Se desarrollan sistemas, pero ¿son útiles en otras esferas?
  - Cuando queremos ayudar en las HD desde el PLN... Nos dicen: "con el departamento de lingüística"

HD: desde el compromiso con la sociedad

- Primera fase:
  - La tecnología ayudará en tareas repetitivas y de cantidad
- Segunda fase:
  - La tecnología condiciona los resultados de la investigación, se debe observar de un modo crítico
  - Las HDs o infraestructuras tecnológicas deben ayudar
    - desde un diseño universal a disminuir y erradicar la brecha digital
    - conservar el patrimonio (digitalizándolo) y la identidad cultural/lingüística

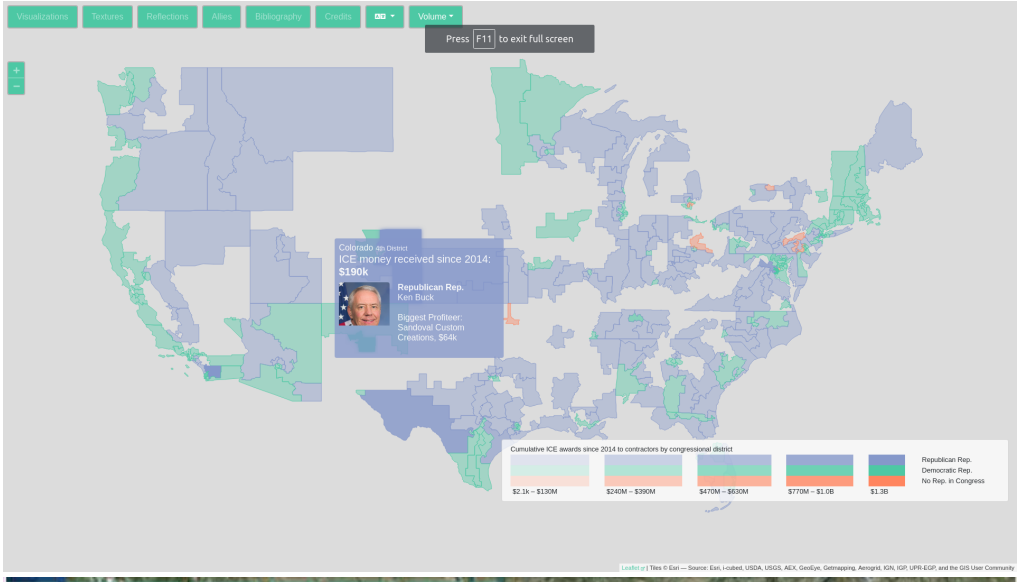
Trabajamos desde la creencia que las HD pueden ayudar en el empoderamiento y la democratización (con sus riesgos y errores)

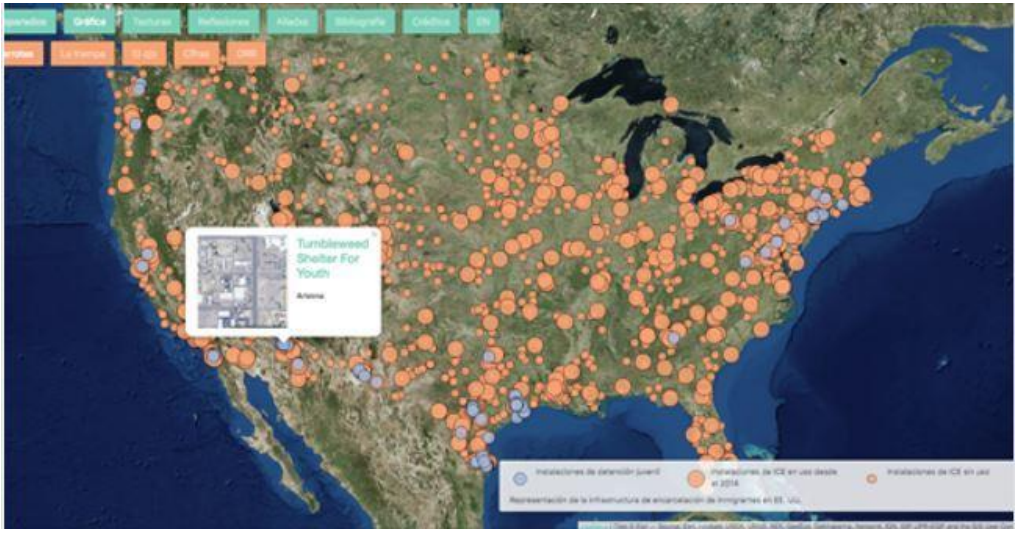
HD: ¿principios y evolución?

- Primer trabajo de HD y primera etapa:
  - Roberto Blusa empezó en 1946 realizando un corpus de Santo Tomas de Aquino y desde el 2006 lo están enriqueciéndolo con información semántica y sintáctica.
- Trabajo de HD en su fase crítica y comprometida
  - Torn Apart / Separados
    - "Política de tolerancia cero" quiere mostrar la crisis humanitaria e incidir en la opinión pública
    - Enseñar donde están los centros de detención, quien los financia (persona y partido)
    - Multilingüe: inglés, francés, castellano

Denuncia social, publicación académica digital, multimodalidad (texto, mapas, imágenes), inter-disciplinar (geografía, política, PLN), compromiso (derechos humanos y opinión pública)

Torn apart


Visualizations
Textures
Reflections
Allies
Bibliography
Credits
Volume
Press F11 to exit full screen
Colorado 4th District
ICE money received since 2014:
$190k
Republican Rep.
Ken Buck
Biggest Profiteer:
Sandoval Custom
Creations, $64k
Cumulative ICE awards since 2014 to contractors by congressional district
$2.1k – $130M
$240M – $390M
$470M – $630M
$770M – $1.0B
$1.3B
Republican Rep.
Democratic Rep.
No Rep. in Congress


Tumbleweed
Shelter For
Youth

HD: referencias importantes

- Libro de HD:
  - *A Companion to Digital Humanities* [Schreibman et al., 2004]
- Asociación de HD:
  - *Alliance of Digital Humanities Organization* (ADHO). 2005
- Fundación de HD:
  - *National Endowment for the Humanities*. 2006
- Y ¿en Cuba?
  - Grupo de HD en Cuba: `http://hdcuba.linhd.es/humanidades-digitales/`
  - ¿Qué HD son posibles en Cuba? [Rodríguez-Roche, 2018]
  - Oportunidades de la triada Bibliotecología-Lingüística-PLN

Humanidades Digitales: antropología y música digital

- En las humanidades digitales además de los estudios humanísticos tradicionales, se estudian nuevos temas, con nuevas metodologías y herramientas
  - Por ejemplo, se puede analizar o comparar al detalle en que se diferencia una misma pieza entre dos personas (o instrumentos)
  - O cómo ha cambiado la música con la digitalización
  - Otro ejemplo, es estudiar los comportamientos de las personas y la comunicación en las redes sociales
  - O se puede estudiar si las herramientas o técnicas empleadas son o debieran ser neutrales

Humanidades y lenguaje: soportes para enseñar (idiomas)

- Pizarra
- Imprenta y texto escrito
- Grabación de voz (casete y CD-ROM)
- Enseñanza de lenguas asistida por computador (CALL)
- Internet: navegadores, correo electrónico, web
- Web 2.0: redes sociales, blogs
- Aparatos móviles: tablets y teléfonos
- LMS: Moodle...
- Corpus, analíticas, métricas, big data

Herramientas en la escuela digital

En qué grado se usan en la escuela? En la enseñanza de lenguas?

- LMS: Moodle
- Web 2.0
- Mobile learning
- Entorno Personal de Aprendizaje (PLE)
- MOOCs
- Learning Analytics
- Aplicaciones: multimedia, gamificación, material interactivo

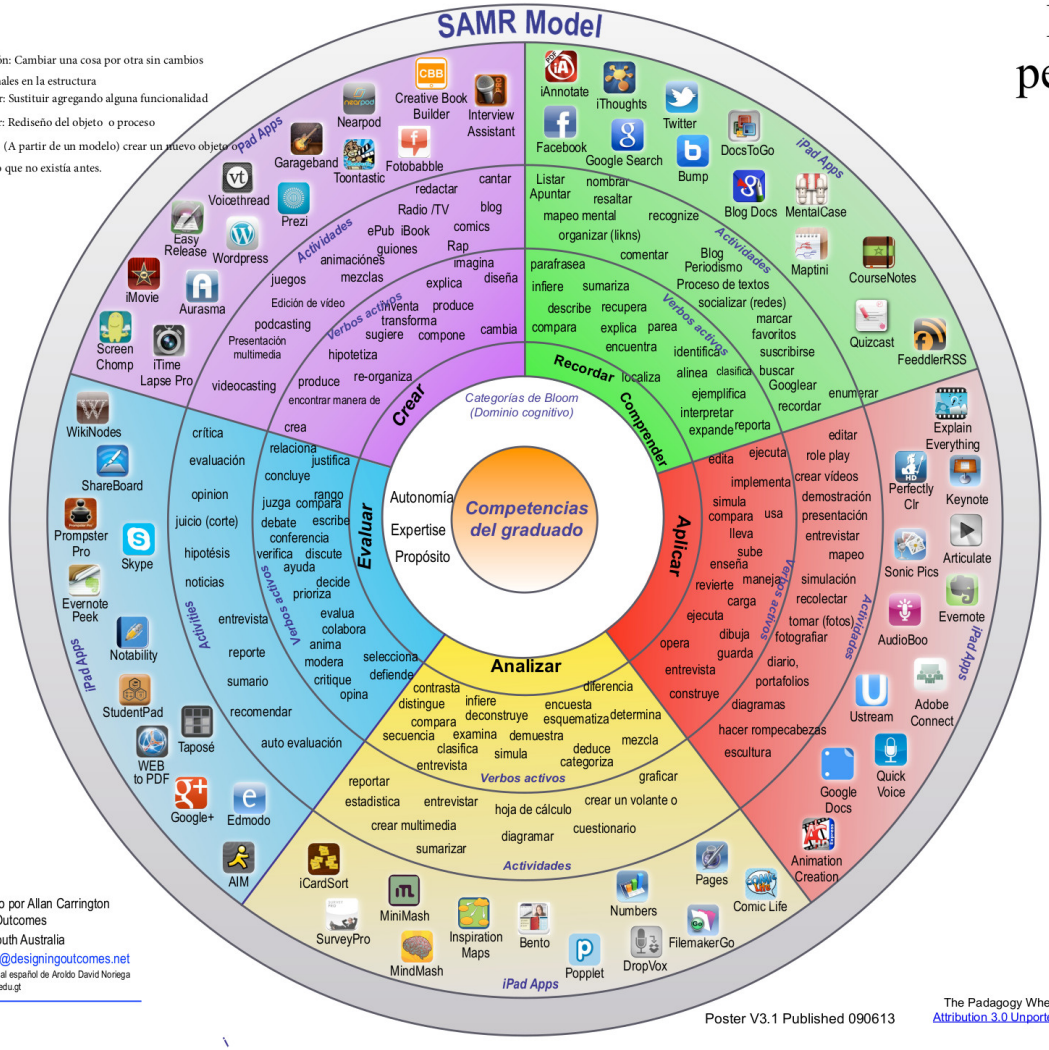
Enseñar Humanidades digitales

- Módulo de investigación digital
  - Métodos de investigación
  - Creación de corpus
  - Técnicas de visualización y estadística
  - Identidad digital y reglamento
  - Conocer modos diferentes de difusión
  - Derechos y obligaciones (combatir la anomia digital)
- Módulo de competencias digitales en la escuela
  - Sociedad y tecnología: riesgos y oportunidades
  - e-learning en contextos diferentes (escuela, empresa, universidad...)
  - Competencias digitales en Educación
  - Enseñanza del lenguaje

Productos de las HDs

- ¿Vale todo lo digital? ¿Qué es una Edición Digital Académica?
  - Un proyecto con objetivos claros
  - Documentación clara de lo que se ha hecho
  - Qué se puede mencionar, re-usar o mejorar (si es que no está finalizado)
  - Qué haga una aportación a su área de trabajo

La rueda de la pedagogía V3.0



Competencias del graduado (a)

Solicitadas por la gente que contrata gente y que decide que quieren de la educación superior.

- Poseer energia, pasión y entusiasmo
- Estad dispuesto a dar el crédito a los demás
- Empatía y trabajo productivo con los demás
- Ser honesto y transparente al negociar con otras personas
- Pensamiento lateral y creativo
- Ser sincero consigo mismo y sus valores
- Escuchar diferentes puntos de vista antes de tomar una decisión.
- Comprender las fortalezas y debilidades de cada uno
- Habilidades para el manejo del tiempo
- Perserverancia
- Aprender de sus errores
- Aprender de la experiencia (propia y de los demás)
- Mantener la calma bajo presión
- Ser capaz de hacer presentaciones efectivas a diferente clase de gentes.
- Identificar los medios de información masiva como oportunidad.

