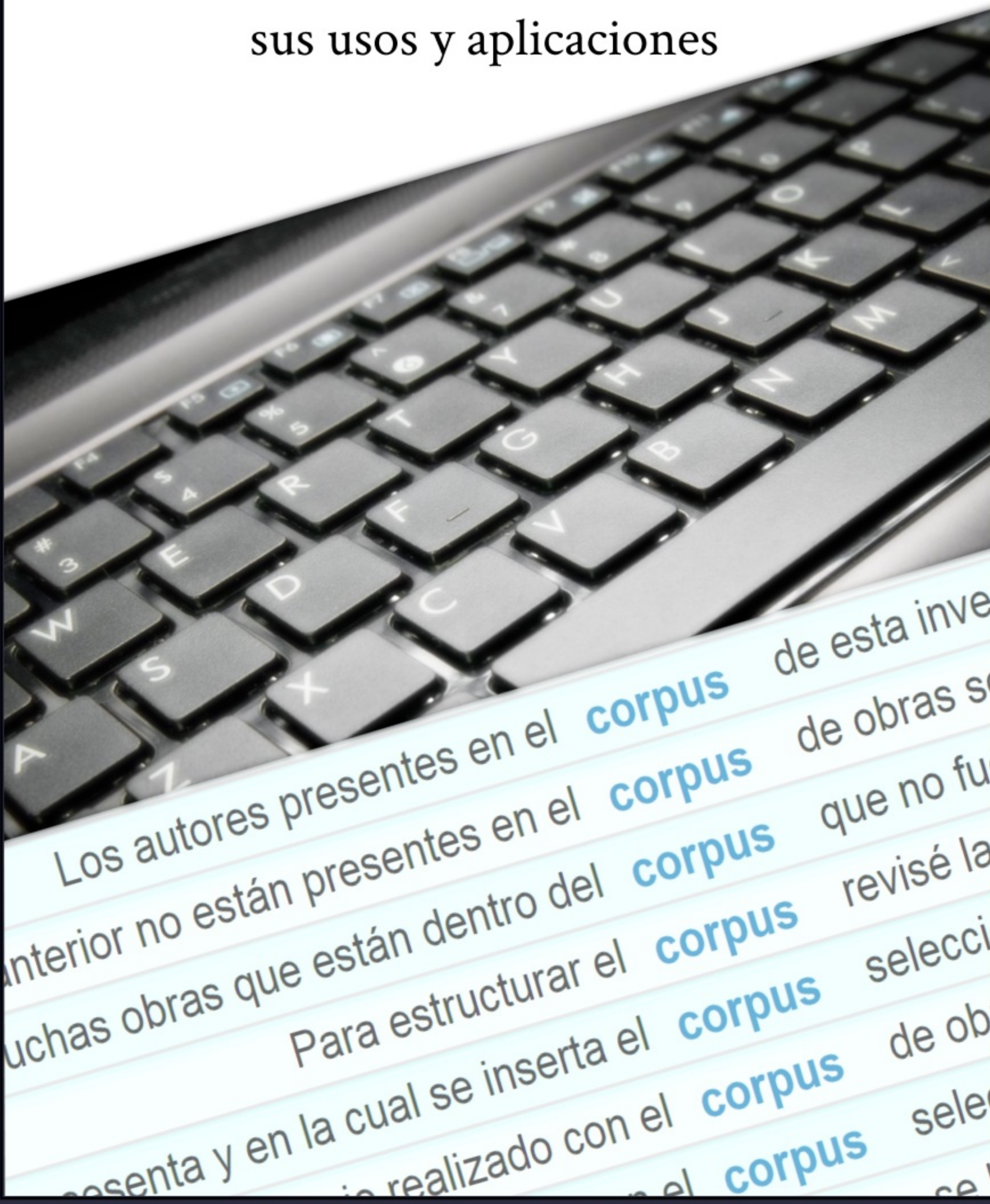
Introducción a la Lingüística de Corpus

sus usos y aplicaciones



Paula J. Liendo, Micaela J. González, Stella M. Maluenda,
Norma A. Maure, Silverio Ortiz, Leticia N. Pisani,
Romina N. Sánchez.

PIN J031 - (2018-2022)

Introducción a la Lingüística de Corpus, sus usos y aplicaciones /
Paula Josefina Liendo ... [et al.] ; dirigido por Paula Josefina Liendo.
- 1a ed. - General Roca : Paula Josefina Liendo, 2022.
Libro digital, PDF

Archivo Digital: descarga y online
ISBN 978-987-88-3921-9

1. Lingüística. 2. Educación Superior. 3. Traducción.
I. Liendo, Paula Josefina, dir.
CDD 410.1

Autores: Paula J. Liendo, Micaela J. González, Stella M. Maluenda, Norma A. Maure, Silverio Ortiz, Leticia N. Pisani, Romina N. Sánchez.

Coordinación: Paula J. Liendo

Asesoramiento editorial: Ma. Palmira Massi

Diseño: Silverio Ortiz

E-book publicado en Biblioteca Digital de Lenguas. (Identificador:

https://bibliotecadelenguas.uncoma.edu.ar/items/show/608)

Realizado en el marco del PIN J031: *Alfabetización académica y tipologías textuales en la enseñanza del inglés para la traducción*, 2018-2022.

Facultad de Lenguas. Universidad Nacional del Comahue, diciembre de 2021.
Av. Mendoza y Perú. Código Postal (8332). General Roca, Río Negro, Argentina.

ISBN 978-987-88-3921-9

9 789878 839219

Índice

Agradecimiento

Nuestro agradecimiento a las colegas y estudiantes que participan o han participado en el proyecto de investigación —PIN I J031, *Alfabetización Académica y Tipologías Textuales en la Enseñanza del Inglés para la Traducción*— que da marco y origen a este manual. Y gracias a nuestras y nuestros estudiantes, en quienes pensamos como receptores y beneficiarios últimos de esta y todos las actividades que realizamos, y a quienes dedicamos este trabajo.

1. Prólogo

Este manual tiene como principal objetivo lograr una aproximación a las distintas definiciones de corpus y de la Lingüística de Corpus (LC), así como ofrecer guías prácticas para su uso. Se produce en el marco de las investigaciones del proyecto PIN I J031, *Alfabetización Académica y Tipologías Textuales en la Enseñanza del Inglés para la Traducción* (2018-2022), Traductorado Público en Idioma Inglés, Universidad Nacional del Comahue, Argentina.

En el transcurso del mencionado proyecto, hemos leído y conversado extensamente sobre la centralidad de los corpus textuales en el ámbito académico en la actualidad, sobre todo gracias a los avances tecnológicos de los últimos veinte años. Y mediante esta práctica, hemos descubierto que estas herramientas son mucho más que instrumentos para reunir datos; se han constituido como ejes temáticos, teóricos y metodológicos de investigaciones de los más variados campos y de corte eminentemente transversal e interdisciplinario.

No obstante, podría decirse que la iniciativa de diseñar esta publicación surge casi por casualidad. En las reuniones de nuestro equipo de trabajo —en un tiempo presenciales, hoy aún virtuales— los intercambios sobre cuestiones académicas de forma y de fondo con respecto al curso de los avances en la investigación van siempre acompañados por el comentario de situaciones relativas a nuestro quehacer en las aulas: la relevancia de ciertos conceptos para la didáctica de una materia, la importancia del desarrollo de alguna estrategia en la labor traductora, la posible aplicación de una herramienta a la enseñanza, por citar solo algunos ejemplos.

Es así que pensamos que podría ser interesante compartir nuestros hallazgos sobre la LC y sus principales usos con la comunidad académica de nuestra universidad, en particular, y hacer extensiva la invitación a miembros de otras instituciones educativas relacionadas con el aprendizaje de lenguas segundas o extranjeras. El punto de partida para la redacción del manual fue la realización de una encuesta. El análisis de las respuestas obtenidas por ese medio nos permitió observar que, si bien la mayoría de las personas encuestadas sabe qué es o para qué se utiliza un corpus de textos (72,9%), menos de un treinta por ciento recuerda fehacientemente que se haya utilizado un corpus en sus clases de idioma extranjero —poco menos que un tercio de las y los docentes, y alrededor de un cuarto de las y los estudiantes—. También es interesante destacar que la mayor parte de las y los estudiantes consultados refirieron a usos prácticos de los corpus en las clases, como resolución de dudas terminológicas o de colocaciones, mientras que entre las ventajas de su utilización señalaron el desarrollo de estrategias, como agilizar búsquedas, identificar géneros o incorporar vocabulario; y competencias, por ejemplo, la traductora, y sus subcompetencias.

Los resultados de la encuesta nos alentaron a pensar en este manual como una herramienta útil para docentes y estudiantes, que permita comprender qué es un corpus, cómo se vincula con el aprendizaje de lenguas extranjeras, y qué usos se les puede dar en las clases. Y con el mismo espíritu de construcción colectiva del conocimiento, los invitamos a volcar sus opiniones en un cuestionario, luego de leer o utilizar este manual. Esperamos sus respuestas y sugerencias para seguir generando otras formas de divulgación y transferencia.

