Aula 10

CONTRIBUCIONES DE LA LINGÜÍSTICA DE CORPUS

META

Presentar conceptos básicos y contribuciones de la Lingüística de Corpus para la enseñanza de lenguas extranjeras.

OBJETIVOS

Al final de esta clase el alumno deberá ser capaz de:
Comprender qué es Lingüística de Corpus, sus objetivos, conceptos básicos y sus contribuciones a las clases de lenguas extranjeras.

PREREQUISITOS

Conocimientos básicos de Lingüística.

Carlos Augusto Santos Vieira
Sandro Marcío Drumond Alves Marengo

INTRODUCCIÓN

La Lingüística de Corpus es una ciencia o metodología (ZAPATA, 2005) relativamente reciente. Posiblemente, por ese motivo, son estudios que no forman parte de modo más profundizado del currículo de muchas universidades.

Imagen 01 – Lingüística de Corpus. Disponible en: https://bit.ly/2MjY96N

Pero… ¿Cuál es la relación entre Lingüística de Corpus y léxico, diccionarios y enseñanza de lenguas extranjeras? ¿Qué entiendes como corpus? ¿Por qué digital? Al final de esta clase se espera que seas capaz de contestar a estas preguntas. En esta clase vamos a revisar las características de la Lingüística de Corpus, sus aportaciones a la enseñanza de lenguas extranjeras y algunos conceptos importantes como corpus, colocación y frecuencia.

TEORÍA, METODOLOGÍA E INCIDENCIA EN LA ENSEÑANZA DE LENGUAS

Inicialmente, vamos a ver algunas definiciones que recibe la Lingüística de Corpus. Estos estudios son definidos como teoría y como metodología. Para Tognini-Bonelli:

> "Mientras que una metodología puede ser definida como el uso de un conjunto de reglas dadas o porciones de conocimiento en una situación específica, (…) la lingüística de corpus está en posición de definir su propio conjunto de reglas y porciones de conocimientos antes de ser aplicados. Esto lleva a los lingüistas a hacer uso de nuevos parámetros para explicar los datos, y esto implica un cambio en lo que puede ser considerado como la unidad de aceptación para

> la descripción lingüística. De este modo, la lingüística de corpus adquiere un estatus teórico que la coloca en posición de contribuir de manera específica a otras aplicaciones (TOGNINI-BONELLI, 2005, apud ZAPATA, 2015, pp. 61-62).

Por otro lado, para McEnery y Wilson (1996, p. 1, apud ZAPATA, 2015), la Lingüística de Corpus es más un estudio lingüístico con base en ejemplos de uso de la 'vida real' y una metodología que un aspecto del lenguaje que requiera una explicación o una descripción (como ocurre con las teorías). De este modo, la Lingüística de Corpus, a partir especialmente de la creación de bancos de datos, aporta a la enseñanza de lenguas una metodología que nos permite reconocer cuáles son los usos y patrones lingüísticos más utilizados por los hablantes.

El reconocimiento de estos usos y patrones confirma la importancia del uso de textos auténticos (muestras reales de lengua) en las clases de lengua extranjera. Los autores de libros didácticos y profesores, si no consideran análisis de corpora, parten de intuiciones de usos que podemos decir insuficientes. En el caso del español, sabemos que es una lengua hablada en 21 países. De este modo, ¿cómo autores de libros didácticos y profesores tendrán condiciones de comprender usos más frecuentes de una palabra o expresión apenas por intuición? Si se trata de un dialogo simple para enseñar presentaciones, ¿cuál debe ser la forma elegida para desarrollar una conversación: a) ¿Cómo está(s)? b) ¿Cómo le(te) va? c) ¿Cómo andas? d) ¿Qué hacés? Además de eso, ¿cuáles serían los criterios? ¿Estos criterios privilegian cuál acento/ área geolectal? ¿Por qué?

De otro modo, a partir de los análisis de corpora, sí sería más fácil reconstruir usos más recurrentes de saludos, por ejemplo, de regiones de habla hispana. Si se trata de un curso de idiomas para alumnos que van a participar de un intercambio en Buenos Aires, el profesor puede reconstruir diálogos con usos más frecuentes de esa ciudad.

De acuerdo con Römer:

> Se puede asumir con toda seguridad que los aprendices podrán desarrollar con facilidad tanto las destrezas receptivas como las productivas cuando se enfrentan con los elementos léxicos más comunes del lenguaje y con los patrones y significados típicos, que cuando lo que se les enseña le da prioridad a palabras y estructuras poco frecuentes que raramente encontrarán en situaciones de la vida real (p. 114 apud ZAPATA, 2015, p.).