Estas son algunas de las competencias que deben ser identificadas como parte de los atributos de un graduado y un mandato en nuestro diseño de cursos.

Necesitamos transformar el fundamento de lo que hacemos como maestros, se trata de los estudiantes. No caigamos en la trampa de diseñar competencias, indicadores de logro, lecciones, y selección de tecnología sin antes pensar en las capacidades que deseamos de un graduado. Si no sabes que quieres que hagan tus estudiantes al terminar el curso, estas perdido.

Sobre los hombros de gigantes

Esta rueda de la Taxonomía, sin las aplicaciones, fue descubierta en el sitio web del consultor educativo Paul Hopkin's mmiweb.org.uk Esa rueda fue producida por Sharon Artley y es una adaptación de la Revisión de la Taxonomía de Bloom (1956) por Kathwohl y Anderson (2001). La idea de adaptarla para equipos móbiles, particularmente iPad debe agradecerse a Kathy Schrock en su website Bloomin' Apps

Desarrollado por Allan Carrington
Designing Outcomes
Adelaide South Australia
Email: allan@designingoutcomes.net
Traducción libre al español de Aroldo David Noriega
direccion@isea.edu.gt

y ¿las tecnologías del lenguaje?, para...

- Crear y gestionar contenidos, seguimiento de adquisición/errores
- Pero a qué se debe que apenas se usen?
  - Falta de estudios basados en datos
  - Marco teórico
  - Poca colaboración inter-disciplinar: pedagogía, lingüística, lingüística computacional
  - Necesidad de transferencia y adaptabilidad a distintas situaciones (*minimal computing*)

Area		Uso
Educación y tecnología	>>	habitual y desarrollado
Tecnología y enseñanza del lenguaje	>>	escaso y menos desarrollado
Tecnologías del lenguaje	>>	casi nulo y poco desarrollado

Uso de las tecnologías

- Uso de tecnología en el día a día
  - Búsquedas en la red, correctores ortográficos, traductores automáticos, mandos de voz...
  - Diccionarios: Diccionario Básico Escolar [Alegria et al., 2006]
- Otras herramientas o recursos interesantes
  - Corpora:
  - Herramientas para explorar un corpus propio: ANALHITZA (analizador léxico general con PoS), Voyant Tools (basado en caracteres)
  - Tareas automáticas

PLN y las habilidades del lenguaje

	Analizar y interacutar	Usar y consultar
Habla	PRAAT: analizador fonético Webex	
Comprensión		TTS
Escritura	correctores ortográficos, ANALHITZA, Voyant Tools, MARKIN, Drive, Wiki	Corpus Ameresco y CORDE + CREA
Lectura	TSS	e-library
Otras aplicaciones	Lextutor, Compress-es, Python y el paquete NLTK, Hot Potatoes	Modela: traductor automático, CRITERION, Seneko...

Procesamiento del lenguaje natural

- 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

Métodos novedosos

- Métodos intuitivos o de autoridad
  - Demostración con ejemplos complejos
- *Corpus Based Analysis* hipótesis y teóricas se validan con la información del corpus
  - Se observa cuantas veces sucede la hipótesis
  - Se pueden re-hacer los experimentos, mejorar, comprobar
- *Corpus Driven Analysis* no se construyen teorías o hipótesis, se observan los datos y se interpretan

Tareas PNL en investigación

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

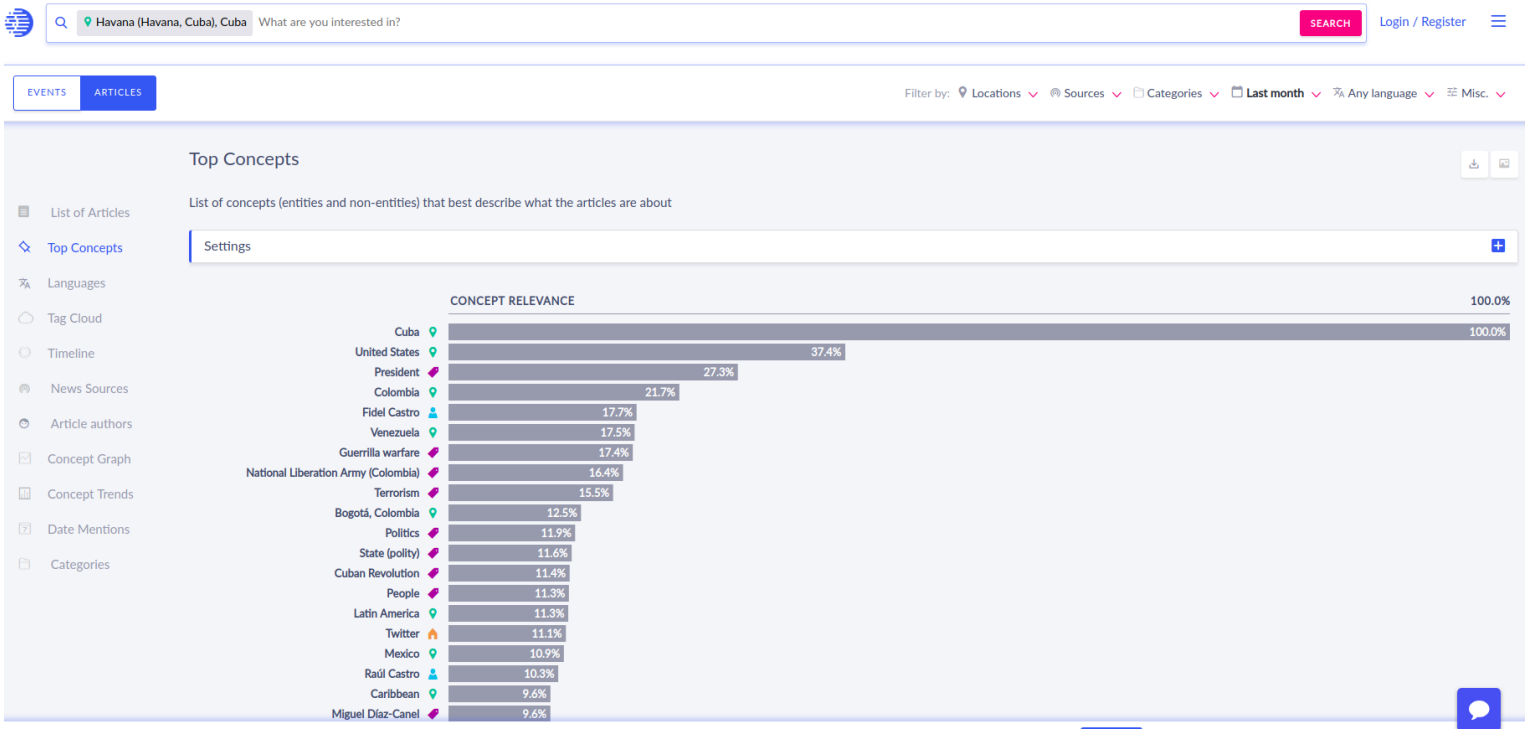
Para realizar tareas que se repiten:

Si x $->$ entonces y

Más ejemplos en: http://nlpprogress.com

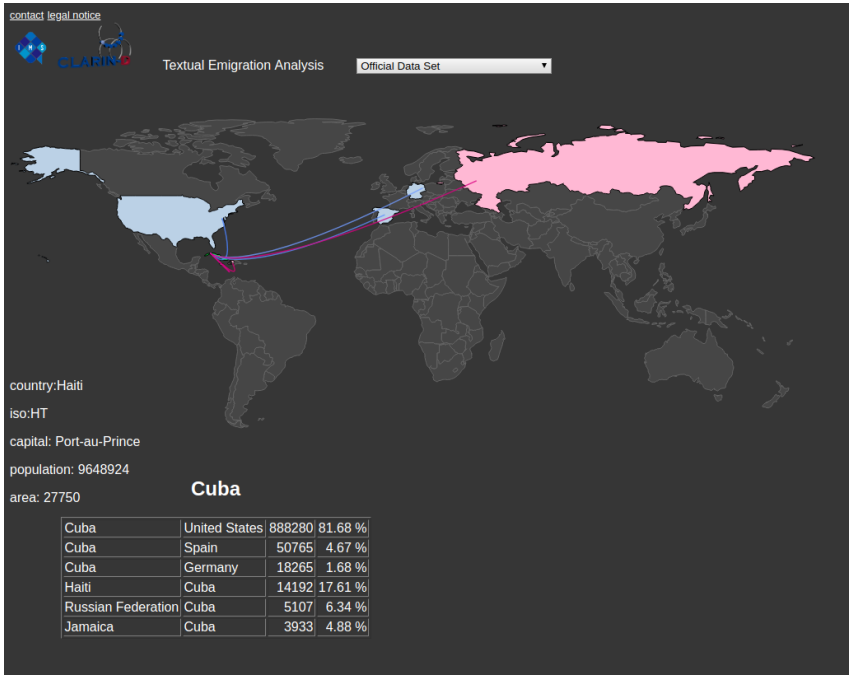
Monitor de noticias y eventos

- En esta web se monitorizan las noticias, hemos buscado los conceptos relevantes que aparecen con “La Habana”
  - http://eventregistry.org



Texto y migración

- Análisis de la migración a partir de la información textual
  - http://clarin01.ims.uni-stuttgart.de/geovis/showcase.html



Web para encontrar y comparar herramientas

- En la web de DiRT (Directory about Research Tools) podemos encontrar herramientas para el uso en la docencia e investigación
- http://dirtdirectory.org/

Analyze data	Interpret data	Annotate	Model data
Archive data	Analyze networks between my data	Capture information	Organize data
Clean up data	Preserve data	Collaborate	Program
Comment	Publish	Communicate	Record audio/video
Analyze the content of my data	Analyze relationships between pieces of data	Contextualize data	Share
Convert files	Analyze the geographical aspect of my data	Create	Store data
Crowdsource data enrichment/analysis	Analyze the structure of my data	Design	Analyze the stylistics of my data
Find information	Theorize	Disseminate data	Transcribe audio, video or manuscripts
Add markup to an object	Translate	Enrich metadata about an object	Visualize data
Collect information	Build a website	Add identifiers to data	Write

Otra web de interés puede ser https://libguides.mit.edu/c.php?g=176357&p=1158575

DiRT Digital Research Tools

Tareas interesantes en educación

- De voz a texto (transcribir) y de texto a voz
- Analizadores básicos: morfológicos, semánticos, sintácticos, discursivos
- Visualizar datos: ANALHITZA, Voyant
- Chatbots
- Creación de ejercicios automáticos
- Uso de corpus reales enriquecidos con técnicas de PLN
- Voz, texto, y multimodal
- Bases de datos, comparaciones automáticas y herramientas de seguimiento

¿Donde guardar todo, para difundir, enlazar y reutilizarlo?

¿En bibliotecas digitales? [Da Sylva, 2013]

- Fuente interesante para PLN y estudios de HD
- Procesar para hacer búsquedas dentro del documento
- Analizar el discurso para resumir
- Análisis de meta-datos
- Datos enlazados (de diferentes formas, para que tengan más interés)
- Clasificaciones de tópicos

Las bibliotecas son infraestructuras para difundir conocimiento

CLARIN

- Common Language Resources and Technology Infrastructure: CLARIN [Váradi et al., 2008])
  https://www.clarin.eu/
- Infraestructura de investigación europea (ERIC) para las Humanidades y Ciencias Sociales que ofrece todos sus servicios sobre tecnología y recursos lingüísticos en una única página
- Es una red organizada administrativamente donde los grupos de investigación promueven, mejoran y comparten el conocimiento, los recursos y la tecnología

PARTHENOS es un proyecto de CLARIN y DARIAH para conservar el patrimonio

¿Qué ofrece CLARIN?