El objetivo último de este manual es proporcionar un andamiaje de los aspectos principales de la LC y sus aplicaciones áulicas para el aprendizaje de las lenguas y la didáctica de la traducción. A tal fin, proporcionaremos algunas definiciones de LC, y su relación con la educación, en particular en la enseñanza de lenguas extranjeras y la formación de traductores (§ 2). Luego compartiremos algunas definiciones de corpus, diseño de corpus, clasificaciones y tipologías textuales, y su vinculación con los géneros textuales (§ 3). A continuación, incluiremos un listado de corpus en línea (en inglés y español) y algunos ejemplos de búsqueda (§ 4), así como una guía práctica para anotar y analizar corpus propios según el paradigma de investigación (§ 5). La lista de referencias al final de cada sección permitirá ahondar en los temas que resulten de interés.

Nos despedimos con el deseo de que este manual resulte relevante para su práctica y formación, y a la espera de sus opiniones.

2. ¿Qué es la Lingüística de Corpus?

Tanto en el ámbito de la lingüística como en otros ámbitos académicos, los avances y las investigaciones interdisciplinarias traen aparejados nuevos *desafíos*: conflictos no resueltos acerca de la terminología a utilizar, múltiples definiciones de conceptos, diversos enfoques similares —pero no exactamente iguales— que coexisten, perspectivas más teóricas que se contraponen a otras más aplicadas. Esta es una situación en la que todos nos encontramos alguna vez, al leer un texto o intentar comprender un concepto de un área de conocimientos nueva para nosotros.

En este sentido, el concepto de Lingüística de Corpus (LC) no es nuevo, pero algunos aspectos son aún controvertidos. En líneas generales, es posible definir a la LC como un *enfoque* para el *estudio de las lenguas* que consiste en la *recopilación y procesamiento de corpus lingüísticos*. Esta forma de trabajo permite utilizar *datos reales y exhaustivos* que reflejan la lengua viva, así como observar el conocimiento lingüístico de los hablantes. En otras palabras, nos posibilita el acceso a producciones lingüísticas reales de quienes usan una lengua en situaciones concretas para estudiarlas. El *análisis* de la información recolectada ofrece *resultados* fehacientes que favorecen el *desarrollo del conocimiento científico.*

2.1 ¿Qué beneficios aporta la LC en el ámbito del aprendizaje de las lenguas?

Entre los aportes más interesantes de la LC a la enseñanza-aprendizaje de lenguas, podemos señalar:

- que se centra en el estudio de las producciones lingüísticas de los hablantes en una *situación concreta*; por lo tanto, ofrece valiosas oportunidades para analizar cómo los alumnos utilizan su lengua o sus lenguas;
- que brinda una base empírica para el desarrollo de materiales educativos y metodológicos de diversa índole, así como para la construcción de gramáticas, tesauros, y diccionarios;
- que ofrece ejemplos más naturales que los que proporcionan muchas veces los libros de textos y las bases de recursos didácticos, porque son una muestra del uso real de la lengua y el bagaje cognitivo de sus hablantes;
- que estos ejemplos permiten, en términos chomskianos, la observación y el análisis de la actuación lingüística (*performance*), en contraposición con la competencia lingüística (*competence*), y así es posible desarrollar teorías lingüísticas sobre el funcionamiento y la adquisición del lenguaje (ver también § 2.3);
- que facilita la descripción, análisis y enseñanza de los distintos tipos de discursos, tanto generales como especializados, orales y escritos;
- que sus aportes han favorecido una mayor centralidad del léxico en la enseñanza-aprendizaje de lenguas extranjeras, en particular en lo que respecta a las colocaciones y las expresiones idiomáticas o «unidades prefabricadas»;
- que todo lo señalado anteriormente ha influido en la evaluación, ya que se observa un mayor énfasis en las estructuras gramaticales, los giros idiomáticos y las palabras clave en contexto en el diseño de pruebas y ejercicios, su input textual y los criterios de valoración utilizados.

2.2 Lingüística de corpus: ¿un marco teórico o metodológico?

El uso de la LC como fuente de evidencias es perfectamente compatible con otras teorías. No obstante, para algunos renombrados lingüistas (Parodi, 2008; Leech, 1992, entre otros) es relevante definir a la LC como una teoría o como una metodología.

Si consideramos que los corpus se usan como *modelos* de uso; es decir, que son representativos de los usos reales de las lenguas en una comunidad determinada, debemos coincidir con Pérez Paredes (2021: 1): «A corpus is used to *model* usage and we can think of a corpus as a *proxy* for usage»[1]. Si pensamos en un corpus como en un *agente* o *representante* del uso real, este se convierte en un *instrumento*, un *método* utilizado por los investigadores para buscar la respuesta a su pregunta.

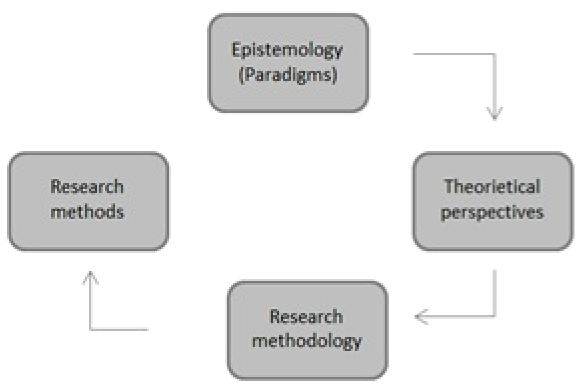
Dado este *enfoque empíric*o de la LC, Parodi (2008), por ejemplo, entiende que la LC no es otra rama de la lingüística, como lo son la sintaxis o la semántica, sino un método de investigación aplicable a otras disciplinas desde enfoques teóricos diversos, que es flexible (porque se basa en datos originales y completos, como unidades de sentido y con propósitos comunicativos específicos), si bien cuenta con principios reguladores bien definidos.

Por otra parte, para Leech (1992), la LC no es un campo de estudio, sino un *área disciplinar* determinada por la centralidad de los corpus que posee metodologías específicas resultantes de la integración de los avances tecnológicos y de ciertas categorizaciones. Otros autores, como Sinclair (1991) y Simpson y Swales (2001), sostienen que la LC es una técnica o una *tecnología*, que tiene al corpus mismo como fundamento y depende principalmente de una construcción adecuada, para que el resultado sea una base de datos representativa.

Una postura conciliadora de estas posiciones es la de Pérez Paredes (2021). Este autor explica que la gran variedad de enfoques y paradigmas utilizados en las investigaciones sobre educación devienen en inconsistencias terminológicas, que resultan de las diferentes perspectivas sobre la realidad y las epistemologías utilizadas. Es por esto que destaca la importancia de vincular las metodologías de investigación con los supuestos filosóficos a los que se adhiere. En general, las investigaciones definen, por un lado, la epistemología y las perspectivas teóricas; por otro los métodos y la metodología. Sin embargo, el uso de métodos distintos, concluye Pérez Paredes, proporcionará explicaciones de la realidad diferentes, ya que estos dependen de metodologías, perspectivas teóricas y paradigmas investigativos diversos. Este autor resume la relación etiológica —o causativa— entre epistemología, teoría, metodología y método, tal como se describe en la Figura 1.

Figura 1

Figure 1.3 *Research in education – Based on Pring (2004) and Gray's (2004)'s adaptation of the work of Crotty* (1998) (Pérez Paredes, 2021: 8)



En este sentido, las y los investigadores que adhieren a un *paradigma cientificista*, que considera que existe una realidad objetiva —independiente, constituida por objetos que interactúan entre sí— desarrollarán teorías, metodologías y métodos diferentes de quienes siguen el *paradigma constructivista* o *fenomenológico*, que sostiene que vivimos en un mundo de ideas, y que son estas las que eventualmente construyen nuestra propia realidad —y no hay una realidad exterior por descubrir sino tantas como investigadores existan.

Teoría o método, tecnología o área disciplinar, no caben dudas de que la LC es un área de investigación reciente, cuyo uso se ha extendido significativamente en las últimas décadas. A modo de resumen, podemos señalar que:

- la LC tiene un carácter eminentemente interdisciplinar ya que el estudio de una lengua en su contexto implica una postura cognitivista y socioconstructivista;
- no existe aún una clara definición de sus procedimientos;
- es probable que los avances en las tecnologías de la información y la comunicación favorezcan la disponibilidad de datos que resulten en más y mejores investigaciones en el área.

Es importante destacar, también, que el uso más extendido de la LC en las últimas décadas se aplica al análisis de lenguajes especializados, es decir a los usos de la lengua enmarcados en comunidades de usuarios en las que tienen un cierto significado. Es aquí donde se encuentra el nexo fundamental entre la LC y el análisis del discurso, en particular la Teoría de géneros discursivos (ver también §3.1).

2.3 Un poco de historia: lingüística, LC y enseñanza de las lenguas

A partir del surgimiento de la Lingüística Generativa (LG) en la década de 1950, y su posterior influencia hegemónica, todas las demás posturas que no sostuvieron una visión idealizada de la lengua o metodologías de índole hipotético-deductivo se vieron debilitadas. Es decir, el giro generativista y racionalista dejó de lado el empirismo, y la LC tradicional —enmarcada en paradigmas socioculturales y contextualistas— se vio opacada por esta nueva corriente.