Sin embargo, se puede defender que estas conclusiones son más adecuadas para gramáticas (especialmente gramáticas descriptivas) y otros tipos de estudios metalingüísticos. Para las clases de lengua, serán más adecuados los textos auténticos. Además de los usos y estructuras, los textos auténti-

cos estarán cargados de características relativas a su tipo y género textual, fundamentales para su comprensión e interpretación.

TIPOS DE CORPORA

El primer corpus digital (University Standard Corpus of Present-Day American English), conocido como Corpus Brown, tiene como fecha de presentación el año 1964 y contiene muestras del inglés americano.

De acuerdo con Sánchez 1995, "un corpus es un conjunto de datos lingüísticos que pueden pertenecer al uso oral o escrito de la lengua, sistematizados según determinados criterios extensos en amplitud y profundidad, de manera que sean representativos en la totalidad del uso lingüístico o de alguno de sus ámbitos, dispuestos de tal modo que puedan ser procesados por un ordenador, con la finalidad de propiciar resultados varios y útiles para la descripción y análisis" (Sánchez, 1995: 8-9).

Vamos a ver el cuadro siguiente algunos tipos de corpora. Obsérvalo atentamente.

Clasificación de los corpora (Torruella y Llisterri, 1999).

<table>
<tr><th colspan="2">Escritos</th><th>Orales</th></tr>
<tr><td rowspan="6">Según el porcentaje y distribución de los textos</td><td>Corpus grande: alberga una cantidad muy elevada de datos.</td><td rowspan="6">- Corpus para la descripción fonética de la lengua: los datos han sido grabados en condiciones acústicas óptimas para su posterior análisis en laboratorio.</td></tr>
<tr><td>- Corpus equilibrado: contiene variedades de textos con porcentajes de datos equilibrados para cada variedad.</td></tr>
<tr><td>- Corpus piramidal: estos corpora tienen varios niveles de textos y porcentajes de datos. A menor variedad de textos por nivel, mayor porcentaje de datos.</td></tr>
<tr><td>- Corpus monitor: contiene un número exacto de datos que se actualizan con el tiempo.</td></tr>
<tr><td>- Corpus paralelo: lo compone una colección de textos traducidos a dos o más lenguas.</td></tr>
<tr><td>- Corpus comparable: se trata de textos del mismo (género) pero escritos en varias lenguas.</td></tr>
</table>

	- *Corpus multilingüe:* colección de textos de diversos géneros en distintas lenguas.	
	- *Corpus oportunista:* colección de textos disponibles sin criterios de selección.	
Según la especificidad de los textos	- *Corpus general:* incluye una variedad muy amplia de géneros discursivos.	**- Corpus para el desarrollo de sistema en el ámbito de las tecnologías del habla:** contienen "…inventarios grabados de unidades de síntesis a partir de los cuales se realiza el paso de una representación ortográfica a una onda sonora…" (p. 14).
	- *Corpus especializado:* incluye solamente textos especializados de un solo tipo, como los poéticos.	
	- *Corpus genérico:* incluye únicamente textos de un solo género, incluyendo los que no son especializados.	
	- *Corpus canónico:* colección de textos de un mismo autor, independientemente del género, como por ejemplo: todas las obras de Shakespeare.	
	- *Corpus cronológico:* incluye datos en secuencia de una época en particular.	
	- *Corpus diacrónico:* los datos en este corpus contienen textos de la misma lengua en diferentes etapas del tiempo con el propósito de observar su evolución.	
Según la cantidad de texto que se recoge de cada documento	- *Corpus textual:* recoge íntegramente todos los textos existentes del tipo de documento que lo constituye.	**- Corpus de transcripciones ortográficas de la lengua hablada:** contienen grabaciones de conversaciones espontáneas o de los medios de comunicación, incluyéndose también (…) discursos políticos, clases, sermones, etc." (p.14).
	- *Corpus de referencia:* a diferencia del textual, este sólo incluye fragmentos de textos de los documentos que lo constituyen.	
	- *Corpus léxico:* recoge sólo fragmentos de textos de igual tamaño, por lo general,	

	muy pequeños (más pequeños que los que se incluyen en el de referencia).	
Según la codificación y la anotación	- *Corpus simple:* corpus cuyos textos no tienen codificación ni etiquetaje. Los textos están en formato neutros.	
	- *Corpus anotado:* textos que incluyen codificación y etiquetaje.	
Según la documentación que acompaña a los textos	- *Corpus documentado:* corpus cuyos textos incluyen datos de identificación (cabeceras).	
	- *Corpus no documentado:* los textos no incluyen datos de afiliación o cabecera.	