- Datos lingüísticos digitales textos escritos u orales, multimodales
  - Herramientas para describir, analizar y comparar textos: **CLARIN language resource switchboard**: http://weblicht.sfs.uni-tuebingen.de/clrs/#
  - Ejemplos divulgativos sobre proyectos interesantes

¿Quién y cómo se realiza ese servicio en CLARIN?
Los Centros de conocimiento o Centros-K

Centros-K de conocimiento

Centros-K	**Tipo**
Spanish CLARIN K-Centre	Language centre: Spanish, Basque, Catalan and Galician
CLARIN K-Centre for Treebanking	Treebanks
Phonogrammarchiv CLARIN K-Centre	Audio-visual fieldwork
CLARIN K-Centre for Speech Analysis	Speech analysis
CLARIN K-Centre DANSK	Language centre: Danish
The CLARIN K-Centre for Language Learning	Language learning and language disabilities
CLARIN K-Centre for Languages of Sweden	Language centre: Swedish language, minority languages
CLARIN K-Centre of Lund University Hum. Lab	Multimodal and sensor-based methods
IMPACT	OCR and document digitalization

Spanish CLARIN K-Centre es un centro de conocimiento que promueve recursos y tecnología básica de las siguientes lenguas: castellano, catalán, gallego, euskera e inglés

Spanish CLARIN-K

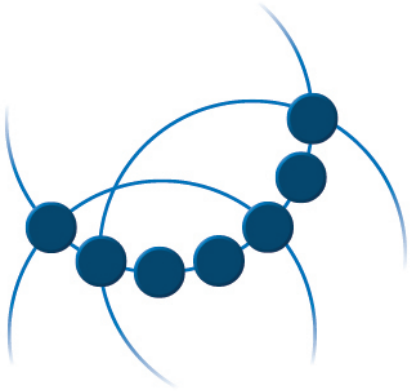
- Spanish CLARIN-K Centre [Bel et al., 2016]
  http://clarin-es.org/
- Grupos que formamos el centro-K
  - IULA-UPF
  - IXA Group
  - LINDH-UNED
  - TALG
- Página web: servicios, noticias y contacto
  http://clarin-es.org/

Objetivo: incrementar y mejorar el uso de PLN en las sociedad, escuelas, grupos de investigación, administración y empresas

¿Cómo podemos ayudar o colaborar?

- Desde Spanish CLARIN-K Centre podemos ayudar si existe
  - Una tarea bien definida
  - Un corpus (procesable)

CLARIN K CENTRE



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.

- Corpus, anotación y desarrollo de herramientas [Ruiz, 1999]
  - Corpus Escolar Cubano: 700000 tokens, anotado manualmente
  - ETICGRAC parser gramatical
  - LEXICON: casi 30000 lemas (base de datos SEPARADA), TERMINA desambiguador de desconocidos (NUEV-LEX)

Búsqueda de ejemplos en corpus de referencia

- Duda "rabo" o "rabos de nube: CREA "rabo∗ de"
- Consultas complejas con su contexto: palabras, lemas, autor, año, país, tema, género textual

Nº	CONCORDANCIA	AÑO	AUTOR	TÍTULO
1	alcanzan vientos poderosos. En Cuba se les llama "rabos de nube". Un tornado de éstos se llevó varias	1985	Valladares, Armando	Contra toda esperanza
2	y crecidas de mayo, sequías perniciosas y varios rabos de nube, y en la que no se habían hecho arreglo	1987	Montero, Mayra	La trenza de la hermosa luna
3	adas de madera, por allá uno de los Bobos, con un rabo de zorra atrapado entre las nalgas, pasaba un	1983	Otero, Lisandro	Temporada de ángeles
4	versitaria frívola que viene al rancho detrás del rabo de su novio. Esa expresión no es mía, la utilizó	1995	Montero, Mayra	Tú, la oscuridad
6	una reunión para otra, discuten aún mordiendo el rabo de la anterior discusión, creen en el mal con	1977	Lezama Lima, José	Oppiano Licario
7	minables. Tres minutos: la eternidad. Al pasar el rabo de nube, el cuartel había sido borrado del mapa.	1992	Alberto, Eliseo	La eternidad por fin comienza un lunes
8	al asta de la bandera, y al acordarse del bendito rabo de nubes, tuvo pena del coronel Justo Mendizábal	1992	Alberto, Eliseo	La eternidad por fin comienza un lunes

Con PLN se anota corpus (PoS), no siempre se puede manualmente

Consultar dudas, en corpus de referencia

- Búsqueda en CORDE (textos cubanos) “rabo”

Pantalla: 1 de 2. Siguiente 1 2 Ver párrafos

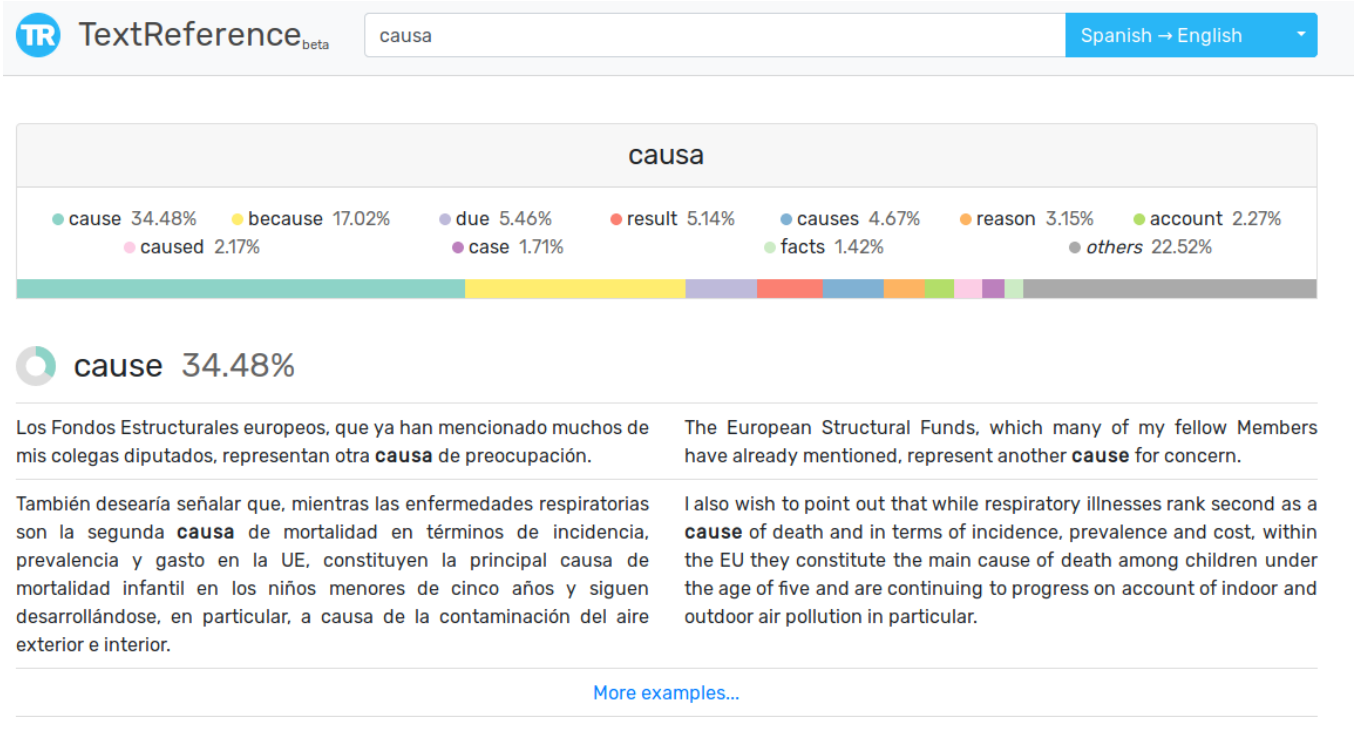
Nº	CONCORDANCIA	AÑO	AUTOR	TÍTULO
1	amos, se decían cosas bajito y nos miraban con el rabo del ojo y todo. Durante toda esa semana Mami me **	1964 - 1967	Cabrera Infante, Guillermo	Tres tristes tigres
2	a Irenita porque era realmente rabirrubia, con su rabo de mula su cola de caballo su moño suelto-amarra **	1964 - 1967	Cabrera Infante, Guillermo	Tres tristes tigres
3	echaba humo por el hocico y por el lomo y por el rabo que era como una antorcha. Ahora había una casa **	1964 - 1967	Cabrera Infante, Guillermo	Tres tristes tigres
4	en las patas, estaba achicharrado y le faltaba el rabo y le faltaban las orejas que debían estar hechas **	1964 - 1967	Cabrera Infante, Guillermo	Tres tristes tigres
5	Qué te pasa? -le pregunté mientras se alargaba el rabo del ojo con un lápiz negro. - A mí nada. Fui y l **	1964 - 1967	Cabrera Infante, Guillermo	Tres tristes tigres
6	golpeando con el garrote de papel que empuñaba el rabo del antílope, que pacía por entre los reflejos d **	1966	Lezama Lima, José	Paradiso
7	las frondosidades de la marquetería, y soltar el rabo del antílope, que se perdía saltando por las roc **	1966	Lezama Lima, José	Paradiso
8	obstante la frase caminando como un ciempiés, con rabo de cabeza de serpiente, y cabeza con entrantes y **	1966	Lezama Lima, José	Paradiso
9	surcos dejados sobre su rostro, los latigazos del rabo del chivo negro que acompaña al diablo. La prime **	1966	Lezama Lima, José	Paradiso
10	cabeza, pero ya de nuevo empieza a madurar por el rabo. Aparecieron después las plantas que necesitan d **	1966	Lezama Lima, José	Paradiso
11	mbién para soltar las amarras y crecer de cirro a rabo cometa. El manajú, servicial príncipe de su rare **	1966	Lezama Lima, José	Paradiso
12	a risotada amistosa. La risotada terminaría en un rabo encintado. Los grupos estudiantiles que se había **	1966	Lezama Lima, José	Paradiso
13	piernas y golpeando en sus cuerpos con su enorme rabo fálico. La carcajada de esos enanos tenía una an **	1966	Lezama Lima, José	Paradiso
14	sus palabras parece un fantasmón, ya no está, es rabo de nube. -En casi todo has acertado -le respondi **	1966	Lezama Lima, José	Paradiso
15	de de vez en cuando un guayabito se llevaba en su rabo el poco fuego sagrado que allí había. Y ya llega **	1966	Lezama Lima, José	Paradiso
16	do la luz sonreído, maliciando con sus ojillos de rabo de nuez el faisán reposado en el extremo del ped **	1966	Lezama Lima, José	Paradiso
17	caballito de bronce en el centro de la isleta, el rabo era de color escarlata y toda la crin del pescue **	1966	Lezama Lima, José	Paradiso
18	a hibernación subterránea. El topo clavado por el rabo, el conejo dominical, el gato moviendo sus bigot **	1966	Lezama Lima, José	Paradiso
19	cabeza, pero ya de nuevo empieza a madurar por el rabo. Seguía su caminata en la medianoche y oyó de pr **	1966	Lezama Lima, José	Paradiso
20	del espejo, lo abandonaba, ya muy mareado, con el rabo enroscado al cuello. Iba saliendo de la duermeve **	1966	Lezama Lima, José	Paradiso
21	da la cuerda, quedándose en el aire, prendida del rabo, como se mordía las uñas, frunciendo el entrecej **	1903	Bobadilla, Emilio	A fuego lento
22	dos en parte a las mejillas terrosas, parecían un rabo de zorra. - Vive le Sha! -gritaron algunos, y el **	1903	Bobadilla, Emilio	A fuego lento
23	a un sendero de marga ocre, pespunteado de malva, rabo de gato, escoba amarga, con sus compactas conste **	1938	Serpa, Enrique	Contrabando. Novela
24	fritos del romerillo en flor. Una iguana, con el rabo ensortijado, corrió ante nosotros. Se oyó el zur **	1938	Serpa, Enrique	Contrabando. Novela
25	en la carnicería. Un pedazo grande. Corría con el rabo entre las piernas. Pero los hombres no son perro **	1938	Serpa, Enrique	Contrabando. Novela

Ir arriba Pantalla: 1 de 2. Siguiente 1 2 Ver párrafos

Nueva consulta: CREA CORDE Nómina de autores y obras Ayuda.

Consulta en TextReference

- Motor de búsqueda en corpus paralelos con palabras alineadas y ejemplos con contexto
- Variedad de direcciones, para el aprendizaje automático en traducción



TTS: Herramientas de texto a voz

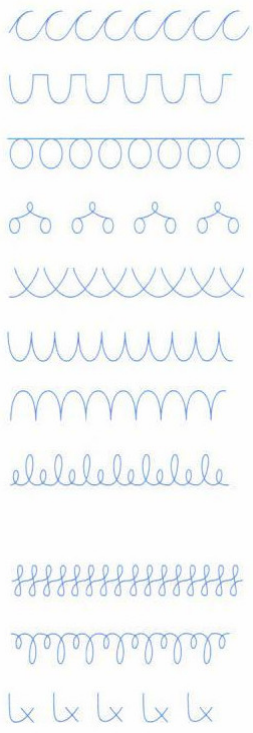
- Aplicaciones para ayudar necesidades diferentes: en diccionarios, periódicos, revistas científicas
- Diccionario Elhuyar: lee las palabras (voz humana o automática)
- Revista www.zientzia.eus: lee el texto

Se puede acompañar la lectura con la voz automática, se puede procesar la voz propia, para oir textos con nuestra voz, se puede escuchar de forma interactiva: con más rapidez o lentitud, elección del sentimiento...