Si bien es innegable que la LG ha realizado aportes fundamentales en materia del desarrollo y adquisición del lenguaje humano, también es cierto que la preponderancia de esta línea hizo que la lengua, como objeto de estudio de las investigaciones, fuera casi exclusivamente concebida con una visión idealizada, focalizada —en términos chomskianos— en la *competencia lingüística (linguistic competence)* en desmedro de la actuación (*performance*) o lenguaje en uso. Esta última era, para la LG, demasiado cambiante e impredecible para ser un objeto de ciencia adecuado.

A comienzos de la década de 1960, se observa un resurgimiento de los estudios basados en corpus, que resultó principalmente de los avances en las tecnologías computacionales y un renovado interés en los usos de las lenguas naturales y cotidianas y su variabilidad en distintos registros y géneros. Al mismo tiempo, surgen en Inglaterra y los países escandinavos importantes proyectos de investigación abocados a la construcción de grandes corpus lingüísticos digitales en idioma inglés.

Alrededor de veinte años después, en la década de 1980, se observa un segundo momento en la LC, y se identifica con la creación de *mega-corpus* de cientos de millones de palabras. También se puede detectar un tercer giro, probablemente aún en curso, que emerge del interés por el estudio de discursos especializados. Estos constituyen normalmente muestras más pequeñas, si se las compara con los corpus generales, pero de mayor focalización temática, estructural y funcional.

En el marco de la Lingüística aplicada a la enseñanza de las lenguas, el resurgimiento en el interés en la LC es importante en función del diseño de enfoques didácticos. Algunos especialistas en pedagogía de las lenguas como Stern (1983, 1992) o Richards y Rodgers (1986) plantean que para diseñar un proyecto didáctico es necesario hacer explícito qué entendemos por una *lengua* —cómo la definimos, qué significa ser un usuario experto, cuáles son las unidades básicas de su estructura—, así como establecer qué significa enseñarla o aprenderla. De tal modo, los enfoques didácticos de los últimos cuarenta años tienen una orientación mayormente comunicativa o socioconstructivista, por lo que parten de una concepción del lenguaje como un fenómeno social, con un propósito comunicativo.

En este contexto, se puede visualizar la importancia de la LC en el diseño curricular para la enseñanza de lenguas. Es decir, un usuario experto de una lengua necesitará poseer conocimientos lingüísticos *sistémicos, discursivos* y *socioculturales*; y la enseñanza-aprendizaje de dicha lengua será un proceso en el que las y los estudiantes elaboran sus propias teorías lingüísticas según el *input* con el que se enfrentan y la negociación de las correcciones que reciben, de acuerdo con el *enfoque comunicativo* (Canale & Swain,1980; Swain, 1985; Krashen, 1987, 1988). El uso de corpus deviene en un contacto con datos sobre el uso real de la lengua que favorece el aprendizaje como construcción social, y prepara al estudiante para desempeñarse con eficacia en un contexto académico o profesional determinado.

2.4 ¿Cómo impacta en la educación superior? La enseñanza del idioma inglés como lengua extranjera (ILE) y la formación de traductores

2.4.1. El corpus en la enseñanza de ILE

En la década de los ochenta el uso de corpus se consolidó como materia prima para la descripción y análisis del uso real de la lengua y se utilizó como insumo para la elaboración de gramáticas y diccionarios bajo el enfoque denominado COBUILD (*Collins Birmingham University International Language Database*). Esto tiene estrecha relación con el enfoque para la enseñanza de lenguas asistida por computadora, conocida como *computer-assisted language learning* (CALL).

Los primeros antecedentes de la enseñanza de idiomas y el uso de corpus son:

- el Enfoque Léxico (*Lexical Approach*), propuesto por Sinclair y Renouf; y
- el enfoque conocido como *Data-Driven Learning* (DDL por sus siglas en inglés) de Tim Johns, quien incorporó el corpus y el análisis de concordancias para la enseñanza del inglés como lengua extranjera, en el que se inspiró el equipo COBUILD (Corpas Pastor, 2012).

Si bien ambos enfoques tenían como objetivo que las y los estudiantes accedan a un entorno real de uso de la lengua, se diferencian en que el Enfoque Léxico utiliza los datos provenientes del corpus para la investigación y el diseño curricular; mientras que el DDL introduce el corpus en el aula y promueve el descubrimiento autónomo a partir de los datos reales. Para ello, utiliza dos tipos de procedimientos::

- inductivos: o aprendizaje por descubrimiento, en el que las y los estudiantes observan, clasifican, generalizan y construyen hipótesis;
- y deductivos: en el que se basan en generalizaciones previas para después clasificar los datos, contrastarlos con las reglas aprendidas y finalmente consolidar y precisar aún más los conocimientos adquiridos.

2.4.2. El corpus en la formación de traductores

En lo que respecta a la enseñanza de la traducción, el uso de corpus tiene una doble relevancia: como herramienta para la enseñanza y el aprendizaje, tal como mencionamos anteriormente; y como una herramienta utilizada por las y los traductores profesionales para las distintas etapas del proceso de trabajo, a saber:

- la obtención del texto base;
- el análisis y la preparación del texto base;
- la traducción propiamente dicha;
- la corrección y optimización;
- y finalmente la entrega (Martín-Mor y otros, 2014).

De hecho, el ámbito de la traducción ha sido pionero al incorporar las tecnologías como su herramienta principal de trabajo, ya que, a medida que han evolucionado las tecnologías de la información y de la comunicación, también lo ha hecho el campo laboral de estos profesionales de la comunicación. Es más, las tecnologías aplicadas a la traducción han cobrado tal relevancia que se ha desarrollado un área específica en torno a esta temática: la *informática aplicada a la traducción*.

A su vez, los avances en los estudios en el área de la didáctica de la traducción indican claramente cuáles son las competencias que se deben desarrollar para adquirir la *macrocompetencia traductor*a, que incluye a las siguientes competencias:

- la *competencia lingüístic*a, que refiere al conocimiento profundo de las lenguas de trabajo;
- la *competencia extralingüístic*a, que abarca el conocimiento sobre aspectos culturales y temáticos;
- la *competencia instrumental o profesiona*l, que integra los saberes sobre el campo laboral, el uso de herramientas, la habilidad para documentarse;
- la *habilidad de transferencia*, que refiere las etapas de comprensión de un texto base y su correcta reexpresión en una lengua meta; y

- la *competencia estratégica*, la capacidad de poder identificar y resolver problemas que surjan durante todo proceso de trabajo.

Tanto la competencia estratégica como la competencia instrumental y la habilidad de transferencia son específicas de la competencia traductora (PACTE, 2005). Es decir, son las que diferencian a traductores profesionales de hablantes bilingües, por ejemplo. Por tanto, en general, estos son los conocimientos y destrezas que se abordan en las asignaturas específicas de traducción.

Dentro de las habilidades que incluye la *competencia instrumental*, se encuentra la *competencia documental o competencia informacional*, que consiste en la identificación de un problema y la correcta solución a partir de fuentes de documentación, principalmente electrónicas.

Aquí el uso de corpus, particularmente en la traducción especializada, adquiere un rol central, y su finalidad varía según la necesidad. Las opciones pueden ser diversas, de un corpus se pueden extraer terminología específica, fraseología, colocaciones, patrones sintagmáticos (léxico-gramaticales), equivalentes, entre otros ejemplos. A su vez, el uso de corpus paralelos y comparables en el aula de traducción tiene otras aplicaciones: por un lado, permite a las y los estudiantes analizar las técnicas y estrategias utilizadas por profesionales, y por otro, acceder a textos auténticos escritos por especialistas.

2.4.2.1. Estudios de Traducción con corpus

En lo que concierne a la investigación, existe una línea descriptiva, llamada Estudios de Traducción con corpus, cuya metodología posibilita el estudio detallado de la lengua traducida y compararla con corpus equiparables escritos en lengua de origen. De esta manera, es posible observar el comportamiento de traductoras y traductores y evidenciar cuáles son las tendencias en la producción de sus textos, establecer universales de la traducción, estudiar normas de traducción y caracterizar distintos estilos traductores.

A modo de conclusión, podemos afirmar que existe un abanico de posibilidades para la aplicación de los corpus en el ámbito de la enseñanza-aprendizaje de las lenguas así como en la investigación y en el ámbito profesional relacionados con ellas. Cualquiera sea el caso, es necesario el desarrollo de un alto grado de competencia tecnológica. En tal sentido, tanto estudiantes como docentes de traducción nos vemos desafiados ante una realidad que implica una actualización constante en lo que respecta a recursos electrónicos, herramientas y métodos de trabajo.

Referencias

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing, *Applied Linguistics, 1* (1), 1-47.

Corpas Pastor, G. (2012). Corpus, tecnología y traducción. En Casas Gómez, M. & García Antuña, M. (Coords.), *XII Jornadas de Lingüística, 2,* 75-98.

Gray, D.E. (2004). *Doing Research in the Real World.* Sage Publications Limited.

