4. Corpus lingüísticos como bases de conocimiento

Contenido

4.1 Conceptos básicos

- 4.1.1 Definición de corpus lingüísticos
- 4.1.2 Clasificación

4.1.1 Definición de corpus lingüísticos

El Procesamiento del Lenguaje Natural abarca múltiples tareas en los diversos niveles de la lengua. En cambio, para realizar estas tareas, para llevar a cabo el análisis lingüístico en general, el investigador debe recolectar una serie de datos lingüísticos. Así pues, en PLN se suele trabajar con corpus lingüísticos.

(Sierra Martínez, 2017)

Corpus lingüístico

- **Corpus lingüístico**: es un conjunto de textos, ya sea escritos o hablados, recolectados debidamente para realizar análisis lingüísticos.
- El **análisis** que se realice puede ser tanto cualitativo como cuantitativo, es decir puede realizarse desde una perspectiva estructural (lingüística) o bien a partir de datos estadísticos.
- Un corpus puede formarse de cualquier tipo de texto, de cualquier género textual, siempre y cuando las muestras sean representativas y se compilen según criterios lingüísticos. De tal forma, un corpus constituye un **modelo** de realidad lingüística, mas no la realidad misma.

4.1.2 Clasificación de corpus

- Los corpus pueden clasificarse en diferentes tipos con respecto a sus elementos y a las características que poseen.

- En la imagen puedes observar una tabla de las distintas clasificaciones que se pueden hacer según cada característica específica. Un corpus puede pertenecer a más de un tipo.

- Por razones prácticas sólo explicaremos los corpus según el origen de los datos y por la especificidad de sus elementos, sin embargo al observar la tabla te habrás dado cuenta que no es nada difícil de entender el resto de las tipologías.

http://www.corpus.unam.mx/cursocorpus/Torruella_Llisterri_99.pdf

Origen de los datos	Orales		Sonoros
			Transcritos
	Textuales		
Espontaneidad	Premeditados		
	No premeditados		
Codificación y anotación	Simples		
	Codificados		
Especificidad de los elementos	Corpus generales		
	Corpus específicos	Literarios	Ensayo
			Narrativa
			Poesía
			Teatro
		Informativos	Periodísticos
			Científicos
			Académicos
			Técnicos
Autoría de los elementos	Genérico		
	Canónico		
	De autoría variada		
Temporalidad de los elementos	Sincrónicos	Contemporáneos	
		Históricos	
	Diacrónico	Cronológicos	
		Periódicos	
Propósito del estudio	Multipropósito	Referencia	
	Específico	Estudio	
		Entrenamiento	
		Prueba	
Lengua	Monolingües	De una variedad dialectal	
		Comparables	
	Multilingües	En distintos idiomas	
		Paralelos	
Cantidad de texto	Grande		
	Pequeño		
	Monitor		
Distribución del tipo de texto	Equilibrado	Piramidal	
	Desequilibrado		
Accesibilidad	Público	No comercial	Acceso restringido
			Acceso libre
		Comercial	
	Privado		
Documentación	Documentado		
	No documentado		
Representatividad	Representativo		
	Oportunista		

Clasificación de corpus

Según el **origen de los datos** los corpus pueden clasificarse como:

- **Textuales**: textos procedentes de la lengua escrita.
- **Orales**: grabaciones o transcripciones de la lengua hablada.

Clasificación de corpus

Según la **especificidad** de sus elementos los corpus pueden ser:

- **Generales:** aportan información general sobre los textos que recogen, y estos pueden pertenecer a cualquier género y tipología textual.
- **Específicos o especializados:** recogen textos que proporcionan información de áreas de especialidad, estos pueden ser textos informativos o literarios.

4.2 Recopilación y creación de corpus

4.2.1 Criterios para la recopilación de corpus

4.2.2 Compilación de corpus textuales

4.2.3 Compilación de corpus orales

4.2.1 Criterios para la recopilación de corpus

Ahora bien, como mencionamos todo corpus lingüístico debe ser debidamente recolectado y cumplir con ciertos criterios. Un corpus lingüístico debe:

Contener datos reales	Ser representativo	Ser selectivo
Tener una muestra variada…	y equilibrada	Ser finito

4.2.1 Criterios para la recopilación de corpus

a) **Contener datos reales:** esto debido a que con él se debe modelar la lengua, mostrar a pequeña escala cómo funciona una lengua natural.

b) **Ser representativo:** la muestra debe representar lo más posible a la población elegida en los aspectos que se desean estudiar, para ello se deberá establecer de antemano los objetivos del proyecto. Para ello la muestra debe ser **variada** y **equilibrada**.

- La variedad y el equilibrio se basan en los criterios específicos de cada proyecto.

4.2.1 Criterios para la recopilación de corpus

c) Variado: se refiere a la variedad de criterios de clasificación y búsqueda de información que se establecen para la creación del corpus, la variedad de información que se espera obtener de los textos y de la población de la que provienen. Por ejemplo, de dónde son (localidad), información personal de los informantes, tema o asuntos de que tratan los textos, tipo de texto, fuente, etc.

d) Equilibrado: debe existir un balance entre los rubros seleccionados para la construcción de el corpus (países, años, tipos de texto y tópicos), es decir, que el material para cada uno de los rubros sea relativamente proporcional.

Criterios para la recopilación de corpus

e) Selectivo: ya que no se puede recopilar todo lo escrito o hablado de una lengua, hay que hacer una selección cuidadosa de materiales, con base en los objetivos establecidos para el análisis.

f) Finito: como se menciona en el punto anterior al ser imposible recopilar todos los textos de una lengua entonces el corpus tendrá que ser finito.

4.2.2 Compilación de corpus textuales

Como sabes en la actualidad se generan una gran cantidad de documentos escritos, es por esto que se deben considerar ciertos criterios para la elaboración adecuada de nuestro corpus, para ello es necesario identificar el objetivo de nuestro estudio, seleccionar y obtener los documentos y conformar un equipo de trabajo. Hablaremos de estos pasos a continuación.