Desarrollo de material basado en datos

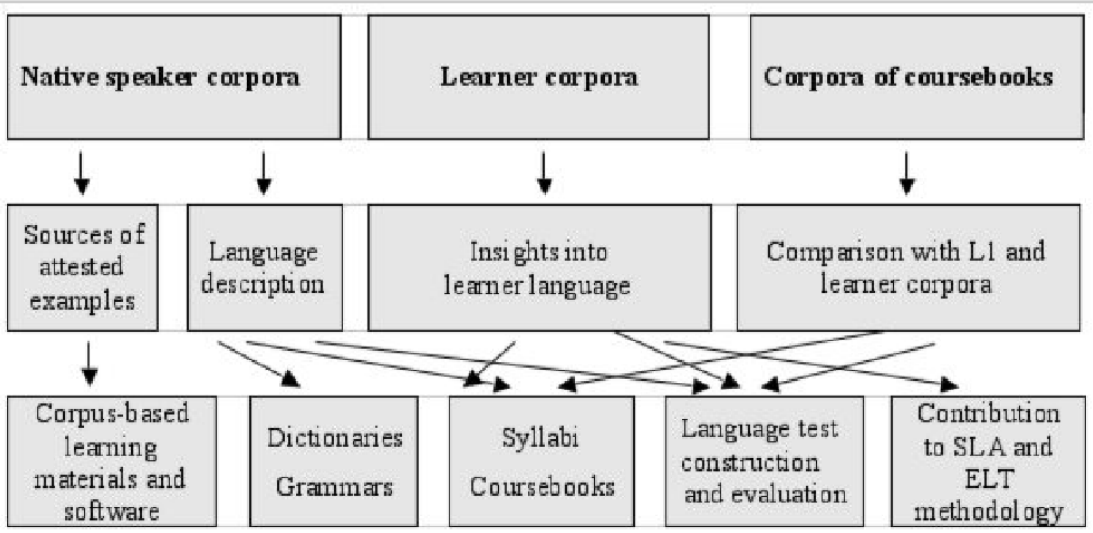
- Cuales son las silabas más frecuentes en los textos infantiles?
- Podemos computar y proponer un sistema de aprendizaje basado en frecuencias y el desarrollo de la grafomotricidad

Letras y trazos
a, d, g, o, q para unir con las anteriores
b, v, w para unir con las posteriores
o para unir con b, f, h, k, l, t
x para unir con posteriores
otras uniones
o unir con m, n, p
e vs l
f
g, j, y
x con la anterior



EPEC	frec.	Cuentos	frec.
bai	3190	bai	1793
bat	1811	bat	1573
ber	1581	gin	1547
bes	1432	han	1281
den	1530	har	1108
gin	2747	hau	752
har	2575	kan	888
kin	1975	kin	1073
kon	1498	men	816
men	1399	nak	1001
nak	1503	ren	2121
ren	6845	rra	1401
rik	1613	rre	1787
rra	2292	rri	3161
rre	4396	san	1405
...	...	...	...

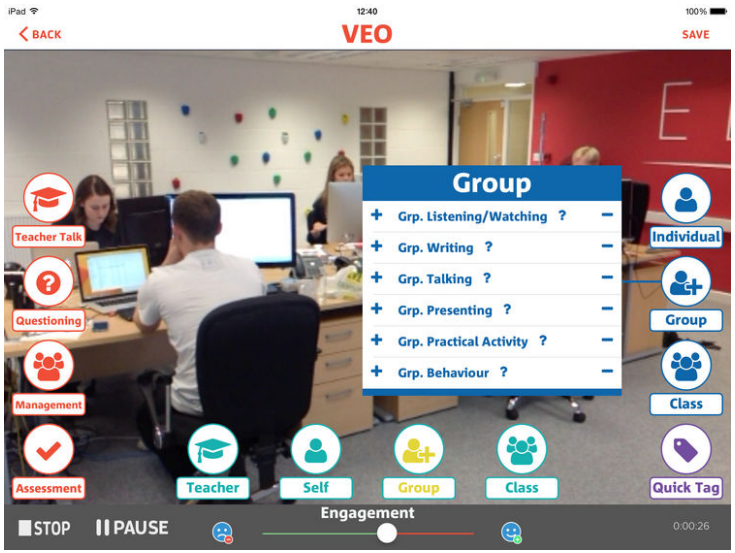
La importancia del aprendizaje con textos reales y de nivel



- El material real es de suma importancia para entender la variedad de la lengua y su importancia en la comunicación
- ¿Cual sería el input más apropiado para el estudio de L2 para el estudiante?
  - ¿Hablante nativo (o de otra generación) o hablante competente de su mismo nivel? ¿O se deberían combinar ambas cuestiones? ¿Cómo?

Intervención multimodal para la mejora del docente

- VEO: herramienta para anotar con etiquetas la dinámica de la clase
- Muy útil para la evaluación entre iguales y desarrollar intervenciones de mejora basado en medios audiovisuales [Rodríguez, 2016]



¿Cómo se podría incluir el PLN en esta situación?

Aprendizaje basado en información de corpus

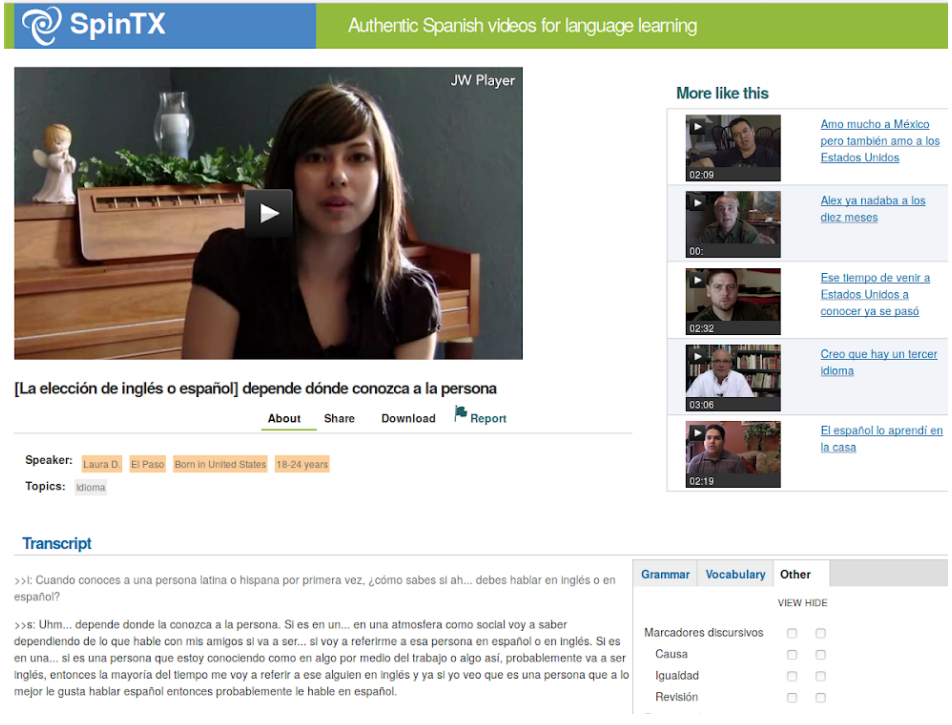
- Cambios desde la enseñanza/aprendizaje basado en información de corpus
  - Desarrollo de material: diccionarios basados en corpus, ejercicios basados en textos reales
  - Cambios en la metodología (acercamiento lexical) y trabajos de comparación
  - Comparación de corpus: hablantes profesionales/nativo VS hablantes del mismo nivel
  - Uso de corpus por parte del alumnado y aprendizaje creando hipótesis
- Lectura práctica: http://www.tesl-ej.org/ej32/a1.html

Uso de material real en clase

- Búsquedas rápidas de material en entrevistas reales
- La búsqueda de material real en un corpus grande, es agotador y hace falta mucho tiempo

Pero y ¿qué tipos de anotaciones pueden ayudarnos a buscar material?
Ejemplo de combinación de automáticos y manuales

Fuente: http://www.coerll.utexas.edu/spintx/video/349



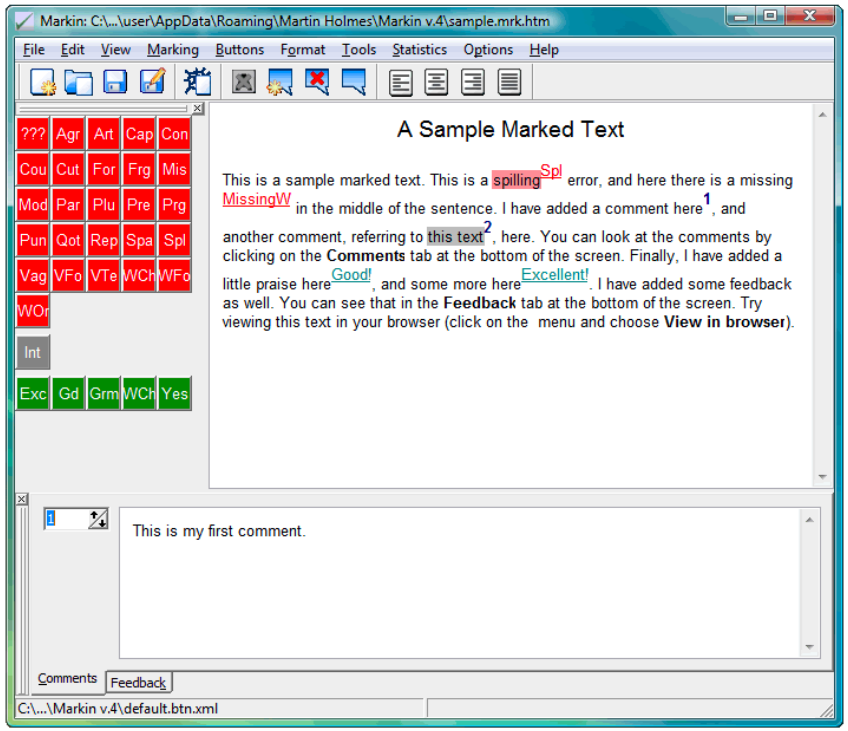
Entornos de corrección semi-automática

- Corrección automática
  - Corrección automática
  - Errores ortográficos, gramaticales (pero ¿cuáles?)
  - Contar palabras, subrayar formas complejas/simples
  - Marcar oraciones extra-largas o formas repetidas
- Describir la adquisición de conocimiento de un estudiante
- Trabajar con las estadísticas de clase y detectar los errores/carencias principales

Corrección de textos digital: MARKIN

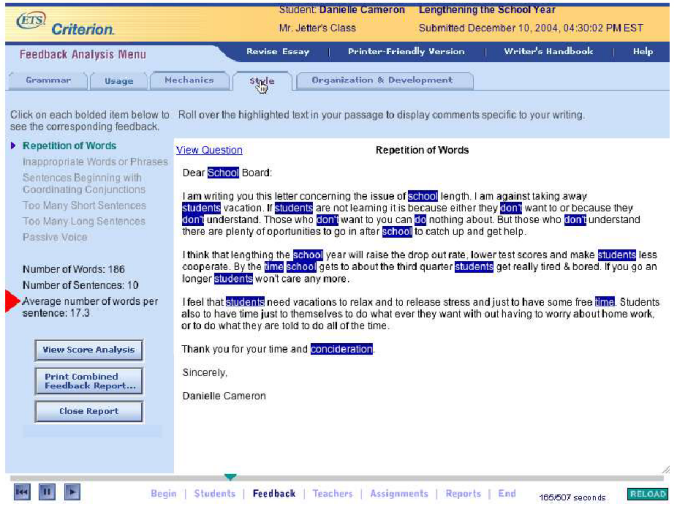
- Tras corregir los textos manualmente
- se pueden contar los errores de la clase y trabajar sobre ellos
- ¿Cómo?, de una forma consciente y colectiva

¿Puede ser un corpus con correcciones una fuente para el desarrollo de correctores automáticos?