Krashen, S. D. (1987). *Principles and Practice in Second Language Acquisition.* Prentice-Hall International.

Krashen, S.D. (1988). *Second Language Acquisition and Second Language Learning.* Prentice-Hall International.

Leech, G. (1992). Corpora and theories of linguistic performance. En Svartvik, J. (Ed.), *Directions in Corpus Linguistics* (pág. 105-122). Mouton de Gruyter.

Martín Mor, A., Piqué Huerta,, R. y Sánchez-Gijón, P. (2014). Cambios en el paradigma de la traducción especializada. *IX Simposio sobre traducción, terminología e interpretación.* La Habana, (1-10).

Parodi, G. (2008). Lingüística de Corpus. Una Introducción al Ámbito, *RLA. Revista de Lingüística Teórica y Aplicada, 46* (1), 93-119.

Pérez Paredes, P. (2021). *Corpus Linguistics for Education. A Guide for Research.* Routledge.

Richards, J. C., & Rodgers, T. S. (1986). *Approaches and methods in language teaching.* Cambridge University Press.

Pring, R. (2004). *The Philosophy of Education.* Bloomsbury.

Simpson, R. & Swales, J. (1991). *Corpus Linguistics in North America.* The University of Michigan Press

Sinclair, J. (Ed.) (1991). C*orpus, Concordance, Collocation.* Oxford University Press.

Stern, H. H. (1983). *Fundamental concepts in language teaching.* Oxford University Press.

Stern, H. H. (1992). *Issues and options in language teaching*. Oxford University Press.

Swain, M. (1985). Communicative Competence: Some roles of Comprehensible Input and Comprehensible Output in its Development. En Gass, S. & Madden, C. (Ed.), *Input in second language acquisition* (pág. 235–253). Newbury House.

---

[1] El uso de las cursivas es nuestro.

3. ¿Qué es un corpus? ¿Para qué sirve? ¿Qué tipos de corpus existen?

En las últimas décadas, muchos lingüistas han propuesto definiciones para corpus tomando en cuenta distintos aspectos. Por ejemplo, Sinclair (1991) define el corpus como: «*a collection of naturally-occurring language text, chosen to characterize a state or variety of a language*». Como vemos, esta definición no dice mucho sobre las características de los textos que componen el corpus ni sobre su almacenamiento. La lingüista especializada en discurso Mona Baker (1995) expande esa definición al considerar que un corpus es una colección de textos completos —en lugar de ejemplos u oraciones—, almacenados de forma electrónica, que se pueden analizar automática o semi-automáticamente (y no a mano). De forma similar, Francis (1982) aporta una definición que considera la naturaleza de los textos y el propósito de su recopilación: «A corpus is a collection of texts assumed to be representative of a given language, dialect, or other subset of a language to be used for linguistic analysis».

En resumen, un corpus puede definirse como una colección de textos escritos u orales —a veces completos, a veces constituída de extractos—, almacenados de forma electrónica. En conjunto, estos textos resultan representativos de una lengua, dialecto, o un aspecto de la lengua, y pueden utilizarse para un análisis lingüístico. Al ser una representación de una lengua, los corpus sirven para ejemplificar y explicar ciertos fenómenos lingüísticos; ayudan a entender cómo los hablantes usan la lengua y el lenguaje, y por qué.

Existen distintos tipos de corpus. Tomando algunas de las clasificaciones que menciona Tolchinsky (2014), podemos decir que la principal distinción se establece entre corpus escritos y corpus orales. Dentro de estos últimos, están los corpus que son *transcripcione*s del habla, y los que tienen tanto la transcripción como el registro en audio. Además, los corpus se pueden clasificar según el *porcentaje* y la *distribución* de los textos que lo componen o según la *especificidad* de los temas de dichos textos —especializados o generales—. Otros se diferencian según la *cantidad* de texto que se incluye en el corpus: algunos incluyen pocos textos completos, y otros incluyen muchos extractos de varios textos. También existen los corpus *anotados*, es decir, aquellos que no solo tienen textos o extractos, sino también etiquetas con información morfológica, sintáctica y semántica (ver también § 3.2.2, 4.2 y 5.1).

En lo que respecta al campo de la traducción, una de las clasificaciones más relevantes es la de corpus *ad hoc*, que puede definirse como una colección de textos compilados para la realización de un trabajo de traducción específico (Rodríguez-Inés, 2008). También se encuentran los corpus *paralelo*s y los *comparables* (Krüger, 2012). Los primeros son una colección de textos en la lengua meta y las traducciones de dichos textos; los segundos consisten en textos traducidos y originales en la misma lengua. Se podrían mencionar clasificaciones de corpus que toman en cuenta otros aspectos, pero no las veremos aquí.

3.1 Corpus, géneros y tipologías textuales

*Géner*o y *Tipología* son dos términos que se usan para clasificar textos en distintos grupos. Diversos lingüistas ofrecen definiciones distintas de ambos términos, y no hay un acuerdo general sobre dónde termina uno y empieza el otro. De forma sencilla, podemos decir que el *género* es una manifestación textual con propósitos comunicativos particulares y reconocidos por una comunidad discursiva (Swales 1990, 2004; Bhatia 1993, 2004). Es decir, la cantidad de géneros posibles es ilimitada; hay tantos géneros como comunidades discursivas.

Los géneros se diferencian entre sí por su propósito comunicativo, organización retórica, realizaciones léxico-gramaticales, dinamismo y evolución. Swales (1990) sugiere que los géneros varían también según otros parámetros, por ejemplo, su intención retórica, medio o modo de comunicación, y grado de atención por parte del emisor a la audiencia potencial. Bhatia (1993), por su parte, hace hincapié en la necesidad de determinar un propósito o finalidad comunicativa para poder definir un género, ya que cualquier cambio significativo en el propósito comunicativo resultará en un género diferente. Así, para Bhatia, un anuncio publicitario y una solicitud de trabajo son un mismo género, ya que ambos tienen como finalidad *promocionar*.

Muchos autores proponen otras definiciones de género que no citaremos en este manual, ya que propician más confusión que claridad. Y si en el marco de esta indeterminación terminológica intentamos definir *tipología textual*, o *tipo textual*, se nos presentan nuevos desafíos. Una definición útil puede ser la de Beaugrande y Dressler (1981), por ejemplo, quienes señalan que las tipologías textuales se conforman según las formas en las que se pueden utilizar los textos pertenecientes a un género. Estos autores mencionan una señal de tránsito, una rima infantil o un artículo periodístico como ejemplos de tipos textuales; no obstante, otros autores (Swales y Bathia, entre otros) dirían que estos son géneros.

En el ámbito del análisis del discurso y la traducción, también existen discrepancias. A la clasificación en textos *descriptivos, narrativos y argumentativos*, Beaugrande y Dressler la definen como tipos textuales, Reiss y Vermeer (1991) como *campo estilístico*, y Trosborg (1997), como *tipo de texto según la finalidad retórica*. A modo de síntesis, se puede afirmar que estos y otros autores concuerdan sobre la importancia de analizar los conceptos de *género* y *tipo de texto*, si bien la mayoría reconoce la imposibilidad de hacer una clasificación acabada de ellos, por ser categorías dinámicas que varían y evolucionan dentro de una comunidad discursiva.

Podemos concluir que, para elaborar un corpus de textos, resulta imprescindible primero adherir a una tipología o definir una que se enmarque dentro de una concepción de *géneros textuales*. Algunas características a tener en cuenta para definir un género son:

- un propósito comunicativo común;
- una comunidad discursiva determinada;
- el medio y modo de comunicación;
- un conocimiento de la audiencia/receptores;
- similitud en propósito y organización retórica;
- un carácter dinámico, y la posibilidad de evolucionar.

3.1.1 Los corpus y la alfabetización académica

La lingüística de corpus ofrece oportunidades revolucionarias para analizar cómo se utiliza la lengua, dado que se centra en el estudio de las producciones lingüísticas concretas de los hablantes en una situación en particular (Tolchinsky, 2014). Como vimos en § 2.3, el uso de corpus influye directamente en el diseño de materiales didácticos y en el desarrollo de actividades relevantes para la enseñanza de la lengua aplicada, por ejemplo, a la traducción, pero principalmente constituye un abordaje integral para la comprensión y redacción de textos en el ámbito académico. La alfabetización académica propone el desarrollo de estrategias de acceso, construcción y difusión del conocimiento en la universidad. Este enfoque resulta un eje central en la formación de grado, ya que implica alentar a docentes y estudiantes a ser partícipes activos de las prácticas discursivas contextualizadas que se rigen por la especificidad de cada disciplina (Liendo et al, 2018).

Dado que los corpus nos aportan ejemplos del uso realista de la lengua, también nos muestran las complejidades y los matices del lenguaje natural. Así es que nos pueden aclarar cuestiones sobre el uso de colocaciones, de terminología especializada, de expresiones idiomáticas y también de unidades prefabricadas, entre otros. Proponer que los alumnos realicen el análisis de los corpus constituye una grandiosa oportunidad para que formulen hipótesis de trabajo, por ejemplo: ¿en qué tipo de texto encontramos este aspecto? ¿Por qué? (Tolchinsky, 2014). Entre sus múltiples aplicaciones debemos destacar que el trabajo directo con corpus de textos propicia el aprendizaje por descubrimiento, la autonomía y la reflexión de los procesos de aprendizaje en el ámbito académico (Elvira-García, 2021).