4.2.2 Compilación de corpus textuales

❶ Identificar el objetivo del estudio

Es importante identificar el propósito de nuestro corpus, este puede ser general o específico, por ejemplo si se utilizara para análisis lingüístico puramente, para enseñanza de idiomas, investigación en PLN, para aplicaciones de ingeniería lingüística, etc. También habrá que pensar en otros aspectos relacionados como el tipo de notación que se usará o los límites que tendrá.

4.2.2 Compilación de corpus textuales

Límites de un corpus

- Límites diatópicos: zonas geográficas.
- Límites dialectales: con relación a un tronco común, que puede determinarse por la geografía.
- Límites de género textual: literarios, técnicos, periodísticos.
- Límites temáticos: específicos vs. generales.
- Límites en tamaño: pequeño, mediano o grande.

4.2.2 Compilación de corpus textuales

❷ **Selección de textos:**

La selección de textos se hará en función de los objetivos del proyecto y habrá que tomar en cuenta el balance y la representatividad. Los corpus con objetivos generales suelen contener gran cantidad de elementos, los cuales se deberán clasificar por temas y contener los datos lingüísticos relevantes.

4.2.2 Compilación de corpus textuales

❸ Obtención de textos:

Para la obtención de textos se deberán buscar los documentos, también se pueden solicitar directamente a quien pertenezcan, y es importante pedir la autorización de uso o cartas de autor.

Una vez obtenidos los materiales, se deberán organizar en una base de datos en la que se indique la procedencia y estado de los documentos.

4.2.2 Compilación de corpus textuales

Internet como fuente de información

Si bien el tipo y la obtención de los documento depende directamente de los objetivos del proyecto, en la actualidad el internet se presenta como una gran base de datos de la cual es posible seleccionar y obtener materiales para conformar un corpus.

Internet no es un corpus como tal ya que sus documentos no cuentan con una clasificación general ni con una codificación estandarizada, sin embargo, son una fuente inagotable de materiales viables para realizar análisis lingüístico y por ello para formar corpus, siempre y cuando se haga una adecuada selección.

4.2.2 Compilación de corpus textuales

❹ Digitalización de los documentos:

Si bien con fines prácticos, es preferible contar con documentos en formato electrónico desde la recolección, no siempre será posible, así que habrá que digitalizar los materiales por nosotros mismos. El proceso de digitalización se divide comúnmente en tres etapas: la digitalización del documento, el reconocimiento de los caracteres y el guardado del documento. Para esto existen dos procesos alternos que convertirán nuestros textos a formato electrónico.

4.2.2 Compilación de corpus textuales

- **Digitalizador de imágenes o escáner**: probablemente ya conoces y has usado uno de estos. Se trata de un equipo que transforma una imagen analógica en una digital, el escáner reconoce los textos como imágenes, no como letras.

4.2.2 Compilación de corpus textuales

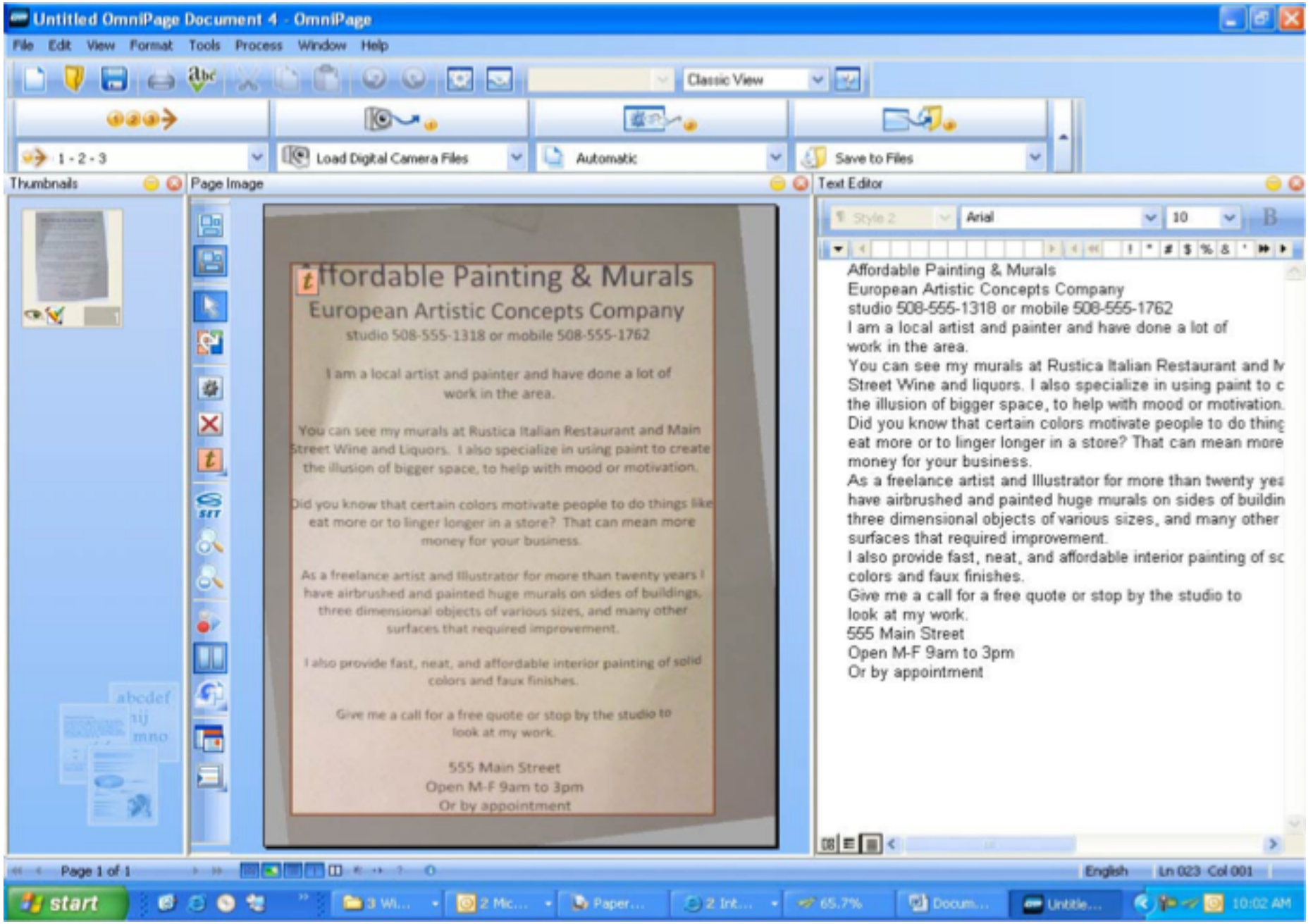
- **Reconocedores de texto**: son programas que leen las imágenes digitales de texto y buscan conjuntos de puntos que se asemejan a letras. Existen dos tipos de reconocedores:

  - ***Optical Character Recognition*** **(OCR):** convierte las imágenes en archivos de texto, reconoce los caracteres tipográficos y los hace legibles para la computadora, funciona con textos escritos a máquina.

  - ***Intelligent Character Recognition*** **(ICR):** es un OCR avanzado, puede usarse con manuscritos, al aplicar inteligencia lógica convierte las imágenes a formato de texto de manera más confiable, pues aplica reglas ortográficas, gramaticales y de contexto.

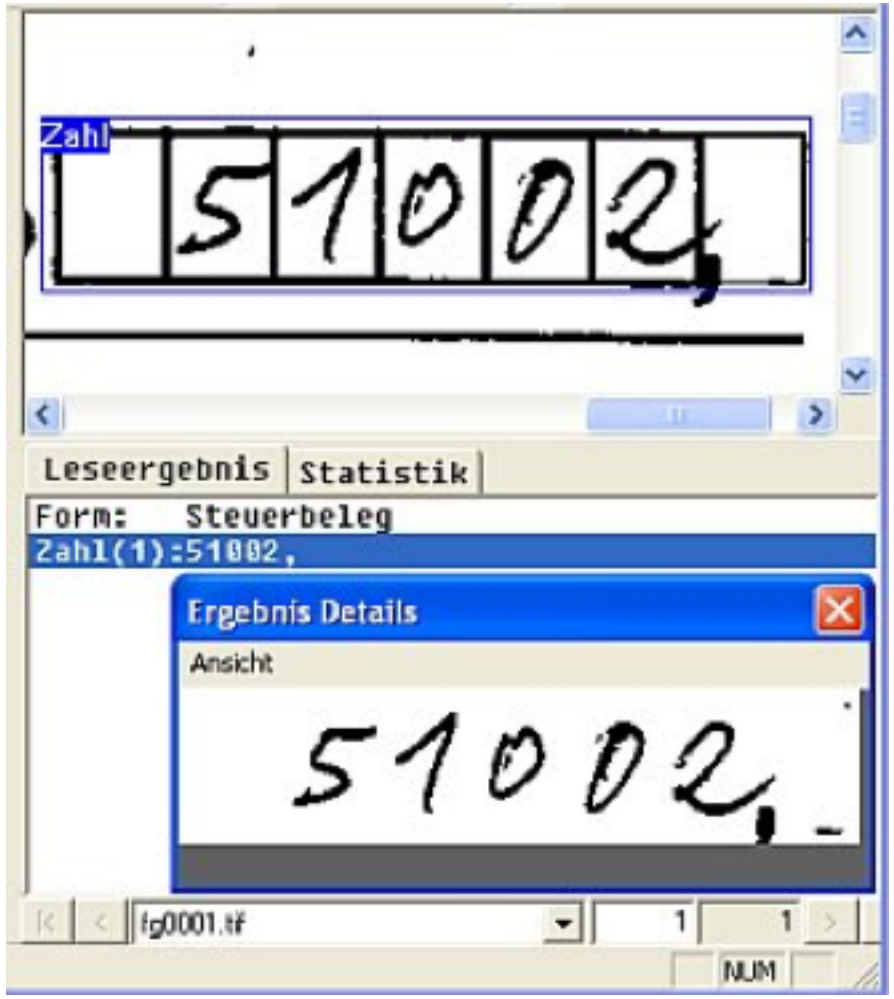
OCR

Optical Character Recognition



ICR

Intelligent Character Recognition



4.2.2 Compilación de corpus textuales

El reconocimiento de caracteres es dependiente del idioma, ya que cada idioma tendrá caracteres especiales.

Entre algunos programas de reconocimiento de texto tenemos el OmniPage, que reconoce texto en 119 idiomas, terminología legal y médica, así como el ABBYY, que reconoce textos con letra manuscrita.

- OmniPage: www.nuance.es/particulares/producto/omnipage/index.htm
- ABBYY: http://es.abbyy.com/

4.2.2 Compilación de corpus textuales

❺ Estandarización del formato

Como mencionamos, los documentos recolectados pueden tener distintas procedencias y por consiguiente diversos formatos (pdf, doc, html, etc.), a fin de optimizar el manejo y procesamiento de estos, será necesario estandarizar su formato a modo que todos tengan el mismo.

Según el tipo de información que se requiere anotar, quizá convenga guardar los archivos en texto plano, aunque cabe mencionar que para este formato también existen diversas codificaciones (ASCII, UNICODE, UTF-8 e ISO-9959-1) las cuales pueden ocasionar serios problemas en la lectura de ciertos símbolos.