Corrección automática parcial

- El PNL se puede usar para realizar correcciones automáticas en entornos virtuales de aprendizaje:
  - Para el docente: Contar palabras, resaltar estructuras repetidas o complejas: CRITERION
  - Para el discente: Subrayar errores, resaltar repeticiones, proponer sinónimos o ejemplos de uso: ARTEXT



Futuro proyecto: digitalizar la escuela y incrustar PLN en colaboración para crear un corrector parcial

Herramientas básicas para su uso amplio

- **ANALHITZA** herramienta de uso fácil
  - resultado en hoja de cálculo, para su posible modificación
  - Desarrollado por el Grupo Ixa para promover el uso de las de herramientas basadas en PLN
  - Basado en información morfosintáctica: procesa castellano, euskera e inglés
  - Minimal computing: no necesita internet y se puede usar con diferente sistemas operativos

Seguramente necesita adaptación para un uso más óptimo con texto cubanos

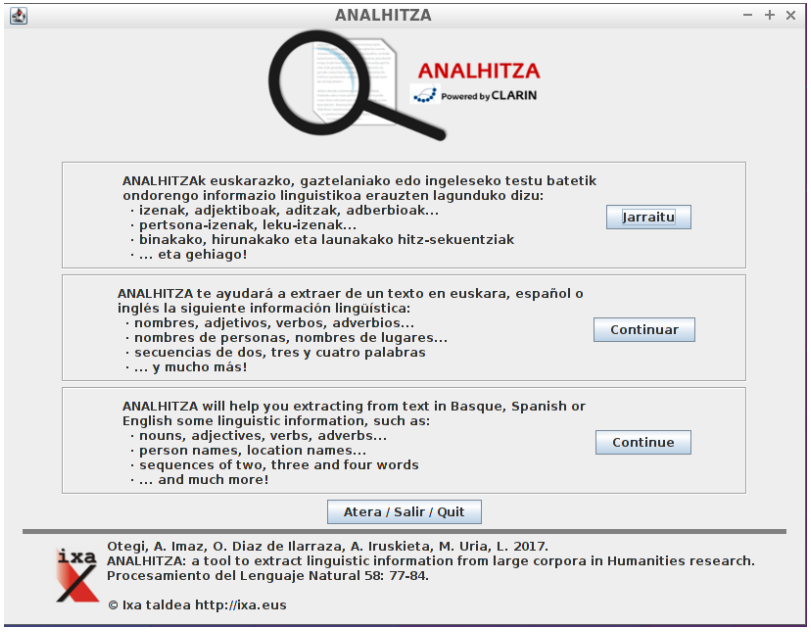
- **VOYANT TOOLS** es una herramienta con una visualización de datos interesante
  - No incorpora información morfológica: procesa muchas lenguas
  - De fácil uso, gratuito y código abierto

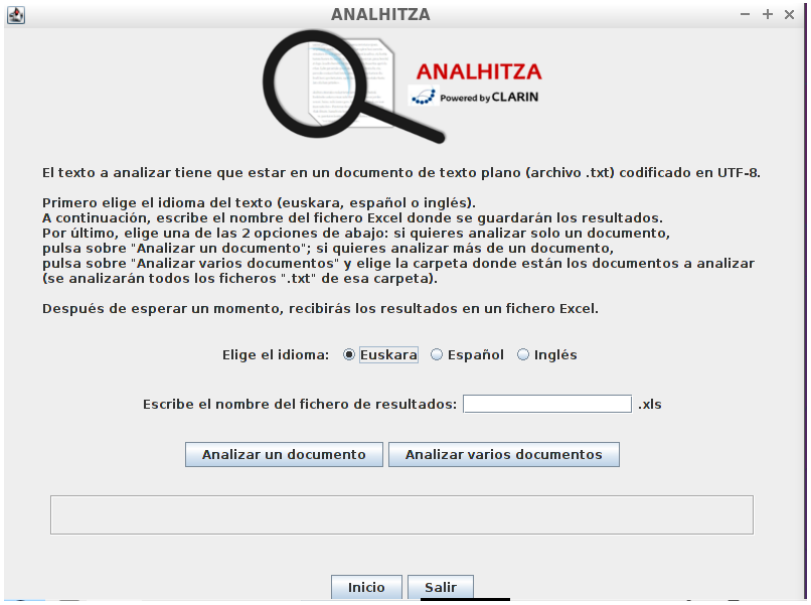
ANALHITZA v2.0.0 (*minimal computing*)

- Versión para este congreso en USB e internet
- Multiplataforma y con versiones diferentes:
  - Para Linux versión de: *a*) docker y *b*) virtual
  - Para Windows sólo la versión de la máquina virtual
- En las carpetas "docker" y "máquina_virtual" más instrucciones para cada versión

En caso de uso, por favor, citar:
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: 77-84.

ANALHITZA v2.0.0 (*minimal computing*)





Corpora utilizado

- La revista PIONERO: www.pionero.cu
- "La Edad de Oro" y "Versos Libres" de José Martí
- Cuentos de la revista "Zunzun" http://www.zunzun.cu/
- Artículos del periódico "Juventud Rebelde"
- Pulgarcito. Libro digitalizado, sin OCR

1.Propuesta combinada utilizando estas herramienta

Empezamos por lo sencillo: un texto y 2 herramientas de PLN

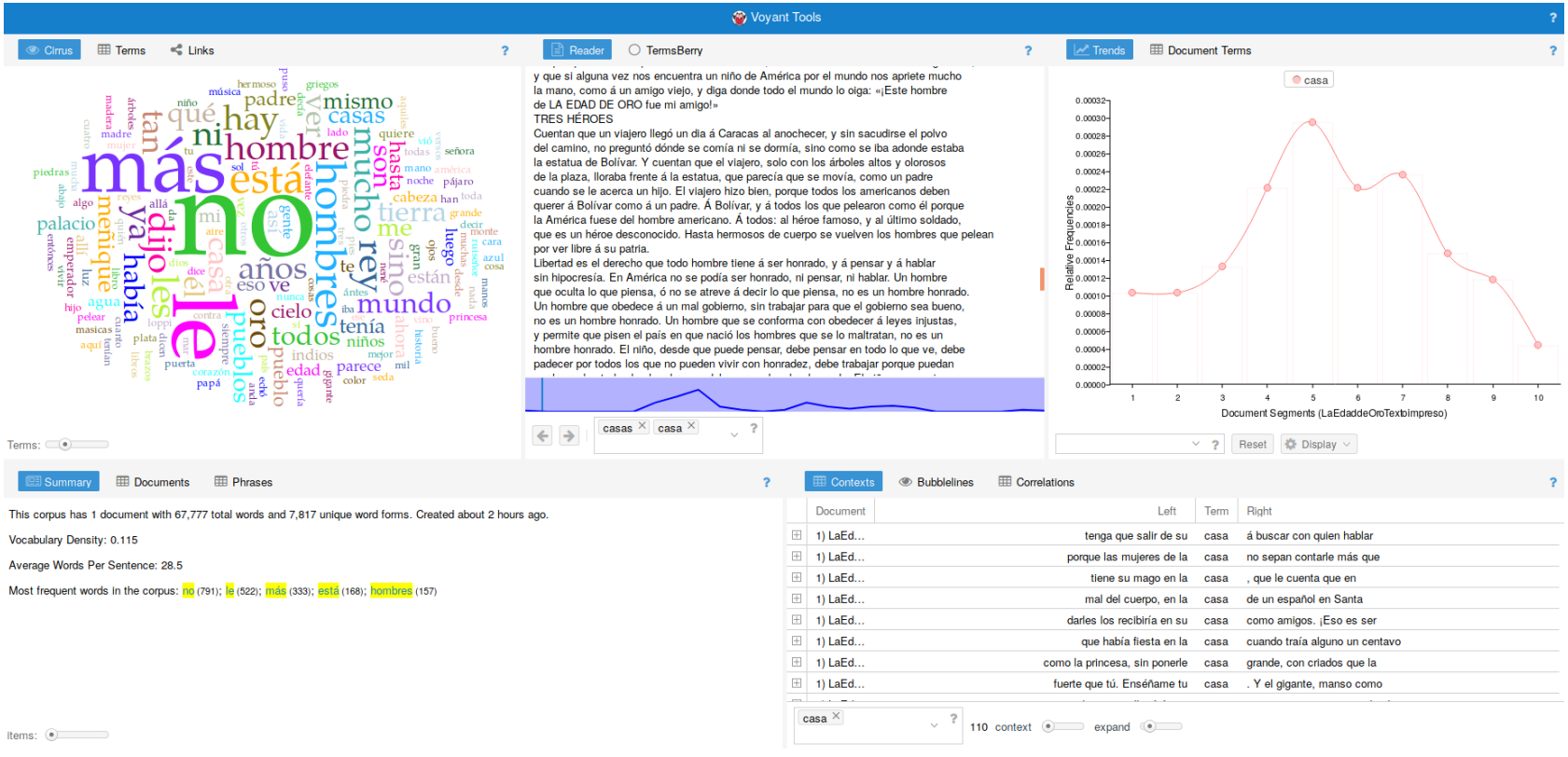
1.*a*. Extraer lemas con ANALHITZA
2.*b*. Leer lista de palabras para buscar desconocidas
3.*c*. Buscar con Voyant Tools el contexto de las palabras desconocidas
4.*d*. Buscar palabras desconocidas en el Diccionario Escolar Cubano

1.*a*. Extraer lemas con ANALHITZA de un libro digital

Martí: Versos					
nombres		Adjetivos		Verbos	
50	vida	23	triste	146	ser
48	hombre	22	nuevo	59	haber
45	alma	17	blanco	37	ver
45	sol	16	muerto	31	ir
43	amor	15	oscuro	27	dar
42	flor	14	vivo	25	vivir
42	tierra	14	negro	23	tener
41	luz	13	roto	21	mirar
38	mar	13	seco	20	decir
37	ojo	13	puro	19	querer
36	mano	12	duro	17	venir
35	pecho	12	pálido	16	morir
33	aire	12	alto	15	saber
30	verso	10	rico	14	echar
29	cielo	10	noble	14	parecer
	...		...		...

1.c.Observar el uso de una palabra

- Descargar el Libro de José Martí desde la www.bne.es
- Observar con Voyant tools como el lema "casa" y "casas" (VOYANT) se menciona muy poco al principio ni al final



2. Buscar con Voyant Tools el contexto de las palabras desconocidas

2.*a.* Jose Martí: Versos Libres

2.*b.* Observar el contexto de las palabras "hombre" y "mujer" con Voyant Tools (Key Word in Context: KWIC)

2.*a*. Extraer bigramas con ANALHITZA

- Jose Martí: Versos Libres
- Observar los adjetivos a la derecha de los lemas "hombre" y "mujer"

frec.	izq	cat.	der.	cat.
1	hombre	N	esclavo	G
1	hombre	N	impasible	G
1	hombre	N	duro	G
1	hombre	N	tenaz	G
1	hombre	N	honrado	G
1	hombre	N	moderno	G
1	hombre	N	libre	G
1	hombre	N	victorioso	G
1	hombre	N	inundo	G
1	hombre	N	triste	G

frec.	izq	cat.	der.	cat.
2	mujer	N	amado	G
1	mujer	N	hermoso	G
1	mujer	N	libre	G
1	mujer	N	libidinoso	G

2.*b*. Buscar con Voyant Tools el contexto de las palabras

- Jose Martí: Versos Libres
- Observar el contexto de las palabras "hombre" y "mujer" con Voyant Tools (Key Word in Context: KWIC)

De donde pasa el	hombre	, por quien muero, Guardo,
Conozco el	hombre	, y lo he encontrado malo
la ancha arena, Y los	hombres	esclavos gladiadores.
¡Que usan los	hombres	hoy oro empañado!
el mar camina, Una afable	mujer	se asomó al ciego:
entrañas encerrado! No es que	mujer	me engañe, o que fortuna
de lacayos Quejarse, y de	mujeres	, Y de aprendices de la
blanco en el original. 35	mujeres	I Esta, es rubia;

3. Otra propuesta comparando 2 textos

Complicamos el uso comparando 2 textos y profundizando en las herramientas de PLN

3.*a*. Comparar términos disjuntos en 2 textos con Voyant Tools

3.*b*. Extraer n-gramas y usar filtros en hojas de cálculo

3.*a*. Nube de palabras y términos disjuntos

- Nube de palabras de las obras de Martí: "La edad de oro" y "Versos libres"

- Palabras frecuentes y distintas:
  - Versos libres: seno (14), torno (9), cráneo (9), tiranos (7), miro (7).
  - La edad de oro: pueblos (98), meñique (92), están (76), indios (68), allá (59).

3.*b*. Buscar el contexto de las palabras desconocidas

- Jose Martí: Versos Libres
- Observar el contexto de las palabras "seno" (14), "**homagno**" (6), "sino" (6), "**hirsuta**" (5), "zorzal" (6), "pomona" (3)... con Voyant Tools (Key Word in Context: KWIC)

la sombra entrevé, ? ¡no son	homagno	! Mis ojos sólo, los mis
las ardientes Manos del triste	homagno	parecían! 24 Y la tierra
dijo: "Flor de mi seno,	homagno	generoso, De mí y de
en el original. 39 CRIN	hirsuta	¿Que como crin hirsuta de
CRIN HIRSUTA ¿Que como crin	hirsuta	de espantado Caballo que en
Que cual tropel famélico de	hirsutas	Fieras saltan de mí buscando
que quería Mucho a un	zorzal	, a quien dejaba libre Surcar
aquellas cercanías,- Y púsole al	zorzal	el buen labriego Sobre sus
el halcón sólo Prendió al	zorzal	, que diestro se le escurre

No es suficiente con una obra, necesitamos más ejemplos

Realizar búsquedas en Voyant Tools

- Extraer la nube de palabras de los objetivos de la profesora con los textos de Zun Zun. (Voyant)
- Quitar palabras no interesantes de la lista (stop words)
- Enseñar sólo las palabras de una lista predeterminada y navegar por el texto sobre ellas

Fábulas y actividades en Zun Zun

a)	lectura (30), revista (28), zunzún (25), texto (23), digital (20)
b)	día (16), entonces (12), boca (10), yo (8), tan (8)

Adivinen que lista corresponde a las actividades y a las fábulas!!!