Son muchas las posibilidades para la construcción de un corpus. De acuerdo con el cuadro, podemos observar que los corpora pueden ser orales o escritos y pueden ser clasificados de acuerdo con el porcentaje y la distribución de los textos, la especificidad de los textos, la cantidad de texto de un documento, su codificación y anotación y también pueden ser clasificados según la documentación que acompaña los textos. Además, es importante observar que el tipo de corpus determina el tipo de investigación relacionada (estudios diacrónicos, fonéticos, ortográficos, discursivo, entre otros.

A continuación, vamos a ver algunos términos básicos de la Lingüística de Corpus. ¿Me sigues?

TÉRMINOS IMPORTANTES

De modo bastante sencillo, podemos entender los términos colocación, coligación, repetición, frecuencia y combinación de la siguiente manera.

- Colocación: asociación entre ítems lexicales o entre léxico y campos semánticos.
- Coligación: asociación entre ítems lexicales y gramaticales.
- Repetición: coocurrencia de una palabra.
- Frecuencia: información sobre el número de repeticiones de una palabra o expresión en determinado contexto.
- Combinación léxica: elección, de acuerdo con algunos criterios, entre las distintas posibilidades de uso de una palabra o expresión.

Se puede observar que colocación y coligación poseen definiciones semejantes. La diferencia, por lo tanto, es que en la coligación hay coocurrencia entre categorías o patrones gramaticales e ítems lexicales (motivación gramatical): adaptarse a, reflexionar sobre, enamorarse de, entre otros. En la colocación la selección lexical tiene motivaciones semánticas (vino tinto, gravemente herido, refrescar la memoria, entre otros).

En la imagen siguiente podemos ver informaciones sobre los ítems lexicales de esta clase.

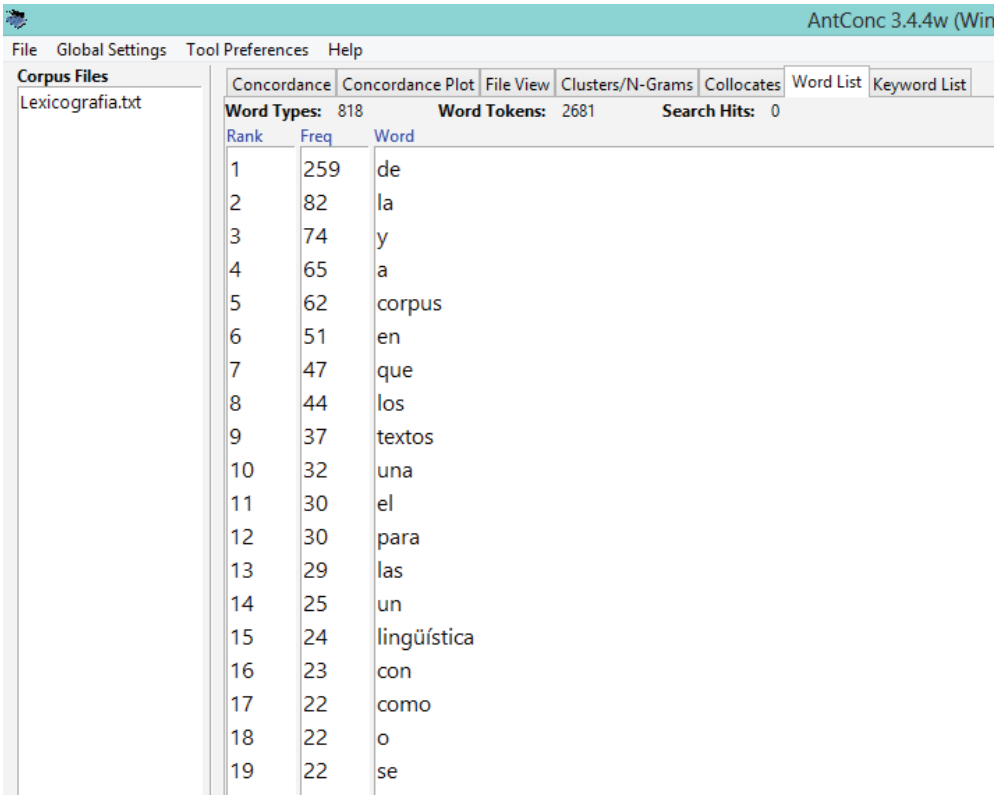

Imagen 02– AntConc (imagen del autor).