4.2.2 Compilación de corpus textuales

❻ Administración del proyecto:

Por último, antes dijimos que en ciertas ocasiones nos veremos en la necesidad de conformar un equipo de trabajo que nos ayude con las tareas de elaboración del corpus, para ello los colaboradores deberán desempeñar alguno de los siguientes roles:

4.2.3 Compilación de corpus orales

Para elaborar este tipo de corpus se consideran los **mismos aspectos que para los corpus textuales** (identificar el objetivo del corpus, seleccionar y obtener el material y tener un administrador del proyecto). Sin embargo, también habrá que considerar la característica de los hablantes o informantes (género, edad, lugar de residencia, etc.), así como de las herramientas para anotar el corpus, el tipo de transcripción y alfabeto que se va a utilizar, además de tomar en cuenta las tecnologías con las que se puede trabajar el habla.

4.3 Anotación de corpus

4.3.1 Conceptos básicos

4.3.2 Lenguajes y principios

4.3.4 Niveles

4.3.1 Conceptos básicos: definición

- La **anotación de corpus** consiste en codificar cierto análisis lingüístico de los datos en un corpus. A menudo se le llama también **etiquetado de corpus.**
- Dicho análisis debe presentarse de manera sistemática y accesible.
- Puede incluirse junto con los datos del corpus o almacenarse por separado pero ligado a estos *(stand-off annotation).*
- Puede variar dependiendo del tipo de análisis para el que se requiera, sin que eso signifique sacrificar la sistematicidad.

(McEnery & Hardie, 2012)

Ejemplo de texto no anotado y anotado

Con diez cañones por banda, viento en popa a toda vela,

no corta el mar, sino vuela, un velero bergantín;

bajel pirata que llaman, por su bravura, "El Temido";

en todo mar conocido del uno al otro confín.

La luna en el mar riela, en la lona gime el viento

y alza en blando movimiento olas de plata y azul;

Con/sps00 diez/dn0cp0 ca\xc3/-None- ±/-None-ones/ncfp000 por/sps00 banda/ncfs000 ,/Fc viento/ncms000 en/sps00 popa/np0000l a/sps00 toda/di0fs0 vela/ncfs000 ,/Fc no/rn corta/aq0fs0 el/da0ms0 mar/nccs000 ,/Fc sino/cc vuela/ncfs000 un/di0ms0 velero/cc bergantĂ/-None- /-None-n/nccn000 ;/Fx bajel/-None- pirata/nccs000 que/pr0cn000 llaman/vmip3p0 ,/Fc por/sps00 su/dp3cs0 bravura/ncfs000 ,/Fc "/Fe El/da0ms0 Temido/vmp00sm ";/-None- en/sps00 todo/di0ms0 mar/nccs000 conocido/aq0msp del/spcms uno/pi0ms000 al/spcms otro/di0ms0 confĂ/-None- /-None-n/nccn000 ./Fp La/da0fs0 luna/ncfs000 en/sps00 el/da0ms0 mar/nccs000 riela/ncfs000 ,/Fc en/sps00 la/da0fs0 lona/ncfs000 gime/aq0cs0 el/da0ms0 viento/ncms000 y/cc alza/ncfs000 en/sps00 blando/aq0ms0 movimiento/ncms000 olas/ncfp000 de/sps00 plata/ncfs000 y/cc azul/aq0cs0 ;/Fx

4.3.1 Conceptos básicos: importanica

La anotación o etiquetado de corpus es vital ya que:

- Facilita la extracción de información.
- Permite el uso de los datos en más de un estudio y en estudios de diversa índole, es decir en diferentes niveles de la lengua.

(McEnery & Hardie, 2012)

4.3.1 Conceptos básicos: tipos y enfoques

❶ **Metadatos**: datos sobre la fuente de la información

- En corpus textuales: el autor del texto, fecha de publicación, etc.
- En corpus orales: identificación de cada hablante en el texto (turnos), información sociodemográfica del informante (género, edad, nivel socioeconómico, etc.)

❷ **Marcado textual *(Textual markup)***:

- XML *(eXtensible Markup Language)*

❸ **Análisis lingüístico**: información sobre los datos en los diferentes niveles de la lengua.

Enfoques en anotación	• **Meramente automática**
	• Automática seguida por corrección manual
	• Meramente manual

4.3.1 Conceptos básicos: elementos

- **Entidad de marcaje:** objeto concreto cuyo marcaje resulta de interés. Por ejemplo, un carácter o conjunto de caracteres, una palabra, una línea, etc.
- **Elemento de marcaje:** está constituido por los elementos del documento que se anotarán para su procesamiento. Se les delimita por una etiqueta de apertura <> y una de cierre </> .

  <título> Canción de Espronceda </título> <autor> José de Espronceda</autor>

- **Atributo de marcaje:** proporciona información adicional de un elemento. Se separan del elemento por un espacio en blanco y se escriben en la etiqueta de apertura.

  <título género="poema corriente"="romanticismo"> Canción de Espronceda </título>

4.3.2 Lenguajes de etiquetado

Como se mencionó anteriormente, las necesidades de extracción de información pueden variar y con ellas, la forma de etiquetado. Esto no significa que se pierda sistematicidad, gracias a los diversos lenguajes de etiquetado.

A continuación se muestran los lenguajes más usados en orden cronológico

4.3.2 Lenguajes de etiquetado

- **GML (*General Markup Language*):** Almacena los metadatos separados del contenido. Describe formato, estructura y contenido.
- **SGML (*Standard Generalized Markup Language*):** Permite al usuario definir las etiquetas. Es sumamente estructurado.
- **TEI: (*Text Encoding Initiative*):** Establece lineamientos para el etiquetado de documentos en lingüística, ciencias sociales y humanidades. Permite etiquetar imagen y sonido.
- **HTML (*HyperText Markup Language*):** Permite señalar referencia cruzada en el mismo documento o en otros, incorporando hipervínculos. Es usado en la creación de páginas web.
- **XML (*eXtensible Markup Language*):** Regula la creación de lenguajes de marcas. Permite al usuario definir las etiquetas. Es el más usado en construcción y manejo de corpus.