3.2 Diseño de corpus: qué es, tipos de corpus

Como vimos en §3, el concepto de corpus es ambiguo y puede referirse a colecciones o inventarios de textos como a corpus informatizados, ya que estos se forman con piezas de una lengua que se seleccionan y ordenan según criterios explícitos, para luego utilizarse como muestras de uso de ese idioma en particular (Tolchinsky, 2014).

En la actualidad, no podemos dejar de hablar de corpus informatizados, que nos permiten recolectar una cantidad extensa de diferentes textos, orales o escritos, para así poder codificarlos y clasificarlos de manera adecuada, y luego hacer diferentes búsquedas entre las grandes cantidades de textos en formato digital que recopilemos.

Los corpus informatizados posibilitan el almacenamiento de los textos en un soporte electrónico, que principalmente permite la recuperación fácil y accesible por parte de los usuarios, así como establecer reglas para su acceso, codificación y clasificación de los tipos de textos (Marín, 1994 en Tolchinsky, 2014). La digitalización de los textos posibilita la sistematización de búsquedas, y el trabajo deja de ser artesanal. Las ventajas son varias, ya que gracias al almacenamiento en soporte electrónico es posible recopilar grandes cantidades de textos, que son accesibles para los usuarios y recuperables cuando lo necesiten (Elvira-García, 2021).

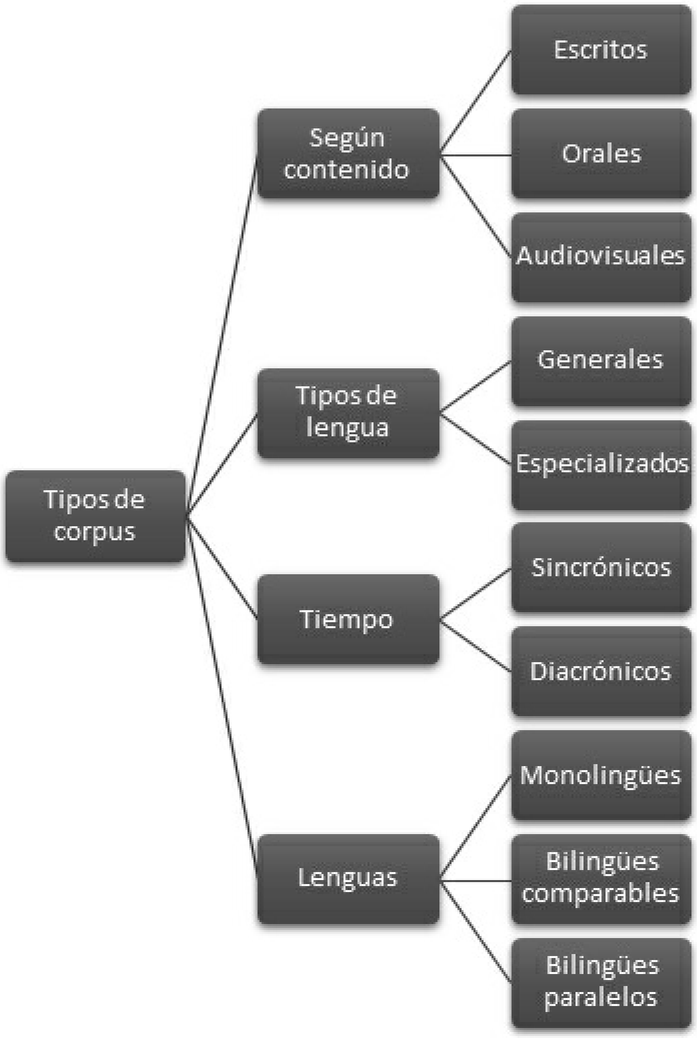
3.2.1 Tipos de corpus

Lo primero que debemos preguntarnos es qué tipo de corpus será necesario para el objetivo del estudio que llevemos a cabo. Dependiendo de nuestro propósito recopilaremos corpus de textos orales o escritos, de fragmentos de textos o textos completos, etc.

Como ilustra Elvira-García en el esquema que sigue (2021:48) (Figura 2), los corpus se pueden *clasificar* siguiendo distintos criterios. Algunos de ellos tienen en cuenta, por ejemplo, su contenido o tipos de textos; el *tipo de lengua*, ya sea general o especializada (por ejemplo lenguaje jurídico); el *tiempo*, es decir, el momento en el que esos textos fueron producidos; las *lenguas* en las que están escritos esos textos —en el caso de corpus bilingües—, etc.

Figura 2. *Tipología de Corpus*

Esquema 4. Tipología de corpus (Elvira-García, 2021:48)



Uno de los aspectos más importantes para tener en cuenta la hora de recopilar un corpus es el concepto de *representatividad*, que se entendería como la cualidad que tienen los textos de reflejar de manera fiel la realidad. Como investigadores, luego debemos definir de qué parte de esa realidad queremos que sea representativo nuestro corpus. Esto también va a determinar el tamaño de la muestra, ya que cuando creamos un corpus, no tenemos acceso a todos los textos existentes. Y además debemos decidir qué datos vamos a recolectar a partir de esos textos.

En ese sentido, resulta interesante la distinción que hace Elvira-García (2021) entre tres categorías de diseño de corpus:

- Un corpus *oportunista* es el que está diseñado para responder una pregunta concreta, para recolectar datos concretos. En este caso, el tamaño de la muestra va a ser más acotado, como por ejemplo, el corpus que recolecta una o un docente para investigar los errores que cometen sus estudiantes.
- Un corpus *monitor* es aquel que no es muy definido, sino que se va modificando de acuerdo a los datos que se quieran recolectar en un determinado momento, por ejemplo cuando se quiere tomar ejemplos reales del habla.
- Un corpus *equilibrado*, en cambio, es aquel cuya muestra será más extensa, y cuyo diseño es bien riguroso, por ejemplo en el caso de que una editorial se propone confeccionar un diccionario.

Una vez que se decidió el tipo de corpus comienza la tarea de compilar los textos.

3.2.2 Importancia de la digitalización

Como se mencionó anteriormente, en la actualidad se confecciona una gran cantidad de corpus de manera digital; y como el acceso a textos en la web es amplio y abierto, para el trabajo de recolección no va a ser necesario digitalizar los textos durante la etapa de compilación.

Una vez que recopilamos nuestros textos, el siguiente paso es organizarlos. Para ello primero es necesario realizar un *marcaje* o codificación, con el simple pero no menos importante objetivo de ordenar los textos que componen nuestro corpus, por ejemplo agruparlos en archivos teniendo en cuenta distintos criterios: autor, año, lugar, etc. Esto va a depender de las características de nuestro corpus (Elvira-García, 2021). El *marcaje* nos da información del contexto en que se ha producido ese texto. La *anotación*, en cambio, es un recurso que nos permite realizar un análisis de la lengua del mensaje en sí (Elvira-García, 2021) (ver también § 3, 4.2 y 5.1). Si trabajamos con corpus digitalizados, los resultados que podemos obtener al realizar las búsquedas nos mostrarán el uso de determinado aspecto que sea de nuestro interés.

Los corpus se pueden anotar en todo los niveles gramaticales, además es posible realizar anotaciones especializadas (ver también § 4.2). El uso de herramientas digitales como CATMA (§ 5.1) permite etiquetar los textos seleccionados que forman parte del corpus, anotando los distintos rasgos que previamente se han establecido o definido. En pocas palabras, podemos resumir que el uso de anotaciones permite explorar los resultados que arrojan, y así realizar un acercamiento más incisivo a los textos objeto de nuestro estudio. Es preciso considerar que el sistema de anotación será el que nos sea útil, y eso debe ser consensuado entre los usuarios del corpus o investigadores de antemano.

Referencias

Baker, M. (1995). Corpora in translation studies: an overview and some suggestions for future research. *Target* 7(2), 223-243.

Beaugrande, R. & Dressler, W. (1981). *Introducción a la lingüística del texto.* Ariel.

Bhatia, V. K. (1993). *Analysing Genre: Language Use in Professional Settings.* Longman

Bhatia, V. K. (2004). *Worlds of Written Discourse: A Genre-based View.* Continuum.

Elvira-García, W. (2021). *Uso de Corpus en la Clase de ELE La Lengua Real* *como modelo. Cuadernos de Didáctica*. Difusión.

Krüger, R. (2012). Working with Corpora in the Translation Classroom. *Studies in Second Language Learning and Teaching*, 2(4), 505-525. https://doi.org/10.14746/ssllt.2012.2.4.4

Liendo, P. (2017). Proyecto de Investigación Alfabetización Académica y Tipologías textuales en la enseñanza del inglés para la traducción. Facultad de Lenguas. Universidad Nacional del Comahue. https://bibliotecadelenguas.uncoma.edu.ar/items/show/332

Liendo, P; Maure, N; Maluenda, S; Salinas, S. (2018). Alfabetización académica: Traducción, investigación y enseñanza. En Actas Congreso El Conocimiento como espacio de encuentro. 5ta edición.

