Introducción a los corpus lingüísticos

Gerardo E. Sierra Martínez
Instituto de Ingeniería, UNAM

18 de agosto de 2015

Prólogo

Como linguistas y como ingenieros nos enfrentamos a situaciones en las que estas áreas se mezclan: requerimos de bases lingüísticas para efectuar algunos desarrollos informáticos, tales como traductores y resumidores automáticos, sistemas de búsqueda para la web o clasificadores de información; o bien, necesitamos automatizar procesos para crear diccionarios, realizar investigaciones filológicas, hacer búsquedas y listas de palabras, entre muchas otras.

Un caso concreto: se necesita un diccionario del vocabulario científico básico del español de México. ¿Cómo sabemos de dónde extraer los términos?, ¿cómo validar tales palabras?, ¿de qué modo se pueden definir?, ¿qué herramientas tenemos a nuestra disposición para facilitarnos la tarea?, ¿qué métodos podemos seguir para efectuar esta labor?

Cuestiones como éstas se tratarán a lo largo de este libro. Los objetivos fundamentales son proporcionar el concepto de corpus, dar una guía de las técnicas que pueden emplearse para operar un conjunto de datos textuales u orales y revisar con ejemplos existentes los resultados que se pueden obtener.

El presente libro se basa en el curso sobre procesamiento de corpus textuales y orales, impartido durante varios semestres en la Universidad Nacional Autónoma de México (UNAM), tanto a nivel licenciatura, en la Facultad de Ingeniería y en la Facultad de Filosofía y Letras, como a nivel de posgrado en los programas de Lingüística y de Ciencia e Ingeniería de la Computación. Asimismo, el libro reúne la experiencia adquirida a lo largo del desarrollo de los proyectos de investigación básica y aplicada en el Grupo de Ingeniería Lingüística (GIL) del Instituto de Ingeniería, UNAM.

Este curso, al igual que otros impartidos en estas licenciaturas y posgrados, se ubica en el área de tecnologías del lenguaje, también conocida como procesamiento del lenguaje natural, lingüística computacional o ingeniería lingüística.

Por tanto, el ámbito que abarca el libro mezcla dos áreas de conocimiento, la lingüísti-

ca y la computación. Gracias a la sinergia de ambas es posible contar con desarrollos complejos y faciltan la labor filológica relegando el uso de fichas al pasado. Si bien, la tradición ha sentado bases sólidas para los estudios lingüísticos, existen actualmente diversos métodos y herramientas que posibilitan procesar la información de manera más precisa y economizan recursos para llegar a resultados mejor cuidados y de manera más rápida.

El libro se organiza en cinco apartados para su mejor comprensión. El primero proporciona los elementos necesarios para entender cabalmente el concepto de corpus lingüístico, desde su definición, la descripción de sus características, su tipología y la descripción de los principales corpus existentes en español, a la vez que se describe cómo puede ser utilizado Internet para formar corpus. Una vez clarificado el concepto de corpus, en el segundo apartado se presenta el proceso necesario para poder compilar un corpus, ya sea textual u oral, desde el diseño del mismo, la obtención de los datos y su registro. Una vez recopilado, en el tercer apartado se describen los elementos básicos de anotación de corpus para distintos fines, destacando el lenguaje de etiquetado XML que es el principal usado hoy en día. Así, una vez tomando en cuenta los aspectos necesarios para la construcción de un corpus, en el apartado cuatro se señalan las técnicas esenciales y algunas herramientas disponibles para analizar corpus. Finalmente, con el fin de ejemplificar la utilidad de los corpus, se mencionan en el apartado cinco diversas aplicaciones tanto para la lingüística y la lingüística aplicada, como para las tecnologías del lenguaje, todas ellas realizadas en el seno del GIL.

Cabe mencionar que si bien la lectura del libro puede ser secuencial de principio a fin, en el seguimiento de un curso conviene tener un panorama general, pero combinar la información vertida en diferentes capítulos. Por ejemplo, empezar con una aplicación concreta, revisar el corpus utilizado, las herramientas o técnicas seguidas, y con ello mencionar la teoría descrita en los tres primeros apartados. Con el fin de facilitar la lectura, se dejó al final de cada capítulo la bibliografía recomendada para ampliar la información.

Agradecimientos

A lo largo de los distintos cursos impartidos sobre lingüística de corpus han transitado muchos alumnos entusiastas en aplicar los conocimientos adquiridos y en aportar cada uno con su experiencia al enriquecimiento de esta obra. Algunos de ellos incluso han continuado la impartición del curso en la licenciatura, en el posgrado y en formación continua; en particular, Carlos Méndez y Javier Cuétara, hoy expertos en corpus textuales y orales, respectivamente.