4.3.2 Principios de la anotación de corpus

- **Inteligibilidad**: El nombre de las etiquetas debe ser claro y descriptivo. Por ejemplo, para etiquetar primera persona del plural, conviene usar 1PL en vez de algo confuso como ppp.
- **Extracción**: Debe ser posible separar las anotaciones del corpus y guardar cada conjunto de datos de independientemente.
- **Intercambio**: Debe ser posible sustituir las etiquetas originales del corpus con otras que se ajusten a otro tipo de análisis.
- **Documentación**: Debe hacerse explícito el sistema de codificación, así como los responsables de esta.
- **Estandarización**: El etiquetado de un corpus debe ceñirse consistentemente a principios claramente establecidos; sin embargo, no existe un esquema de anotación único que sirva para todos los tipos de análisis, por lo que se permite ajustarlos por razones prácticas.

4.3.2 Estandarización

Algunas instituciones han buscado estandarizar la anotación de corpus.

- **LDC**: Linguistic Data Consortium

https://www.ldc.upenn.edu/

- **EAGLES**: European Advisory Group on Language Engineering Standards

http://www.ilc.cnr.it/EAGLES/intro.html

- **TEI**: Text Encoding Initiative

http://www.tei-c.org/Vault/P4/doc/html/

Text Encoding Initiative

The XML Version of the TEI Guidelines

4.3.3 Niveles

1. Textual
2. Ortográfico
3. Fónico: fonético, fonológico, prosódico
4. Morfológico
5. Morfosintáctico
6. Sintáctico
7. Semántico
8. Discursivo
9. Pragmático y estilístico

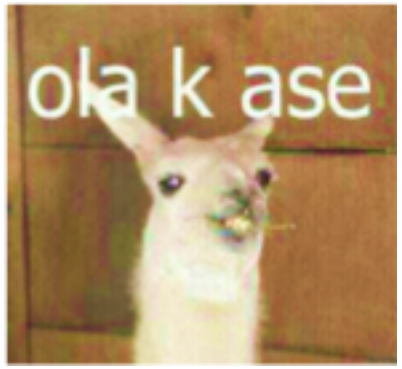
4.3.3 Etiquetado textual y ortográfico

- **Etiquetado textual**
  - Distingue unidades en las que se divide el texto (capítulos, secciones, párrafos, oraciones).
  - Etiqueta ciertos elementos como títulos y subtítulos.
  - Puede etiquetar el tipo, el tamaño de letra y la fuente.
  - Etiqueta por tipo de texto (artículo de revista, tesis, novela, etc.)
- **Etiquetado ortográfico**
  - Es llamado también “transliteración”.
  - Asocia escritura común (ortográfica) a los elementos del corpus.
  - Incluye mayúsculas y minúsculas, acentuación, signos de puntuación, etc.

(Sierra,)

4.3.3 Etiquetado fónico

- **Etiquetado fonético**
  - Describe las manifestaciones físicas de los sonidos.
- **Etiquetado fonológico**
  - Considera la función de los sonidos dentro del sistema lingüístico.
  - Establece valores distintivos dentro del inventario de sonidos de una lengua.
  - Es útil para textos sin norma académica, como los hallados en redes sociales o textos antiguos.



- **Etiquetado prosódico**
  - Incluye acento, melodía, entonación, pausa, velocidad, ritmo y cualidad de la voz.

(Sierra,)