AntConc es un programa de computador que nos permite verificar la frecuencia de una determinada palabra en un documento, su colocación, el número de apariciones, entre otras informaciones.

En la imagen 02, podemos verificar que en esta clase una de las palabras que aparecen más veces es "textos". Tal palabra fue escrita treinta y siete veces y fue utilizada principalmente después de la preposición "de" y del artículo "los". Después de la palabra "textos", aparece tres veces (resultado más frecuente) la palabra "auténticos", explicitando cierta insistencia metodológica por el uso de ese tipo de texto.

Podemos observar que las preposiciones y artículos aparecen entre las unidades más frecuentes. La preposición "de" aparece en esta clase dosci-

entos cincuenta y nueve veces. Entre las veinte palabras más frecuentes, aparecen apenas tres sustantivos.

Sigue verificando otras informaciones en AntConc a partir de las unidades léxicas de esta clase y después vamos a las actividades, ¿sí?

ACTIVIDAD

1. ¿Qué es la Lingüística de Corpus? ¿Qué aporta a la enseñanza de lenguas extranjeras?
2. ¿De qué manera softwares como AntConc pueden ser utilizados en sala de clase?
3. Selecciona un documento de tu preferencia y obsérvalo en AntConc. Luego, contesta a las siguientes preguntas:

a) ¿Cuáles son las dos palabras que más aparecen?
b) ¿Cuáles son las concordancias más frecuentes de estas dos palabras?
c) ¿Qué aspectos te parecieron interesantes en este rápido análisis?

RESUMEN

En esta última clase, revisamos aspectos introductorios de la Lingüística de Corpus. Estudiamos especialmente sus objetivos, los tipos y definiciones de corpus y sus aportaciones a las clases de lenguas extranjeras.

El reconocimiento de patrones y usos lingüísticos representa una importante contribución a la enseñanza de lenguas extranjeras. A partir de los corpora, autores de materiales didácticos (manuales, diccionarios, entre otros) tienen mejores condiciones de seleccionar textos e informaciones más relevantes para los alumnos. Del mismo modo, profesores de lenguas extranjeras tienen en la Lingüística de Corpus una valiosa herramienta de análisis lingüístico.

De acuerdo con Sánchez (1995), adoptamos como definición de corpus "un conjunto de datos lingüísticos que pueden pertenecer al uso oral o escrito de la lengua, sistematizados según determinados criterios extensos en amplitud y profundidad, de manera que sean representativos en la totalidad del uso lingüístico o de alguno de sus ámbitos, dispuestos de tal modo que puedan ser procesados por un ordenador, con la finalidad de propiciar resultados varios y útiles para la descripción y análisis" (Sánchez, 1995: 8-9).

A partir de Torruella y Llisterri (1999), visualizamos las diversas posibilidades de la elaboración de corpus: de acuerdo con el porcentaje y la distribución de los textos, la especificidad de los textos, la cantidad de texto

de un documento, su codificación y anotación y según la documentación que acompaña los textos.

Por fin, estudiamos algunos conceptos importantes para los estudios iniciales en Lingüística de Corpus, tales como colocación (asociación entre ítems lexicales o entre léxico y campos semánticos), coligación (asociación entre ítems lexicales y gramaticales), frecuencia (información sobre el número de repeticiones de una palabra o expresión en determinado contexto), repetición (coocurrencia de una palabra) y combinaciones léxicas (elección, de acuerdo con algunos criterios, entre las distintas posibilidades de uso de una palabra o expresión).

COMENTARIOS SOBRE LAS ACTIVIDADES

Algunos autores consideran la Lingüística de Corpus una ciencia y otros la entienden como una metodología (ver ZAPATA, 2005). Para McEnery y Wilson (1996, p. 1, apud ZAPATA, 2015), la Lingüística de Corpus es más un estudio lingüístico con base en ejemplos de uso de la 'vida real' y una metodología que un aspecto del lenguaje que requiera una explicación o una descripción (como ocurre con las teorías). De este modo, a partir de la Lingüística de Corpus, es posible identificar informaciones y datos fundamentales para el análisis de textos orales y escritos y para el reconocimiento de usos y patrones de escrita y de habla.

Softwares como AntConc pueden ser muy útiles especialmente para la verificación de frecuencia de palabras y concordancia. Estas informaciones pueden ser muy relevantes para la comprensión de textos, percepción de mecanismos de cohesión textual, identificación de ideologías, entre otros aspectos.