Marin, Marcos (1994). Informática y Humanidades. Gredos.

Reiss, K. & Vermeer, H. (1991). *Grundlegung einer Allgemeinen Translationstheorie* (2da edición). Niemeyer.

Rodríguez-Inés, P. (2008) *Uso de corpus electrónicos en la formación de traductores (inglés-español-inglés)* (ISBN 9788449043307), [Tesis doctoral, Universitat Autònoma de Barcelona]. Tesis Doctorals en Xarxa. https://www.tdx.cat/handle/10803/286111#page=1.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford University Press.

Swales, J. M. (1990). *Genre Analysis: English in Academic and Research Settings.* Cambridge University Press.

Swales, J. M. (2004). *Research Genres: Explorations and Applications.* Cambridge University Press.

Tolchinsky, L. (2014). El uso de corpus lingüísticos como herramienta pedagógica. *Textos. Didáctica de la Lengua y de la Literatura* (65), pág. 9-17.

Trosborg, A. (1997) (Ed.). *Text Typology and Translation.* John Benjamins.

4. Herramientas informáticas para la gestión de corpus digitales

Los corpus digitales o informatizados representan una herramienta sumamente importante para la investigación lingüística ya que permiten ahorrar tiempo y trabajar de manera más ordenada y exhaustiva (Tramallino, 2021).

Es pertinente mencionar que un corpus lingüístico informatizado o digital no es meramente una recopilación de documentos; debe cumplir ciertos criterios de organización interna de los datos con un propósito lingüístico. En esta línea, Torruella y Listerri (1999) señalan que un corpus informatizado es «una recopilación de textos seleccionados según criterios lingüísticos, codificados de modo estándar y homogéneo, con la finalidad de poder ser tratados mediante procesos informáticos y destinados a reflejar el comportamiento de una o más lenguas» (p. 7). Parodi (2008) agrega que el corpus debe estar disponible para que pueda ser reutilizado en otras investigaciones o estudios de una lengua, y contar con datos precisos sobre la procedencia y recopilación de las muestras.

En los últimos años se ha desarrollado una gran cantidad de programas para el estudio de corpus que pueden utilizarse en línea o mediante la descarga e instalación en nuestras computadoras. Estos programas actúan principalmente como interfaz entre los lingüistas y las computadoras para extraer del corpus elementos de información relevantes en una investigación. Más adelante veremos de manera general el funcionamiento de dos ejemplos de software diseñados para tal fin (ver § 5.2).

Una de las ventajas de utilizar herramientas o programas informáticos es que permiten recopilar grandes cantidades de documentos que, mediante una correcta codificación, ordenación y organización de los datos, permitirá que quien investiga pueda aprovechar al máximo la información contenida en el corpus (Torruella & Listerri, 1999).

En nuestra opinión, según los programas más comunes, en el proceso de trabajo con corpus electrónico, podríamos identificar cinco etapas o fases de aplicación de herramientas informáticas:

1. Creación
2. Anotación
3. Análisis
4. Visualización
5. Exportación

Estas etapas pueden completarse y reiniciarse de forma consecutiva, como se ilustra en la Figura 3.

Figura 3.

Fases del trabajo con corpus electrónico

La implementación de herramientas informáticas en cada etapa dependerá de las prestaciones del software utilizado y de los objetivos de cada investigador. En estas fases de trabajo, es común encontrar terminología referida a programas específicos:

1. Creación: compiladores, conversores, rastreadores web (*web crawlers*), removedores, procesadores de texto, entre otros.
2. Anotación: codificación, asignación de etiquetas, anotación sintáctica y semántica (*parsing*).
3. Análisis: de concordancia, colocación, co-ocurrencia, frecuencia, listas de palabras.
4. Visualización: nubes de palabras, redes de términos, árbol doble, gráficos de distribución, exploración.
5. Exportación: datos en diferentes formatos de intercambio de archivos capaces de ser interpretados por otras aplicaciones informáticas tales como: xml, csv, txt, rdf, html, etc.

4.1 Programas y aplicaciones para la gestión de corpus

A continuación, presentamos una recopilación de herramientas que pueden utilizarse según la etapa o fase de trabajo en la que se encuentre la investigación. Cada programa de la lista contiene un enlace al sitio del propietario, para quienes deseen conocer en detalle su funcionamiento. Se incluye una etiqueta según la función o el tipo de análisis a realizar, aunque algunos consisten en paquetes de programas que posibilitan un trabajo integral con múltiples funciones para analizar un corpus digital:

Fase de Creación

- Online Convert.com -- conversor
- AntFileConverter -- conversor
- BootCat -- compilador

- Corpus Text Processor -- compilador
- CLaRK -- compilador
- SpiderLing -- rastreador, compilador

Fase de Anotación

- ANVIL -- anotador
- ANNIS -- anotador
- CorefAnnotator -- anotador
- Dexter -- anotador
- DisMO -- anotador
- UAM CorpusTool -- anotador
- UAM ImageTool -- anotador

Fase de Análisis

- aConCorde -- concordancias
- AntConc -- concordancias
- AntPConc -- concordancias
- Concordancer -- concordancias
- ConcGramCore -- colocación, concordancias
- Pareidoscope -- colocación
- KHCoder -- analizador, compilador
- KWords -- palabras clave
- OneClick Terms -- palabras clave
- kfNgram -- n-gramas

Kits de Investigación

- ATLAS.ti -- paquete de software
- CATMA -- paquete de software (on-line)
- CorpKit -- paquete de software
- CorpusExplorer -- paquete de software
- NOOJ -- paquete de software

4.2 Corpus informatizados de consulta en línea

Del mismo modo, recopilamos algunos enlaces a los corpus más conocidos que pueden consultarse de manera directa en internet. Algunos de estos corpus pueden descargarse para trabajar sin conexión:

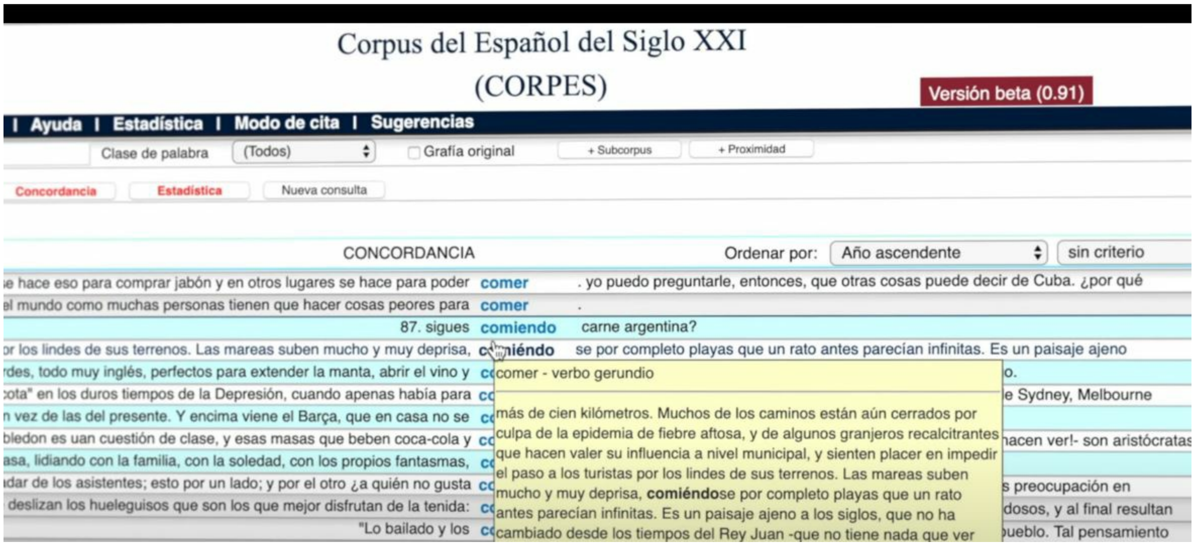
Corpus de Referencia

- English-Corpora.org -- Mark Davies -- inglés
- CREA -- Real Academia Española -- español
- CORPES XXI -- Real Academia Española -- español
- Corpus del español -- Mark Davies -- español
- CORDIAM. Corpus Diacrónico y Diatópico del Español de América -- Academia Mexicana de la Lengua -- español
- El Grial -- Pontificia Universidad Católica de Valparaíso -- español

Corpus de Aprendientes

- CHILDES (Child Language Data Exchange System) -- Brian MacWhinney & Catherine Snow -- inglés
- International Corpus of Learner English (ICLE) -- Sylviane Granger (University of Louvain) -- inglés
- SPLLOC (Spanish Learner Language Oral Corpus) -- Universidad de Southampton, Universidad de Newcastle, Universidad de York -- español
- AACFELE (Adquisición y aprendizaje del componente fónico del español como lengua extranjera) -- Universidad de Alcalá -- español
- CEDEL2 (Corpus escrito del Español como L2) -- Cristobal Lozano -- español
- CAES (Corpus de Aprendices de Español como Lengua Extranjera) -- Instituto Cervantes -- español

En internet, existen múltiples ejemplos de consulta y tutoriales para el uso de corpus en línea. CORPES XXI ofrece un video[2] que describe de forma breve una búsqueda y visualización de resultados para el análisis de concordancia:



(ver en Youtube: https://www.youtube.com/watch?v=m6e_NUJI_rM)

Referencias

Parodi, G. (2008). Lingüística de Corpus. Una Introducción al Ámbito, RLA. En *Revista de Lingüística Teórica y Aplicada*, 46 (1), 93-119.