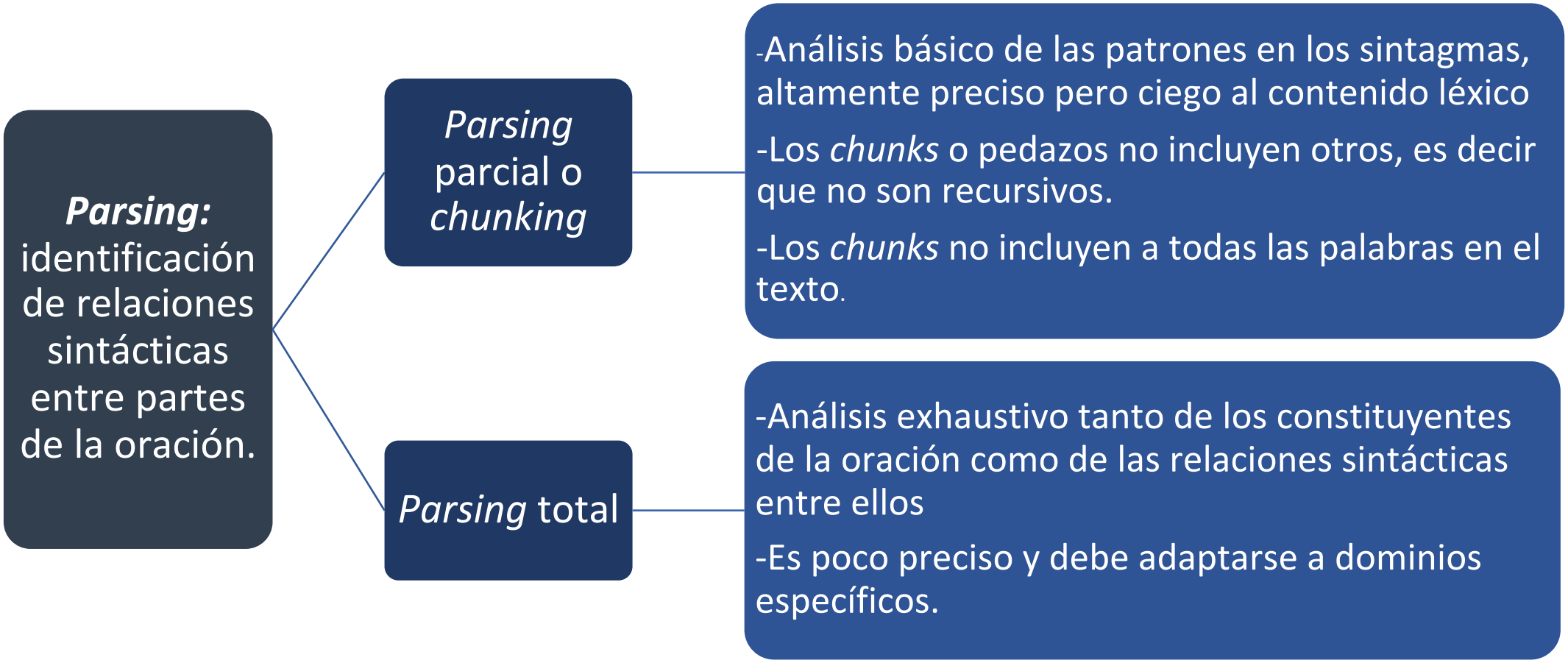
4.3.3 Etiquetado morfológico

- Identifica morfemas dentro de una palabra.
- En los sustantivos se etiquetan morfemas de género (-a,-o, -e) y número (-s, -es)
- En los verbos se etiquetan morfemas de tiempo, aspecto, modo y persona (conjugación).
- También se puede obtener el lema (forma de diccionario) de una palabra mediante la lematización, abordada en la sesión 3. Esto facilita la indización y la recuperación de la información.

4.3.3 Etiquetado morfosintáctico

- Es el etiquetado de las partes de la oración y también se le conoce como etiquetado POST (*Part Of Speech Tagging*).
- Consiste en etiquetar el tipo de palabra o categoría gramatical (sustantivo, verbo, adjetivo, etc.)
- Es el tipo de etiquetado más usado.
- Puntos esenciales para el etiquetado POST:
  - Identificación de palabras o unidades léxicas
  - Definición de las categoría gramaticales
  - Definición de etiquetas con las que se anotarán las clases de palabras
  - Métodos para etiquetar partes de la oración

4.3.3 Etiquetado sintáctico



4.3.3 Etiquetado semántico

- **Objetivos**

• Desambiguar el sentido de las palabras	• Identificar relaciones léxico semánticas

- **Tipos de anotación semántica**
  - **Características semánticas:** etiqueta no sólo los significados, sino también otras características, como su pertenencia a un marco semántico.
  - **Anotación ontológica:** describe formalmente lo conceptos de un corpus y las relaciones entre estos. Es usado en documentos de internet para facilitar la búsqueda.
  - **Anotación de relaciones semánticas**: marca relaciones léxico-semánticas (sinonimia, antonimia, polisemia) y relaciones entre elementos del texto, como los participantes de una acción (agente, paciente, etc.)

4.3.3 Etiquetado discursivo

- Tiene como objetivo identificar:

• Emisores y receptores	• Normas que regulan la situación
• Temas en construcción	• Efectos de la comunicación

- Si bien este tipo de etiquetado no ha sido estandarizado, la Rhetorical Structure Theory que se abordó en la sesión anterior permite hacer una descripción de la estructura discursiva de un corpus.
- Resulta útil para la resolución de anáforas, pues permite etiquetar expresiones anafóricas y antecedentes.

Ejemplo de etiquetado de relaciones anafóricas

- Nótese que se marca tanto el inicio como el final de los antecedentes y las expresiones anafóricas.
- Además estas se numeran, compartiendo número los antecedentes y las expresiones anafóricas que comparten referente.

<ANT id=?1?>Las zonas costeras</ANT> han tenido importantes actividades agropecuarias, por lo que a cortas distancias <ANAF id=?1?> disponen </ANAF> de <ANT id=?2?>alimentos</ANT>, a <ANAF id=?2?>los </ANAF> que es necesario adicionar <ANT id=?3?>los productos pesqueros</ANT> que <ANAF id=?3?>los </ANAF> tienen a la mano y al no necesitar enlatar<ANAF id=?3?>los</ANAF> ni congelar<ANAF id=?3?>los</ANAF>, <ANAF id=?3?>su</ANAF> costo es inferior a un 50% de <ANAF id=?3?> lo</ANAF> que cuestan en el altiplano.

4.3.3 Etiquetado pragmático y estilístico

- No está estandarizado y depende del propósito del corpus.
- Permite identificar
  - Ironía
  - Sentimientos
  - Opiniones
- Puede marcar:
  - Actos de habla
  - Entoncación
  - Gestos durante la emisión
  - Intensidad de la emisión
  - Pausas
  - Polaridad

4.4 Herramientas informáticas especializadas y técnicas de análisis

4.4.1 Herramientas de recopilación

4.4.2 Herramientas de etiquetado

4.4.3 Técnicas de análisis

4.4.4 Herramientas de análisis

4.4.1 Herramientas de recopilación

De corpus textuales:

- Digitalizador o escáner
- Reconocedores de texto
  - *Optical Character Recognition* (OCR)
  - *Intelligent Character Recognition* (ICR)

De corpus orales:

- Speech Viewer
- Praat
- Sound Forge
- Speech Tools
- WaveLab
- CSLU ToolKit

4.4.2 Herramientas de etiquetado

- Herramientas de etiquetado morfosintáctico:
- Analizador morfológico, etiquetador y parser del Grupo de PLN de la UPC disponible en: http://www.lsi.upc.edu/~nlp/SVMTool/demo.php
- Memory Based Tagging Demo disponible en: https://github.com/LanguageMachines/mbtserver
- Freeling disponible en: http://nlp.lsi.upc.edu/freeling/

4.4.3 Técnicas de análisis

- Sin importar qué tan robusto sea o qué tan bien haya sido recopilado un corpus, sus datos no son de utilidad más que con un análisis adecuado. Para ello existen las siguientes técnicas:

❶ Conteo de palabras

❷ Concordancias

❸ Colocaciones

4.4.3 Conteo de palabras

- Para esta técnica de análisis es necesario tokenizar el texto. Recuerda que los conceptos de *token, type* y tokenización se abordaron en la sesión 4.