PARA CONCLUIR

La formación de profesores de lenguas extranjeras implica el conocimiento y el reconocimiento de diversos modos de pensar y el acercamiento a otras áreas de estudio. Observa como FIORIN (2008) ejemplifica el carácter interdisciplinar de los estudios lingüísticos.

> Podemos debruçar-nos sobre as diferenças entre as línguas e então a linguística faz fronteira com a antropologia e a etnologia. Podemos ocupar-nos da variação no espaço, como fazem a dialetologia e a geolingüística, e aí a linguística acerca-se da geografia. Podemos examinar a variação de grupo social para grupo social e, nesse caso, a linguística limita-se com as teorias sociológicas. Podemos observar a variação de uma situação de comunicação para outra e

> então a linguística faz limites com a teoria da comunicação. Podemos pesquisar a mudança linguística e a evolução de uma língua ou de uma família de línguas e aí a linguística avizinha-se da história. Podemos analisar a aquisição da linguagem e aí, dependendo da posição teórica com que se faz a análise, a linguística confina com a biologia ou a antropologia. Podemos ver a linguagem como um sistema formal e então a linguística se aproxima da matemática e da computação. Podemos investigar as unidades maiores do que a frase, isto é, o discurso e o texto. Nesse caso, quando se põe acento na dimensão linguística, os estudos do discurso têm vizinhança com a retórica, com a dialética, com a teoria da literatura. Quando se enfatiza a dimensão histórica do discurso, a análise do discurso é limítrofe da história (FIORIN, 2008, p. 01).

Del mismo modo, a partir de la Lingüística de Corpus, son posibles variados tipos de trabajos: elaboración de diccionarios y de traductores en línea, creación de nubes de palabras, análisis lingüístico y sociolingüísticos, incluso la verificación de hashtags en internet. De esta manera, se espera que los profesores de lenguas extranjeras busquen siempre una formación actualizada y no apenas conozcan las nuevas contribuciones de las ciencias lingüísticas, sino también las utilicen y potencialicen sus prácticas docentes.

SUGERENCIA DE ACTIVIDAD

Te sugerimos la lectura del texto La utilización de la lingüística de corpus en la enseñanza de español: usos de pero y sino, de Iandra Maria da Silva, disponible en nuestro Ambiente Virtual de Aprendizaje.

AUTOEVALUACIÓN

¿Comprendo qué es Lingüística de Corpus, sus características y sus términos básicos? (Sí/ No)
¿Reconozco la relación entre Lingüística de Corpus y enseñanza de lenguas extranjeras? (Sí/ No)
¿Hice las actividades y busqué orientaciones con mis tutores sobre las dudas? (Sí/ No)
¿Busqué otros textos para complementar mis estudios? (Sí/ No)

REFERENCIAS

BERBER SARDINHA, Tony. **Linguística de corpus**. Barueri: Manole, 2004.

FIORIN, José Luiz. Linguagem e interdisciplinaridade. **Alea**, Rio de Janeiro , v. 10, n. 1, p. 29-53, June 2008. Disponible en: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1517 106X2008000100003>. Fecha de acceso: 02 oct. 2018.

ROBLES GARROTE, Pilar. Aportaciones de la Lingüística de Corpus al estudio de la conferencia como género académico de divulgación científica / Corpus Linguistics contributions to the study of conference presentations as an academic discourse. **CHIMERA: Romance Corpora and Linguistic Studies**, [S.l.], v. 3, n. 1, p. 1-21, apr. 2016. ISSN 2386-2629. Disponible en: <https://revistas.uam.es/index.php/chimera/article/view/2282/4527>. Fecha de acceso: 10 oct. 2018.

SILVA. I. M. **La utilización de la lingüística de corpus en la enseñanza de español**: usos de pero y sino. Disponible en: <https://cvc.cervantes.es/ensenanza/biblioteca_ele/publicaciones_centros/PDF/rio_2006/26_dasilva.pdf>. Fecha de acceso 12 nov. 2018.

SINCLAIR, J. **Corpus Concordance Collocation**. Oxford: Oxford University Press, 1991.

ZAPATA, Chinger. 3. La Lingüística de Corpus (lc) y su Incidencia en la Enseñanza de Lenguas Extranjeras. **EDUCARE**, [S.l.], v. 19, n. 2, p. 53-75, jun. 2016. ISSN 2244-7296. Disponible en: <http://revistas.upel.edu.ve/index.php/educare/article/view/3680>. Fecha de acceso: 15 dic. 2018.