Aprendizaje entre iguales

- La coevaluación es posible con la redacción y corrección digital de los textos, en un entorno adecuado.
- Puede haber diferentes correctores
- Se pueden comparar muy fácilmente los marcadores de discurso de dos estudiantes
- Lo interesante es poder comparar el texto con otro estudiante que lo ha hecho con éxito (se puede graduar), si ha escrito el mismo texto y se puede ver cómo lo ha resuelto

Explicación teórica y inducción de errores

- Pedir a los estudiantes que analicen sus propios textos
- Proponer un error que realizan muchos estudiantes

Observamos como los diferentes estudiantes utilizan formas correctas e incorrectas

Disjuntos: palabras exclusivas en cada texto

- Listar las palabras que solo aparecen en cada cuento de la siguiente colección de cuentos de la revista Zun Zun

Palabras de cada texto: fábulas en Zun Zun

1	ternerito (4), amarillo (3), color (3), verde (2), pintó (2)
2	manos (2), tenían (1), supo (1), salvarla (1), resultado (1)
3	paloma (5), hormiga (5), manantial (2), vio (2), vuelo (1)
4	vara (4), boca (7), cocodrilo (5), me (6), dentista (3)
5	zorras (6), campesino (5), burro (5), sogas (4), hacia (3)
6	tiddalick (5), empezó (4), agua (6), rápido (2), mucha (2)
7	otros (3), piel (3), cocodrilo (3), animales (3), poco (2)
8	anansi (6), tigre (8), caballo (3), cosas (2), chicas (2)
9	simón (5), perro (5), olvidó (4), granjero (4), gallina (4)
10	fico (5), padres (2), después (2), amigos (2), ni (2)

¿Más ejercicios?, extraer formas de uso de escolares

- Usando ANALHITZA extraer de 15-25 textos los marcadores del discurso
  - Observar cuál es el más usado
  - Observar cuál de los marcadores explicados por el docente no se han usado en los textos con una hoja de cálculo
- Observar qué formas léxicas interesantes/nuevas han usado los estudiantes y enseñar el contexto y su uso a otros alumnos que no lo han usado

¿Más ejercicios?, comparar textos diferentes

- Comparar el lenguaje utilizado en la revista "Zun Zun" con la de "Juventud Rebelde", para un análisis diacrónico
- Buscar textos comparables
- Conformar corpora de tamaño similar
- Utilizar ANALHITZA y Voyant Tools

Os toca: ¿podemos colaborar en otras comparaciones interesantes? **¿Hace falta alguna otra herramienta para vuestras investigaciones?**

Descarga de documentos

Objetivo: analizar un libro de una biblioteca digital: problemática para observar las dificultades y posibilidades que tiene el PNL

- 1. Descargar un libro digitalizado que contiene textos escritos a mano y a máquina
  - Problemas de conexión para bajar con un explorador de internet (mala conexión?)

```
wget imagenes.sld.cu/download/pulgarcito/volumen-2.pdf
```

Comandos útiles para la creación de corpus

- Pasar todos los PDF de una carpeta a documentos TXT desde la línea de comando:

```
for file in *.pdf; do pdftotext "$file" "$file.txt"; done
```

- Unir diferentes ficheros:

```
cat *.txt > corpus.txt
```

Pasos

- Para transcribir documentos escritos a mano: https://transkribus.eu/Transkribus/
- Para pedir ayuda: centro de competencia IMPACT https://www.digitisation.eu/

¿Cómo digitalizar un libro de imágenes?

- Hemos usado un OCR comercial, pero los resultados han sido muy pobres
- El centro IMPACT nos ha ayudado gratuitamente a usar y configurar OCRs de mayor calidad (9 días)
- Hemos logrado archivos con mayor calidad
  - Todo el corpus en formato TXT (con errores de OCR)
  - Todo el corpus en XML (con errores de OCR)

Ayuda de centro de competencia de CLARIN IMPACT

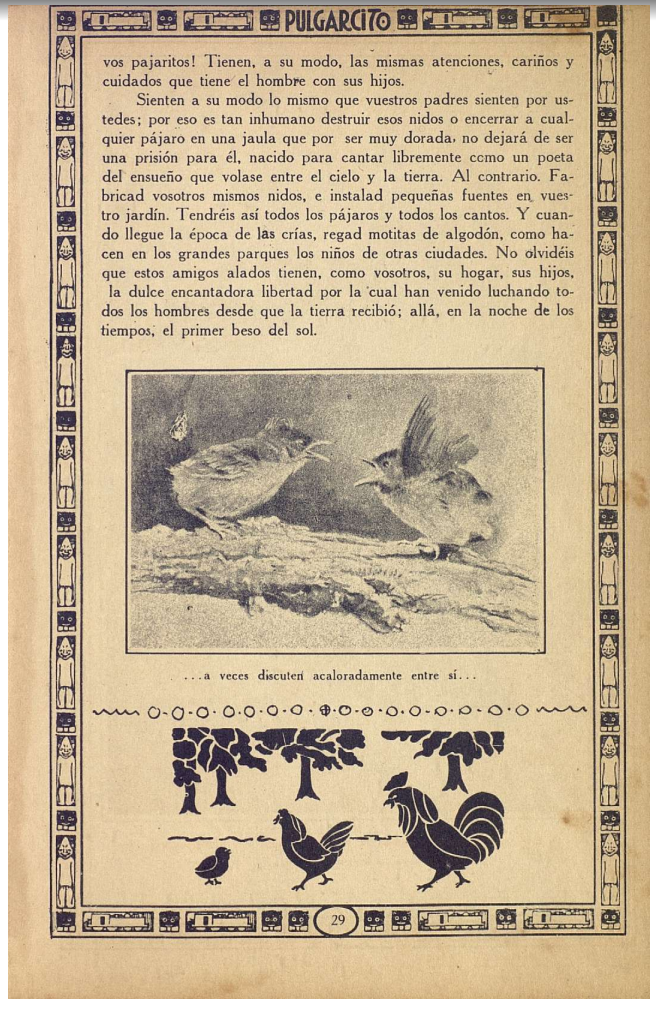
- La digitalización se ha realizado con la colaboración de IMPACT Centre of Competence y se han usado las siguientes herramientas
  - Extracción de imágenes a partir de pdf: comando pdfimages en Linux.
  - Binarización: FineReader 11 SDK versión
  - OCR: FineReader 11 SDK versión con idioma español y tipos de letra de las pruebas: normal y handprinted salida en ALTO XML y Text Unicode Defaults.

Ejemplo de un texto de PDF a TXT (IMPACT)

CUANDO UN N1NO
<<SÍ POEAA?
M$&ECE, UMBETRAfO
conminas y cía

Cursos en Programming Historian sobre OCR:
https://programminghistorian.org/en/lessons/generating-an-ordered-data-set-from-an-OCR-text-file

De PDF a TXT (IMPACT)


PULGARCITO

vos pajaritos! Tienen, a su modo, las mismas atenciones, cariños y cuidados que tiene el hombre con sus hijos.

Sienten a su modo lo mismo que vuestros padres sienten por ustedes; por eso es tan inhumano destruir esos nidos o encerrar a cualquier pájaro en una jaula que por ser muy dorada, no dejará de ser una prisión para él, nacido para cantar libremente ccmo un poeta del ensueño que volase entre el cielo y la tierra. Al contrario. Fabricad vosotros mismos nidos, e instalad pequeñas fuentes en vuestro jardín. Tendréis así todos los pájaros y todos los cantos. Y cuando llegue la época de las crías, regad motitas de algodón, como hacen en los grandes parques los niños de otras ciudades. No olvidéis que estos amigos alados tienen, como vosotros, su hogar, sus hijos, la dulce encantadora libertad por la cual han venido luchando todos los hombres desde que la tierra recibió; allá, en la noche de los tiempos, el primer beso del sol.

...a veces discuten acaloradamente entre sí...

29

NoBin_pul-074.txt	Bin_pul-074.txt
vos pajaritos! Tienen, a su modo, las mismas atenciones, cariños y cuidados que tiene el hombre con sus hijos. Sienten a su modo lo mismo que vuestros padres sienten por us- tedes; por eso es tan inhumano destruir esos nidos o encerrar a cual- quier pájaro en una jaula que por ser muy dorada, no dejará de ser una prisión para él, nacido para cantar libremente ccmo un poeta del ensueño que volase entre el cielo y la tierra. Al contrario. Fa- bricad vosotros mismos nidos, e instalad pequeñas fuentes en vues- tro jardín. Tendréis así todos los pájaros y todos los cantos. Y cuan- do llegue la época de las crías, regad motitas de algodón, como ha- cen en los grandes parques los niños de otras ciudades. No olvidéis que estos amigos alados tienen, como vosotros, su hogar, sus hijos, la dulce encantadora libertad por la cual han venido luchando to- dos los hombres desde que la tierra recibió; allá, en la noche de los tiempos, el primer beso del sol.	vos pajaritos! Tienen, a su modo, las mismas atenciones, cariños y 4 cuidados que tiene el hombre con sus hijos. Sienten a su modo lo mismo que vuestros padres sienten por us- tedes; por eso es tan inhumano destruir esos nidos o encerrar a cual- quier pájaro en una jaula que por ser muy dorada, no dejará de ser una prisión para él, nacido para cantar libremente ccmo un poeta del ensueño que volase entre el cielo y la tierra. Al contrario. Fa- bricad vosotros mismos nidos, e instalad pequeñas fuentes en. vues- tro jardín. Tendréis así todos los pájaros y todos los cantos. Y cuan- do llegue la época de las crías, regad motitas de algodón, como ha- cen en los grandes parques los niños de otras ciudades. No olvidéis que estos amigos alados tienen, como vosotros, su hogar, sus hijos, la dulce encantadora libertad por la ‘cual han venido luchando to- dos los hombres desde que la tierra recibió; allá, en la noche de los tiempos, el primer beso del sol. ...a veces discuten acaloradamente entre sí... O-O-O-O’O’O-O - $-0.0-0.0-0-0 - 0*0

Expresiones regulares: limpiando el ejemplo

Bin_pul-074.txt

vos pajaritos! Tienen, a su modo, las mismas atenciones, cariños y 4 cuidados que tiene el hombre con sus hijos. Sienten a su modo lo mismo que vuestros padres sienten por us- tedes; por eso es tan inhumano destruir esos nidos o encerrar a cual- quier pájaro en una jaula que por ser muy dorada, no dejará de ser una prisión para él, nacido para cantar libremente ccmo un poeta del ensueño que volase entre el cielo y la tierra. Al contrario. Fa- bricad vosotros mismos nidos, e instalad pequeñas fuentes en. vues- tro jardín. Tendréis así todos los pájaros y todos los cantos. Y cuan- do llegue la época de las crías, regad motitas de algodón, como ha- cen en los grandes parques los niños de otras ciudades. No olvidéis que estos amigos alados tienen, como vosotros, su hogar, sus hijos, la dulce encantadora libertad por la 'cual han venido luchando to- dos los hombres desde que la tierra recibió; allá, en la noche de los tiempos, el primer beso del sol. ...a veces discuten acaloradamente entre sí...
O-O-O-O'O'O-O - $-0.0-0.0-0-0 -0*0

- Se puede usar el editor de texto o el programa Geany (multiplataforma) https://www.geany.org/, para limpiar el texto de OCR y con expresiones regulares

	Orig.	Result.	Coment.
1	cuan- do	cuan- do	Mantener original
2	cuan- do	cuando	procesar texto
3	y 4 cuidados	y cuidados	
4	O-O-O-O'O'O-O - $-0.0-0.0-0-0 -0*0		Limpiar texto

Expresiones regulares: limpiando el ejemplo

- También se puede realizar en block de notas con el este ejemplo
  1. No hacer nada
  4. $O-O-O-O'O'O-O\backslash s-\backslash s\backslash\backslash\$-0.0-0.0-0-0\backslash s-0\backslash *0$
  3. $\backslash d\backslash n$
  2. $-\backslash s$

Hay que tener mucho cuidado con la orden, pero se puede probar muchas veces guardando el original e integrarlo en Python para automatizar el proceso

- Más en *Programing Historian*
  You could isolate all the punctuation, spaces, and numbers this way: $[[:punct:],,0-9]$ limpiando OCR, ordenando datos

- Una vez que tenemos el libro en formato TXT (UTF8) hemos utilizado ANALHITZA para extraer el diccionario reconocible, nombres propios (NERC) y algunos n-gramas, entre otras cuestiones

Freq.	Nouns	Freq.	Adjectives
255	niño	160	bueno
194	año	124	gran
159	hombre	99	grande
154	día	75	nuevo
148	padre	62	viejo
148	rey	57	blanco
134	hijo	51	pobre
131	vez	48	mayor
114	libro	48	largo
106	casa	45	azul
103	tiempo	44	mejor
...	...	...	...