Temprano *levantó la muerte el vuelo*

temprano *madrugó la madrugada*

temprano *estás rondando por el suelo*

3 tokens, 1 type

- El número total de *tokens* determina el tamaño del corpus.

4.4.3 Conteo de palabras

- **Frecuencia absoluta:** número de ocurrencias (*tokens*) de cada *type.*
- **Frecuencia relativa:** número de ocurrencias (*tokens*) de cada *type* en relación al número de palabras en el corpus.
- **Riqueza léxica:** relación entre los *types* y *tokens.* Esta se puede obtener fácilmente con Python.

```
tokens_conjunto=set(tokens)

palabras_totales=len(tokens)
palabras_diferentes=len(tokens_conjunto)

riqueza_lexica=palabras_diferentes/palabras_totales

print(riqueza_lexica)
```

4.4.3 Conteo de palabras

Las palabras (*types*) junto con sus frecuencias absolutas pueden enlistarse de diferente manera y siguiendo distintos órdenes.

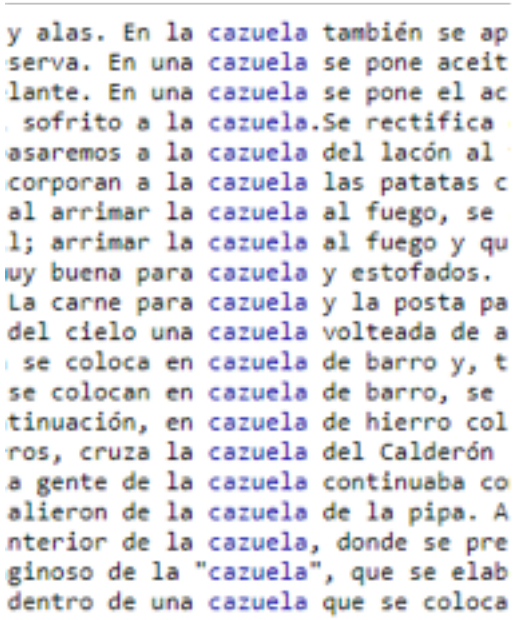
Listas de palabras		Orden de las listas	
• Simple	• Lemas	• Alfabético	• Alfabético inverso
• De formas canónicas	• Dos o más palabras	• Frecuencias	• Categoría gramatical
• Partes de la oración		• Longitud	

4.4.3 Concordancias

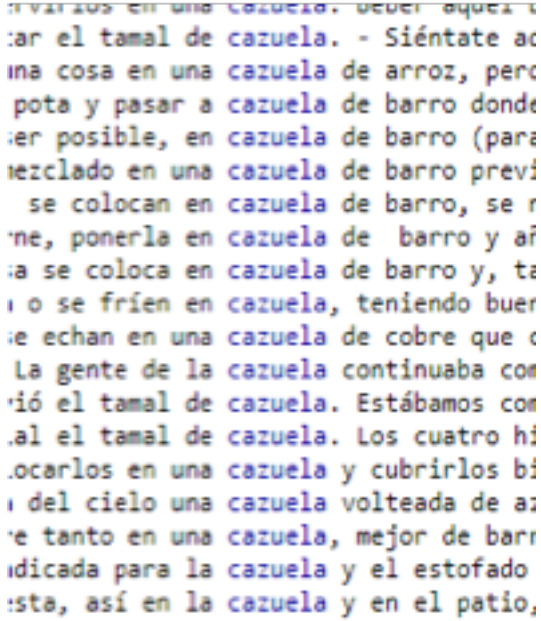
- **Concordancias:** palabras de un corpus que aparecen en un contexto, llamadas también KWIC (*Key Words in Context*).
- Cuando se consultan las concordancias de una palabra en un corpus, estas se visualizan en una columna con las palabras que le anteceden a la izquierda y las que le suceden a la derecha.
- **Ventana:** cantidad de texto que puede visualizarse al recuperarse las concordancias. Puede establecerse por número de caracteres o números de palabras.
- Además en posible ordenar las concordancias por orden alfabético de las palabras de su contexto.

4.4.3 Concordancias

En la primera imagen las concordancias están ordenadas a partir de las dos palabras que ocurren a su izquierda.



Mientras que en la segunda, aparecen ordenadas a partir de las dos palabras que ocurren a su derecha.



Actividad formativa

- Visita el sitio del Corpus de Referencia del Español Actual: http://corpus.rae.es/creanet.html

- Realiza la búsqueda de las concordancias de las expresión *fucú.* ¿Cuántas concordancias hay en dicho corpus? ¿En qué país se produjo el texto en el que aparece y qué temática tiene este?

- Realiza la búsqueda de la expresión *tirar la casa por la ventana.* ¿Cuántas concordancias hay en textos mexicanos?

 Respuesta

- Hay sólo una concordancia de *fucú* y es de República Dominicana.
- Hay 27 concordancias de *tirar la casa por la ventana* y ninguna se encuentra en un texto mexicano.

4.4.3 Colocaciones

- **Colocaciones:** en PLN, se refiere a dos o más palabras que, además de encontrarse cercanas en un texto, tienen la tendencia a ocurrir en contextos específicos. Sus tipos y elementos ya fueron abordados en la sesión 4.
- Se cuenta las palabras cercanas a la izquierda y a la derecha del nodo (palabra cuyas colocaciones se busca determinar) y se realiza un conteo de sus concordancias.