RAEInforma. (2021). *El CORPES XXI, en un minuto.* https://www.youtube.com/watch?v=m6e_NUJI_rM

Rodríguez Ines, P. (2008). *Uso de corpus electrónicos en la formación de traductores* [Tesis Doctoral, Universitat Autónoma de Barcelona]. https://www.tdx.cat/bitstream/handle/10803/286111/pri1de2.pdf?sequence=1&isAllowed=y

Berberich, K. & Ingo Kleiber (2020). Tools for Corpus Linguistics. https://corpus-analysis.com/ Torruella, Joan & Llisterri, Joaquim. (1999). Diseño de corpus textuales y orales. En *Filología e informática.*

https://gramatica.usc.es/~gamallo/aulas/lingcomputacional/biblio/LinguisticaD
Tramallino, C. P. (2021). Avances en el tratamiento computacional en corpus de aprendientes de español como lengua segunda y extranjera. *Quintú Quimün. Revista de Lingüística,* 5. http://revele.uncoma.edu.ar/htdoc/revele/index.php/lingustica/article/view/31

Vivaldi Palatresi, J. (2009). Catálogo de herramientas informáticas relacionadas con la creación, gestión y explotación de corpus textuales. *Revista Tradumàtica: tecnologies de la traducció*, 7, 1-9.

[2] Disponible en https://www.youtube.com/watch?v=m6e_NUJI_rM

5. ¿Qué herramientas existen para anotar y analizar un corpus?

5.1 ¿Para qué analizar un corpus? La importancia de definir paradigma y método de investigación.

Como se observó anteriormente, los objetivos de la utilización de corpus pueden ser muy amplios. Por ello, es fundamental que los investigadores definan a qué paradigma de investigación responden antes de decidir sobre el método que se va a seguir. Como vimos en § 2.2, los investigadores pueden adherir a un paradigma *cientificist*a, en cuyo caso parten del supuesto de que hay una realidad objetiva que no depende del punto de vista de las y los investigadores; o tener una postura *constructivista*, que sostiene que se parte de las ideas para construir la propia realidad.

En línea, también, con lo expuesto en § 2.2, el método de investigación, y luego la metodología, dependerá del paradigma elegido. Para anotar y analizar textos de un corpus, podemos elegir entre dos métodos de análisis:

- El *método cuantitativ*o: que utiliza etiquetas *predefinidas* para identificar rasgos o propósitos comunes de un tipo de texto (como frecuencia, distribución, colocaciones, palabras clave, entre otras). Su objetivo es obtener datos estadísticos para analizar fenómenos *preexistentes*, y así poder hacer *generalizaciones* de los resultados.
- El *método cualitativo*: que utiliza etiquetas existentes o creadas especialmente para hacer anotaciones o realizar un análisis con el objeto de registrar *fenómenos* mediante la observación de la *asociación* o relación entre

variables en un *contexto situaconal* determinado. El fin último es identificar la *naturaleza* profunda de las realidades, sus sistemas de relaciones, y su dinamismo.

Algunas de las herramientas, entre las que se encuentran las que introduciremos con mayor detalle más adelante (ver § 5.3), son consideradas *híbridas*, es decir que permiten llevar adelante análisis cualitativos y cuantitativos. En otras palabras, quienes utilicen esta herramienta podrán:

- explorar la causalidad entre variables, la operacionalidad y la medición de conceptos; usar grandes muestras y lograr obtener una generalización (*métodos cuantitativos*); o
- desarrollar teorías y modelos a partir de las anotaciones, los datos obtenidos y el análisis realizado; evaluar los fenómenos complejos que requieren múltiples métodos de análisis; utilizar muestras pequeñas analizadas en profundidad (*métodos cualitativos*).

En resumen, el gran beneficio de este tipo de herramientas es que brindan la posibilidad de estudiar datos sobre la lengua mientras es utilizada en actos reales de comunicación, en forma de textos.

5.2 ¿Cómo anotar y analizar un corpus?

Para anotar y analizar un corpus, es posible trabajar con textos escritos y transcripciones de textos orales. En ambos casos, los investigadores en lengua y lingüística pueden seleccionar textos especializados, auténticos o con características particulares, según el propósito de la investigación, para marcarlos.

Estos textos, luego, se cargan en la interfaz de la herramienta para llevar a cabo un análisis, que se almacena y organiza en un mismo proyecto con distintas *colecciones* (carpetas con sus correspondientes textos) organizadas por temas. Es importante considerar que en cada proyecto puede haber un único usuario o varios, quienes pueden utilizar la aplicación de forma simultánea con algunas herramientas. Una vez que se accede a la colección, cada usuario puede *anotar* los textos con las etiquetas previamente definidas, por ejemplo, indicando o *marcando* características propias del tipo de texto, del uso de la lengua, etc.

5.2.1. ¿Qué significa marcar en este contexto?

La *anotación* consiste en señalar o marcar aspectos *lingüísticos* particulares de cada texto con *etiquetas*. Desde luego, el número y tipo de etiquetas dependerá del propósito de la investigación. Por ejemplo, es posible anotar rasgos discursivos de un texto, que se pueden dividir en distintos sub-rasgos: comunicativos, formales y cognitivos y que, a su vez, se pueden subdividir en distintas subcategorías y sub-subcategorías determinada por los usuarios. Una vez *anotados* los textos, estos *códigos* pueden relacionarse para formar *familias y subfamilias*; es decir, *analizarse*.

5.2.2. ¿Cómo se visualizan los resultados?

Estos vínculos pueden visualizarse de distintas formas, como ya se mencionó, según el propósito con el que se quieran analizar los datos obtenidos; por ejemplo:

- Nube de palabras: sirve para evaluar la frecuencia con la que ciertos rasgos o ciertas palabras aparecen en determinados textos (ver también §4).
- Gráfico de distribución: una opción de visualización de algunas herramientas para crear gráficos de distribución para palabras, frases, grupos de palabras o etiquetas a partir de uno o más textos y en uno o más diagramas.
- Árbol doble (*double tre*e): una opción interactiva para visualizar las palabras clave en contexto.

- Palabra clave en contexto (*KWIC*, por sus siglas en inglés): otra de las opciones de visualización de algunas herramientas, que muestra las palabras que aparecen delante y detrás de la palabra que fue anotada como palabra clave.

5.3. Algunos ejemplos del uso de herramientas para anotar y analizar corpus

Como se explicó anteriormente, un corpus puede adaptarse a distintos propósitos de investigación. Esta característica hace posible realizar, por un lado, un análisis cuantitativo para identificar rasgos o propósitos comunes o frecuentes de un tipo de texto y, por otro lado, un análisis cualitativo para registrar ocurrencias lingüísticas, como colocaciones, metáforas, terminología, etc.

En paralelo con la LC, la tecnología computacional también avanzó vertiginosamente en el desarrollo de *hardware* (sistemas físicos) y de *software*, (programas computacionales), lo cual permitió que, a mediados de la década de 1980 y en poco tiempo, surgieran los programas informáticos de ayuda al análisis cualitativo de datos —también conocido como QDA, por sus siglas en inglés— y así sus primeros usuarios, los analistas cualitativos. No obstante, desde entonces, su uso ha aumentado considerablemente entre académicos que realizan investigaciones cualitativas y cuantitativas. Estos notables avances tecnológicos contribuyeron a la construcción y el almacenamiento de estas bases de datos computarizadas, así como al desarrollo de sistemas de interrogación y recuperación de la información contenida en dichos sistemas (Parodi, 2008) (ver también §4).

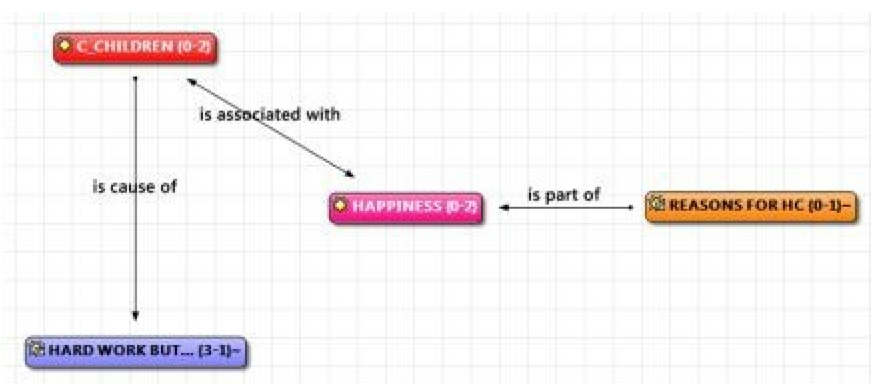
A continuación, se detallarán dos herramientas de marcado de textos, que permiten realizar tanto un análisis cualitativo como cuantitativo y que ofrecen distintas opciones de visualización de resultados, como ya se mencionó en §5.2.2.