Nombres propios

Freq.	w1	Type
8	alemania	LOC
2	dinamarca	LOC
2	alejandro	PER
1	16 de mayo de 1703	DATE
1	cataluña	LOC
...	...	...

Bigramas

Freq.	W1	cat	w2	cat
846	de	P	el	D
692	en	P	el	D
565	a	P	el	D
388	y	C	el	D
245	de	P	su	D
229	por	P	el	D
226	el	D	que	Q
224	todo	D	el	D
206	con	P	el	D
204	a	P	su	D
202	que	C	el	D
201	de	P	uno	D
165	ser	V	el	D
151	el	D	niño	N
...	...	...	...	...

Nube de palabras

Palabras en su contexto

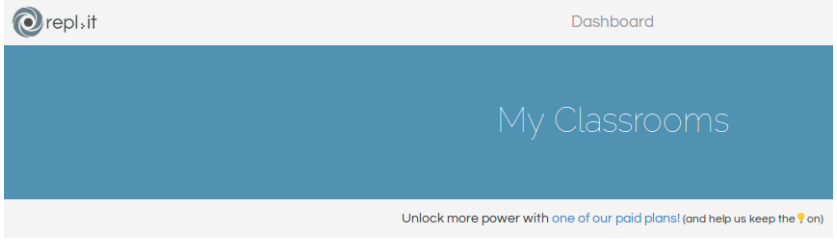
- El KWIC de la palabra "niña" extraído con Voyant Tools

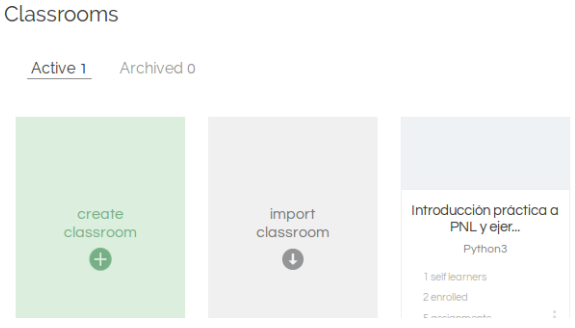
Left	Term	Right
tenía, a su vez, una	niña	, que era dulce y bon
las excelentes cualidades de aquella	niña	. La encomendó las tareas más
pies a cabeza. La pobre	niña	todo lo sufría con paciencia
g n □w- canzaria. La	niña	perdió uno de sus zapatos
meses regalaremos al niño o	niña	que mayor número de ellas
ha pensado mucho en la	niña	! El dice que siempre que
y escribe mejor- Y la	niña	se va, se va adespacio
tropieza con todo! Pero la	niña	no se ha des- pertado
de olor: y es una	niña	de sombrero colorado, que trae
hoy en casa por mi	niña	", le dijo su padre, "y
...	...	...

Se puede navegar por el el documento gracias a los programas de visualización

PLN con Python (en web sin instalar nada) [Iruskieta and Otegi, 2018]

- Entorno de aprendizaje: para empezar de cero o tener feedback (tanto automático y manual)
  - Entorno `https://repl.it/`
  - Recibir feedback o ayuda del docente o ver soluciones o hacer cambios sobre la solución





Instalar Python

- Entorno habitual: para realizar programas más complejos
- Ubuntu 16.04 ya viene con Python 3 y Python 2. Para asegurarse: apt-get
  - $ sudo apt-get update
- Ver si la versión de Python 3 está instalada:
  - $ python3 -V
- Para instalar paquetes, usaremos pip:
  - $ sudo apt-get install -y python3-pip
  - $ sudo apt-get install python3-pip
- Para Windows:
  https://www.python.org/downloads/windows/

Ejemplos con Python para lingüistas

- Imprimir: https://repl.it/@ixaixa/imprimir
- Introducir datos: https://repl.it/@ixaixa/Introducir-datos
- Contar: https://repl.it/@ixaixa/contar
- Iteraciones: https://repl.it/@ixaixa/iteraciones
- Abrir URLs:https://repl.it/@ixaixa/abrir-URL
- Extraer titulares: https://repl.it/@ixaixa/Extraer-titularBeatifulSoup
- Escribir en un fichero https://repl.it/@ixaixa/escribir-en-fichero
- Guardar un titular en XML: https://repl.it/@ixaixa/T6-Insertar-URLsguardar-titular-XML
- Comparar frecuencias de hombres y mujeres: https://repl.it/@ixaixa/T7comparar-frecuencias-hombres-mujeres

Más complejo pero interesante con Python para lingüistas

- **¿En que partido pone más el foco los siguientes periódicos?**
- Comparar las frecuencias de las siguientes palabras: Athletic de Bilbao, Bilbao, Real Sociedad, Real Madrid, Barcelona y Barça.
- Para conseguir estos datos hemos mirado en más de 1000 HTMLs (3/07/2018). Fuentes: DV, Correo, Marca, Sport

	DV	Correo	Marca	Sport
Athletic de Bilbao	0	0	0	0
Barça	0	0	4	7
Barcelona	5	5	2	5
Bilbao	0	27	0	0
Real Madrid	1	1	17	6
Real Sociedad	7	0	0	0

Otro ejemplo interesante con Python para lingüistas

- Extraer n-gramas y sus frecuencias de los siguientes pronombres personales para ver cual es el más habitual: "la" (femenino), "el" (masculino) o "lo" (neutro).
- RESULTADOS parciales de https://labur.eus/z5BSG

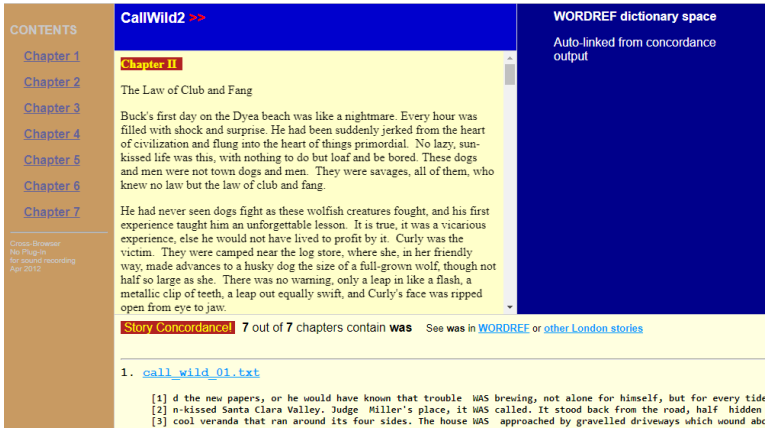
fem.	term.	frec.	masc.	term.	frec.	neu.	term.	frec.
la	persona	22	el	caso	16	lo	concerniente	1
la	suscripción	17	el	pago	5	lo	considere	1
la	web	15	el	momento	4	lo	aceptes	1
La	persona	10	el	primer	4	lo	leas	1
la	empresa	7	el	acceso	3	lo	establecido	1
la	société	7	el	IVA	3	lo	tanto	1
la	campaña	6	el	segundo	3	lo	que	2
la	información	4	el	servicio	3			
la	loi	4	el	formulario	2			
la	manière	4	El	IVA	2			
...	...	...	...	...	....			

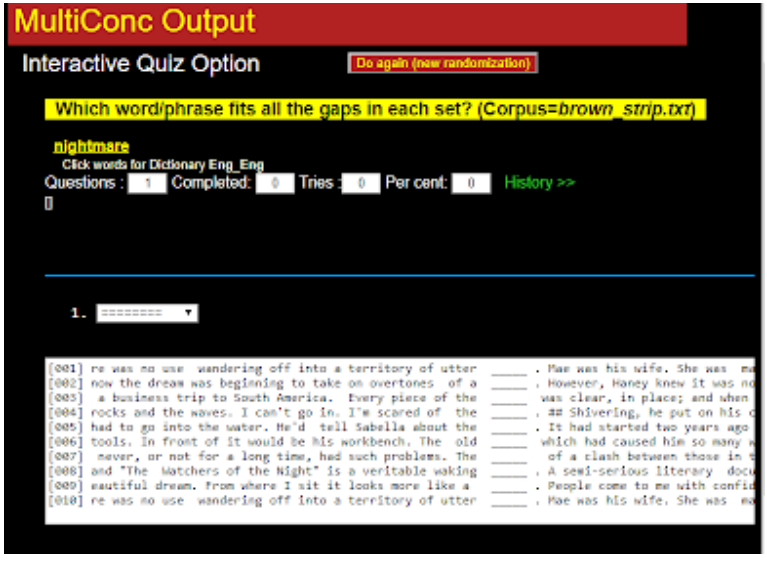
Más Phyton

- Curso de iniciación de PLN y Python para lingüistas
  https://ixa.si.ehu.es/node/11537
- Manual en castellano:
  https://wiki.python.org/moin/SpanishLanguage
- En inglés: http://humanitiesprogramming.github.io/
- Estilometría con Python
  https://programminghistorian.org/en/lessons/introduction-to-stylometry-with-python

Recursos web para auto-aprendizaje o preparar materiales

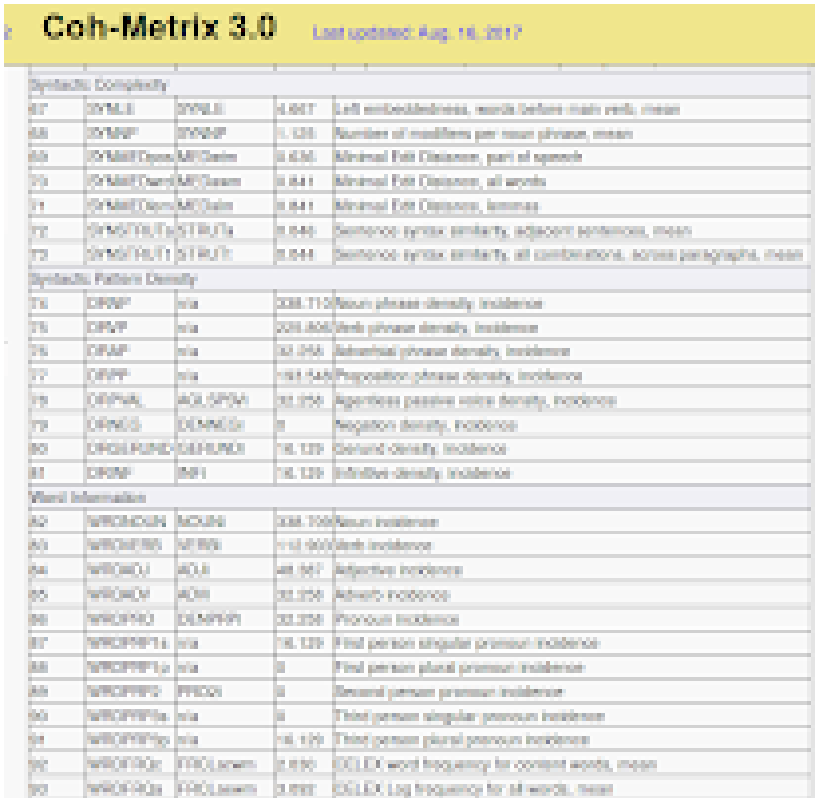
- Lextutor es una web con tecnología del lenguaje y recursos interesantes:
  - Leer texto y oírlo
  - Elegir una palabra en el texto y observar el uso de esa palabra en todo el libro
  - Crear ejercicios, huecos (borrando la palabra elegida...)