4.4.3 Colocaciones

- **Información mutua:** medida estadística que permite medir la fuerza de asociación entre dos palabras. Calcula la probabilidad de que dos ocurran juntas y la compara con la probabilidad de que aparezcan aisladas.

$$MI(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)}$$

Esta medida y otras usadas en la extracción de colocaciones se abordarán más a fondo en la sesión 10 – Extracción de Información.

4.4.4 Herramientas de análisis

Estas son tres de las herramientas de análisis textual más usadas para analizar corpus:

- WordSmith Tools
- T-Lab
- Goldvarb

4.4.4 WordSmith Tools

- Software comercial para el estudio del comportamiento de las palabras en un corpus desarrollado en la Universidad de Liverpool
- Consta de tres herramientas básicas que se pueden complementar con otros instrumentos:
  - Wordlist
  - KeyWords
  - Concord

4.4.4 WordSmith Tools

- **Wordlist**
  - Enlista las palabras de un corpus ya sea en orden alfabético o por frecuencias.
  - Analiza estadísticamente la distribución de las unidades léxicas.
- **KeyWords**
  - Identifica las palabras clave y su distribución en un corpus.
- **Concord**
  - Da información sobre el contexto de palabras o frases.
  - Permite identificar colocaciones.
  - Permite identificar el significado de una palabra a partir de sus contextos de aparición.

4.4.4 T-Lab

- Software comercial desarrollado por Franco Lancia
- Extrae, compara y representa de manera gráfica las características lingüísticas de textos de diversos tipos.
- Dos de sus ventajas es que puede adaptarse dependiendo de las necesidades de análisis y tiene versiones para diferentes lenguas, incluyendo el español.

Etapas de trabajo con T-Lab

❶ Procesamiento del corpus

- Estandarización, segmentación, lematización, selección de palabras clave

❷ Análisis de coocurrencias

- Identificación contextos de las expresiones, palabras con las que coocurre una expresión y distribución de la asociación entre dos o más palabras

❸ Análisis temático

- Extracción de contextos significativos para resumir el contenido, identificación de relaciones entre los temas principales, agrupación de documentos temáticamente

❹ Análisis comparativo

- Comparación de textos a partir de sus temas y expresiones típicas o exclusivas

4.4.4 Goldvarb

- Creado en el Departamento de Lenguas y Ciencias Lingüísticas de la Universidad de York
- Software de análisis estadístico multivariable diseñado para la sociolingüística variacionista.
- Se basa en el modelo logístico de variación.
- Permite determinar las variables lingüísticas y extralingüísticas que condicionan la ocurrencia de un fenómeno de variación y la manera en la que estas interactúan.
- Realiza tanto análisis binomial de un solo nivel, como de ascenso y descenso.

Links a las herramientas

En las siguientes páginas podrás encontrar mayor información sobre el uso y aplicaciones de las herramientas que se presentaron, así como instrucciones para obtenerlos.

- WordSmith: http://www.lexically.net/wordsmith/
- T-Lab: http://www.tlab.it/default.php
- GoldVarb: http://www.individual.utoronto.ca/tagliamonte/goldvarb.html

4.5 Corpus existentes

- El uso de la informática para la elaboración de corpus inició en 1949 con la transcripción de la obra de Santo Tomás de Aquino a tarjetas perforadas, lo que constituyó el ***Index Tomisticus.***
- En 1960 se realizó el ***Survey of English Usage*** para la descripción del habla culta británica.
- Simultáneamente se creó el ***Brown Corpus*** para el inglés americano.

Actualmente existe una gran variedad de corpus orales y textuales, tanto especializados como generales que permiten describir el español de diferentes regiones y extraer información sobre ámbitos específicos. A continuación se enlistan algunos de los corpus informatizados para el español más representativos.

4.5 Corpus orales

- Corpus Diálogos Inteligentes Multimodales en Español (DIME) y DIMEx100
- Corpus de Investigación en Español de México el Posgrado de Ingeniería Eléctrica y Servicio Social (CIEMPIESS)
- Corpus Sociolingüístico de la Ciudad de México (CSCM)

4.5 Corpus textuales

- Corpus del Español Mexicano Contemporáneo (CEMC)
- Corpus Diacrónico y Diatópico del Español de América (CORDIAM)
- Corpus Textual Especializado Plurilingüe
- Corpus Lingüístico en Ingeniería
- Corpus de las Sexualidades en México (CSMX)
- Corpus Histórico del Español en México (CHEM)
- Corpus de Contextos Definitorios (COCORDE)
- RST (Rhetorical Structure Theory) Spanish Treebank
- Corpus Electrónico para la Enseñanza de la Lengua Escrita

4.5 Corpus textuales

- Corpus Científico del Español de México (COCIEM)
- Corpus Electrónico del Español Colonial Mexicano (COREECOM)
- Archivo de Textos Hispánicos de la Universidad de Santiago de Compostela (ARTHUS)
- Corpus de Aprendices del Español (CAES)

4.5 Corpus orales y textuales

- Corpus Diacrónico del Español (CORDE)
- Corpus de Referencia del Español Actual (CREA)
- Corpus del Nuevo Diccionario Histórico del Español (CDH)
- Corpus del Español de Mark Davies
- Corpus de Ah-hocracia (CAC)

Referencias

- Bird, S. and Klein, E. and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media ISBN: 9780596555719 (Chapter 2: Accessing Text Corpora and Lexical Resources & Chapter 11: Managing Linguistic Data)
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press. (Chapter 1: Goals and Methods of the corpus-based approach)
- Manning, C. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA. (Chapter 4: Corpus-based work)
- McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press. (Chapter 1: What is Corpus Linguistics? Chapter 2: Accessing and Analysing Corpus Data)
- Sierra Martínez, G.E. (2017). *Introducción a los Corpus Lingüísticos*. Universidad Nacional Autónoma de México