Asimismo, en el desarrollo de los proyectos del Grupo de Ingeniería Lingüística ha estado involucrado en mayor o menor medida el diseño, construcción y empleo de corpus de diversa naturaleza. Dichos proyectos han sido realizados en el Instituto de Ingeniería y han sido patrocinados principalmente por el Consejo Nacional de Ciencia y Tecnología y por la Dirección General de Asuntos del Personal Académico de la UNAM.

Para la preparación de este libro ha participado un grupo importante de magníficos colaboradores, entre ellos: Valeria Benítez, Yolanda Bravo, Paulina de la Vega, Marina Fomicheva, Jorge Lázaro, Gabriela Méndez, Ariana Paz, Teresita Reyes y Patricia Velázquez-Morales.

El último empujón ha sido obra de un proyecto de Abel Herrera, profesor de la Facultad de Ingeniería, con un proyecto PAPIME para la edición del libro, así como una estancia sabática en la Universidad de Aviñón.

Sea la contribución de todos para la formación de investigadores, profesores, profesionales y estudiosos del lenguaje y de las tecnologías del lenguaje que aprovechen la experiencia de varios años a nuevas aplicaciones de los corpus lingüísticos.

Indice

Parte I

Introducción a corpus

Capítulo 1

El concepto de corpus lingüístico

Ya desde los estudios tempranos en adquisición del lenguaje, con el fin de hacer un análisis cualitativo de la información, el investigador ha tenido que recolectar una serie de datos lingüísticos. Se tiene registrado entre los primeros trabajos, a fines del siglo XIX, la obtención de un corpus para hacer análisis cuantitativo de la frecuencia y secuencia de las letras en alemán. Tanto en los análisis cualitativos como en los cuantitativos se requiere trabajar con muestras que brinden datos reales de diferentes textos, ya sea orales o escritos. Con cualquiera de sus nomenclaturas, los corpus lingüísticos o los corpus de textos constituyen, de algún modo, un modelo de la realidad lingüística que se quiere observar.

Cualquiera que sea el tamaño o extensión del corpus, el procesamiento de estos textos fue, hasta antes del uso de la informática en la década de los cincuenta, de manera manual. Los investigadores tenían que hacer fichas de papel, de manera que tareas simples, como la búsqueda de ocurrencia de palabras, resultaba ser un trabajo muy laborioso que se realizaba mediante lectura directa en cada una de las fichas.

1.1. Definición de corpus lingüístico

Definir el concepto de corpus en la actualidad no es tan simple como parece. Desde el punto de vista etimológico, corpus proviene del latín y significa cuerpo, donde el cuerpo es texto. Antes de entrar en la definición, conviene hacer mención al uso del término mismo. Es curioso que proviniendo del latín, la Real Academia de la Lengua acepte el nombre de manera indistinta para el singular y para el plural, en tanto que en el inglés se diferencia corpora para el nombre plural. Si bien algunos colegas de habla hispana diferencian corpus de corpora, en este libro se adoptará la postura general para el español, no obstante se conservará el término corpora cuando venga así referenciado de algún otro autor.

Ahora bien, de manera amplia se puede definir que un corpus lingüístico consiste en un conjunto de textos de materiales escritos y/o hablados, debidamente recopilados para realizar ciertos análisis lingüísticos. Una definición de esta naturaleza, sin embargo, requiere realizar algunas precisiones.

Conjunto de textos

Los textos que conforman un corpus deben ser representativos y se compilan según criterios lingüísticos que les permiten ser analizados.

Retomando la definición general, en primera instancia se habla de un conjunto de textos, de manera que puede constituirse por uno o varios libros, una revista o simplemente por un artículo de periódico; puede tratarse de un texto científico, de uno literario o incluso de los mensajes enviados entre dos personas mediante computadoras en línea; se puede hacer un compendio sincrónico o diacrónico de la lengua, referirse a la obra entera de un autor, a solo una de sus obras, o bien al habla de un niño en la etapa temprana de adquisición de su lengua materna; es decir, los corpus pueden formarse con cualquier tipo de texto. A menudo se usan textos individuales para muchos tipos de análisis literario y lingüístico, como ocurre con el análisis estilístico de un poema o con el análisis de la conversación de una muestra de charla de televisión. Con este último caso se aborda otro punto interesante de la definición, pues los textos también pueden proceder de material hablado, aunque en este libro se dedicará especial atención a los corpus de procedencia escrita.

Un corpus puede estar formado por cualquier tipo de texto, es decir, por uno o varios libros, una revista o simplemente por un artículo de periódico. Asimismo, los textos que lo conforman pueden pertenecer a cualquier género textual, esto es, pueden ser literarios, científicos o de lenguaje coloquial.