5.3.1. ATLAS.ti

ATLAS.ti[3] es un programa informático de ayuda al análisis cualitativo de datos (QDA). En resumen, el programa permite *segmentar* y *codificar* datos específicos, determinar *relaciones* o diálogos entre distintos *códigos* y desarrollar *anotaciones* para llevar adelante una revisión precisa del sistema que se utiliza. Es decir, el análisis se almacena y organiza en un único archivo, llamado *Unidad Hermenéutica*, donde se encuentran los documentos primarios —en el caso de los investigadores en lengua y lingüística, los textos—, de los que se extraen citas (*segmentos*) a los que se aplican los códigos (etiquetas). Estos *códigos* son las unidades básicas de análisis, que luego pueden relacionarse para formar familias (y subfamilias) (Muñoz Justicia & Sahagún Padilla, 2017). Estas relaciones, llamadas *vínculos*, pueden visualizarse en *vistas* de red, como se muestra en la Figura 4.

Figura 4.

ATLAS.ti 7 Quick Tour - Guía Rápida (Friese).



Originalmente el programa se diseñó teniendo en cuenta las necesidades de las investigaciones en áreas relacionadas a las ciencias sociales, pero luego su uso se extendió a muchas otras ciencias que encontraron útil su aplicación.

El objetivo principal de ATLAS.ti fue el desarrollo de una herramienta efectiva para que los usuarios pudieran analizar grandes cantidades de material de investigación, notas y teorías relacionadas. Si bien el programa facilita muchas de las actividades vinculadas con el análisis cualitativo y la interpretación de datos —como la selección, el etiquetado y la anotación—, es importante destacar que su propósito final no es la automatización completa de estos procesos. En resumen, estas son las principales ventajas del uso de Atlas.ti:

- No se limita al análisis de una rama científica en particular.
- Permite un análisis cualitativo, es decir que determina los elementos que conforman la fuente primaria de datos, por ejemplo, datos textuales, gráficos y audiovisuales e interpreta su significado.
- Hace posible la transformación de datos en conocimiento útil, en otras palabras, la administración del conocimiento.
- Ofrece versiones de prueba gratuitas y planes económicos para estudiantes, instituciones educativas, entre otros. Además, los usuarios se pueden formar a través de seminarios en línea gratuitos sobre la herramienta.

5.3.2. CATMA 6

CATMA FOR UNDOGMATIC TEXTUAL MARKUP AND ANALYSIS

CATMA[4] es un software de código abierto para el análisis y marcado de texto asistido por computadora que se creó en la Universidad de Hamburgo y que se encuentra en continuo desarrollo desde 2008.

Este programa tiene un enfoque «no dogmático», es decir que el sistema no prescribe un conjunto de reglas o esquemas de *anotacione*s, tampoco obliga a los usuarios a aplicar taxonomías rígidas a los textos, sino que los alienta a explorar los fenómenos textuales desde sus necesidades: es posible crear, ampliar y modificar continuamente los conjuntos de *etiquetas* creados. De esta manera, si existe más de una interpretación para un texto determinado, el sistema no interfiere de ninguna forma en la selección de múltiples *etiquetas* ni en las *anotaciones* incluidas.

El programa combina tres módulos de funciones interactivas: el módulo *Anotar* permite la creación flexible de conjuntos de *etiquetas* y el *etiquetado* digital de textos; el módulo *Analizar* ofrece una serie de opciones de análisis predefinidas, así como la posibilidad de introducir consultas individualizadas; y con el módulo *Visualizar*, se obtiene una disposición gráfica de los datos.

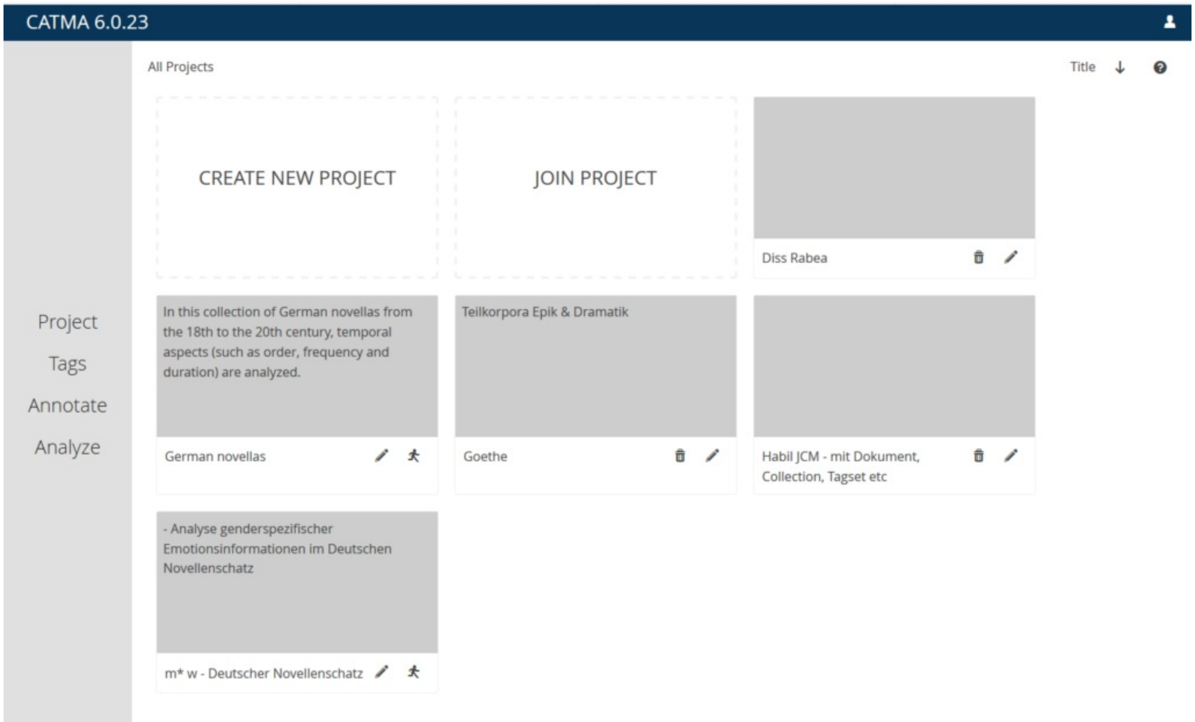
Estos son, brevemente, los principales beneficios que ofrece el uso de CATMA 6:

- Habilita la exportación de los datos marcados y su reutilización en otros contextos.
- Aunque puede utilizarse para realizar un análisis cualitativo, también es posible realizar un análisis cuantitativo mediante el módulo *Analizar*.
- La arquitectura modular permite que los usuarios aborden el texto desde distintos parámetros. Por ejemplo, en el módulo *Anotar* se puede realizar una lectura minuciosa y en el módulo *Analizar* podemos obtener un análisis cuantitativo de los datos y anotaciones.
- Es posible trabajar en línea, en equipos y de forma simultánea mediante la creación de *proyectos*. Es decir, no es necesario descargar una aplicación o programa, sino que se puede acceder de forma grupal e individual a los

distintos *proyectos*, generar cambios, guardarlos y sincronizarlos para que el resto del equipo los visualice, como se muestra en la Figura 5.

Figura 5.

Visualización de proyectos en CATMA 6



Referencias

Friese, Susanne (2021). ATLAS.ti 9 Windows User Manual. ATLAS.ti Scientific Software Development. https://doc.atlasti.com/ManualWin.v9/ATLAS.ti_ManualWin.v9.pdf

Muñoz Justicia, J & Sahagún Padilla, M. (2017). *Hacer análisis cualitativo con Atlas.ti 7. Manual de uso*. Versión 11. Licencia Creative Commons. Atribución 4.0 Internacional. https://manualatlas/psicologiasocial.eu/atlasti7.html

Parodi, G. (2008). Lingüística de Corpus. Una Introducción al Ámbito, RLA. En *Revista de Lingüística Teórica y Aplicada*, 46 (1), 93-119.

[3] Versión de prueba disponible en https://atlasti.com/free-trial-version/

[4] Disponible en https://catma.de/

Esperamos que la lectura de este manual haya resultado útil. De ser así, agradeceremos su difusión, y los invitamos a contactarse con nuestro equipo para enviarnos consultas y sugerencias. Por último, nos será de gran ayuda recibir sus comentarios. Los alentamos a completar el cuestionario.

¡Muchas gracias!

Paula J. Liendo, Micaela J. González, Stella M. Maluenda, Norma A. Maure, Silverio Ortiz, Leticia N. Pisani, Romina N. Sánchez.

PIN J031: *Alfabetización académica y tipologías textuales en la enseñanza del inglés para la traducción*. Facultad de Lenguas. Universidad Nacional del Comahue, 2018-2022.