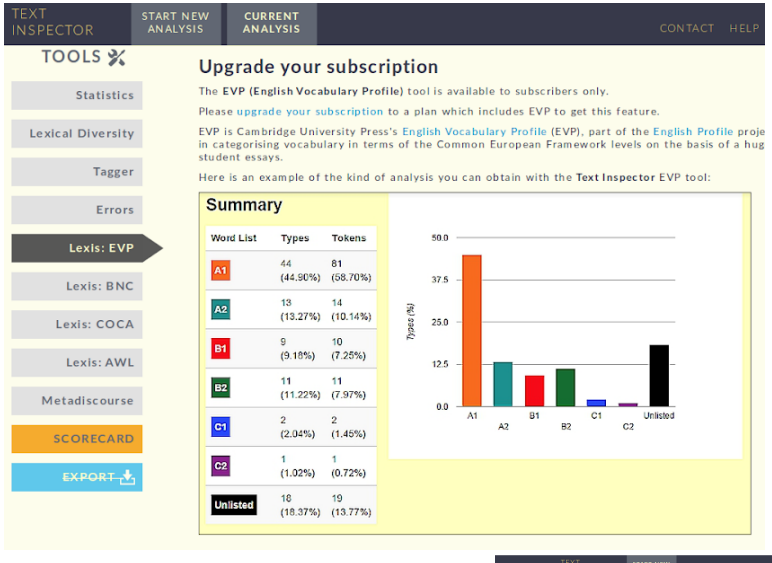
Extraer información de textos (I) (eng)

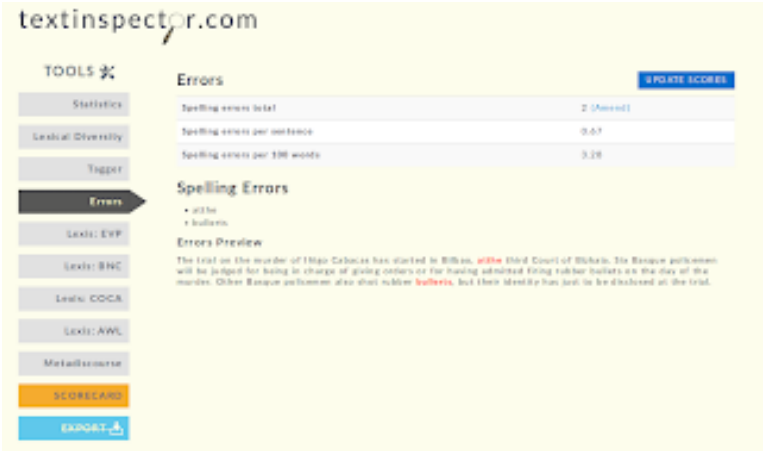
- La herramienta Coh-Metrix 3.0:
  http://tool.cohmetrix.com/
- Mide la complejidad de las palabras, oraciones, la cohesión y la coherencia textual

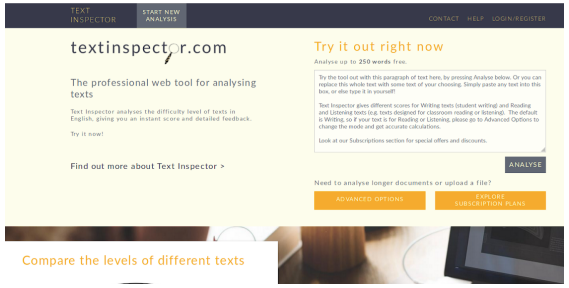


Extraer información de textos (II) (eng)

- Textinspector: https://textinspector.com/
- Mide la complejidad textual, determina el nivel y describe aspectos interesantes







Tecnologías del lenguaje y las learning analytics

- Learning analytics sirve para mejorar las interacciones en entornos virtuales de aprendizaje
  - Por ejemplo, en Moodle el plugin SmartKlass
- Partiendo de lo que ha pasado (acciones, comportamiento, decisiones y comunicación) pueden ayudar proponiendo actividades de éxito que sirvieron en el pasado
- Las tecnologías del lenguaje pueden "entender" el lenguaje del estudiante y pueden clasificar la complejidad léxica y sintáctica o partir de la similaritud textual [McNamara et al., 2017]

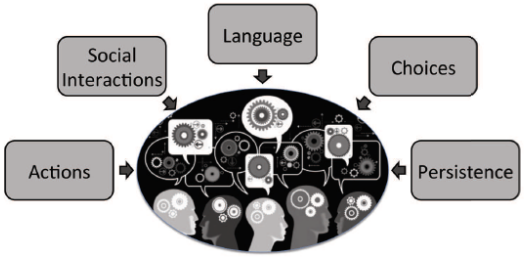


Figure 8.2. Predicting educational outcomes will require the integration of multiple sources of data.

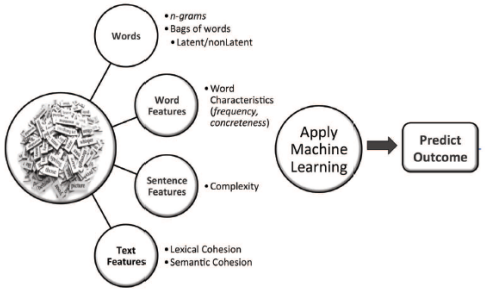


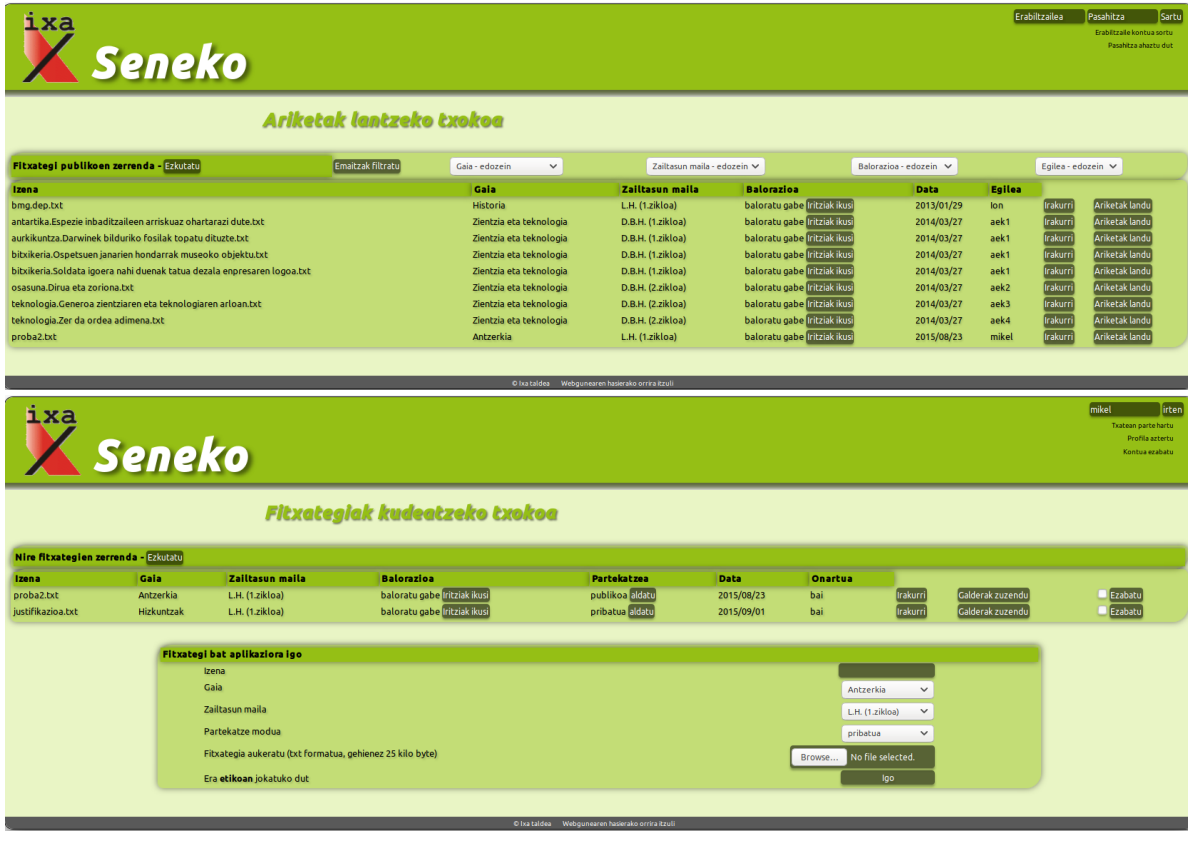
Figure 8.1. Developing algorithms using NLP requires machine-learning techniques applied to various sources of information on the text, including information from the words, sentences, and the entire text.

Evaluar los resúmenes de clase automáticamente

- Clasificación automática de resúmenes escolares: Compress-eus [Atutxa et al., 2017] o Compress-es
- En clase no se puede corregir todo semanalmente
  - Consecuencia: sólo se elaboran mapas mentales
  - El resumen y su feedback a un lado
  - El estudiante agradece observar las mejores versiones de otros o del docente
  - También se agradece una corrección automática en el momento de corrección y no una semana más tarde

Ejercicios automáticos

- SENEKO [López Gazpio, 2013] sistema para hacer preguntas y respuestas (con distractores)



Ejercicios orales (en un futuro?)

- En un futuro, los entornos de aprendizaje, deberán gestionar ejercicios orales y comunicativos
- Oír un ejemplo y poder grabar la del estudiante, comparando automáticamente estrategias de traducción: entonación, cambios de volumen, ritmo y pausas...
- Transcribir automáticamente de voz a texto
- Analizar automáticamente la voz y el texto
- Feedback automático y manual
- Chatbots para que el estudiante conteste de forma coherente pero con improvisación

Los grupos de investigación o los centro de conocimiento como CLARIN-K deben colaborar para la creación, uso y gestión de todas estas tecnologías

Futuro

- Colaboración entre los diferentes agentes en el aprendizaje de la lengua: HiTZ y CLARIN-k (entre otros)
- Promover investigaciones con experimentación y tomar decisiones basadas en datos
- Colaboración entre pedagogía y tecnologías del lenguaje
- Corpus multimodal y tecnologías del lenguaje

Carencias (por lo menos en nuestro entorno)

- Existen reticencias
- Existen hoy en día muchas posibilidades... pero también carencias
- Existen prácticas obsoletas en formato digital
- No hay un punto de vista crítico y pocas decisiones basadas en datos objetivos
- Falta análisis crítico de lo que más se necesita

Retos (por lo menos en nuestro entorno)

- Adecuar (en cierto grado) la investigación y la docencia al entorno virtual
- Adecuar las tecnologías del lenguaje a un entorno amigable para especialistas en humanidades
- Hacer HD con humanistas
- Formación de los humanistas en tareas multidisciplinares y multimodales
- Competencia digital e identidad digital

Preguntas

- ¿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?

Preguntas

- ¿Sabríais 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.
- ¿Cómo podemos contribuir?
  - Rellene esta Encuesta anónima
  - ¿Tienes textos?
  - Escríbenos a clarinkcenter@gmail.com

Bibliografía I

Alegria, I., Arregi, X., Artola, X., Astiz, M., and Ruiz, L. (2006).
Building an electronic version of the cuban basic school dictionary.
In *Proceedings XII EURALEX International Congress*, pages 243–250.

Atutxa, U., Ansa, O., Iruskieta, M., and Molina, A. (2017).
Compress-eus: i(ra)kasleen laburpenak lortzeko tresna.
In *EUDIA-6*, pages 1–8. EUDIA.

Bel, N., García, E. G.-B., and Iruskieta, M. (2016).
Clarin centro-k-español.
Procesamiento del Lenguaje Natural, (57):151–154.

Da Sylva, L. (2013).
Nlp and digital library management.
In *Emerging Applications of Natural Language Processing: Concepts and New Research*, pages 265–290. IGI Global.

Iruskieta, M. and Otegi, A. (2018).
Infraestructuras de pln y tareas básicas con python: una introducción práctica.
In *29ª edición Cursos de Verano UNED*. https://www.youtube.com/watch?v=RkKPK9pA2U0.

López Gazpio, I. (2013).
Seneko: galderak automatikoki sortuz testuak lantzeko aukera ematen duen aplikazioa.

Bibliografía II


McNamara, D. S., Allen, L. K., Crossley, S. A., Dascalu, M., and Perret, C. A. (2017).
Natural language processing and learning analytics.
Handbook of learning analytics, page 93.


Rodríguez, J. B. (2016).
An interview with paul seedhouse on video enhanced observation (veo): A new tool for teacher training, professional development and classroom research.
Bellaterra Journal of Teaching & Learning Language & Literature, 9(3):90–97.


Rodríguez-Roche, S. (2018).
¿qué humanidades digitales son posibles en cuba?
In *6th International Conference on Language Resources and Evaluation (LREC 2008)*.


Ruiz, L. (1999).
Primeros pasos de la etiquetación automática en cuba.
In *Actas del VI Simposio Internacional de Comunicación Social*, pages 710–714.


Schreibman, S., Siemens, N., and Unsworth, J. (2004).
A Companion to Digital Humanities.
Oxford: Blackwell.


Váradi, T., Wittenburg, P., Krauwer, S., Wynne, M., and Koskenniemi, K. (2008).
Clarin: Common language resources and technology infrastructure.
In *6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Tecnologías del lenguaje para la enseñanza e investigación en Humanidades Digitales

Universidad de La Habana, Facultad de comunicación La Habana, 26 /Enero / 2019 / Cuba

Para cualquier duda o ayuda con ANALHITZA
o sobre alguna otra cuestión de la charla
escriban a mikel.iruskieta@ehu.eus

FACULTAD
DE COMUNICACIÓN
UNIVERSIDAD DE LA HABANA

ixa