Debidamente recopilados

Un segundo punto importante para la definición de corpus lingüístico se relaciona con la recopilación de los textos, pues el concepto de corpus como tal puede resultar ambiguo. Una biblioteca, sea cual sea su forma, su tamaño o su tipo, sea digital o de material impreso, no constituye un corpus como tal, independientemente de que se trate de textos escritos bien diferenciados por estratos o marcas culturales, geográficas, dialectales, temáticas o de cualquier otra naturaleza. No obstante, una selección cuidadosa de documentos de esta biblioteca, escogidos con criterios bien delimitados para su posterior análisis, como se verá más adelante, y con los textos capturados adecuadamente, constituirá el primer paso en la construcción de un corpus.

Para análisis lingüísticos

El tercer aspecto a señalar en la definición de corpus se refiere al objetivo, esto es, a la realización de análisis lingüísticos. Los análisis que pueden realizarse en un corpus son de dos tipos: cualitativos y cuantitativos. En los primeros se estudian las características de una lengua o algún fenómeno ocurrido en ésta. En los segundos, todo lo competente a cifras numéricas, frecuencias de aparición, etc. del fenómeno que se estudie en el corpus que se tenga. Por ejemplo, si se quisiera hacer un estudio cualitativo de la correlación grafía-sonido del fonema /s/ en el español del siglo XVI, tendría que describirse en qué contextos se empleaba determinada grafía; en tanto que un estudio cuantitativo indicaría la frecuencia de aparición de cada tipo de grafía.

Tanto en los análisis cualitativos como en los cuantitativos está muy difundido el término corpus para describir el material sobre el cual se realizan las investigaciones: es un conjunto de datos reales y aceptables, debidamente ordenado, codificado y organizado, de diferentes textos recopilados, pertenecientes a un código lingüístico determinado, oral o escrito. Con cualquiera de sus nomenclaturas, los corpus lingüísticos o los corpus de textos constituyen, de algún modo, un modelo de la realidad lingüística que se quiere observar, mas no la realidad misma. Es mediante criterios lingüísticos pertinentes que un corpus no sólo contiene el texto en sí mismo, sino que además proporciona información que facilita su análisis.

Con el innegable valor que constituyen los corpus para las investigaciones lingüísticas, cabe extender su empleo más allá del alcance de las mismas. Como se verá ampliamente en el último capítulo, su uso es tanto para la lingüística teórica como para la aplicada, pero también para las investigaciones y desarrollos en las llamadas tecnologías del lenguaje. Hoy por hoy podemos decir que prácticamente no hay estudios de dichas tecnologías sin el uso de corpus. Cada vez el uso es más extendido y es posible encontrar diversos corpus para distintas áreas en particular. En tanto que antes era un concepto que se limitaba al ámbito de la ciencia del lenguaje, ahora abarca otras disciplinas y se aplica en múltiples tareas.

Ya después de varios años, hoy existen corpus muy potentes, algunos de los cuales pueden ser consultados gratuitamente para múltiples investigaciones, desarrollos y aplicaciones:

En lingüística. El uso de corpus permite realizar estudios que abarcan los distintos niveles de análisis de la lengua: fonético, fonológico, morfológico, sintáctico, semántico y pragmático.

En lingüística aplicada. Los corpus se emplean en áreas como enseñanza de lenguas, análisis del discurso, patologías del lenguaje, lexicografía, terminología, tra-

ducción y lingüística forense, por sólo mencionar algunas.

En tecnologías del lenguaje. Dado que las tecnologías del lenguaje procesan voz y texto, el uso de corpus se extiende para crear sistemas de diálogo, generadores de documentos, recuperadores y extractores de información, traductores y resumidores automáticos, entre muchos otros.

1.2. La lingüística de corpus

A pesar de que el concepto corpus no tiene que ver directamente con la informática, en la actualidad está estrechamente relacionado con ella. Los progresos computacionales y tecnológicos han permitido manejar los corpus con programas informáticos apropiados, lo que proporciona un excelente material para el trabajo de investigación. Por ello, es pertinente hablar de corpus informatizados, es decir, un conjunto de textos elegidos y anotados con ciertas normas y criterios para el análisis lingüístico, de forma que se sirven de la tecnología y de las herramientas computacionales para generar resultados más exactos.

Tomado en cuenta la labor de la lingüística basada en textos, conviene especificar que un corpus es, sobre todo, una colección de textos en soporte informático. Esta colección, si bien llega a ser muy extensa, incluso de varios millones de palabras, puede ser también de dimensiones mínimas. Por extensión, se ha llamado lingüística de corpus a la parte de la lingüística en la que se estudian con medios informáticos de diferentes tipos grandes masas de datos, inabordables de otro modo, para obtener de su análisis, por ejemplo, las características lingüísticas de una lengua en un cierto momento de su historia, de cierto tipo de textos, de un conjunto de autores o un autor determinado, etc.

Con el nacimiento de la informática es posible el análisis de textos de manera eficaz, ya que permite llevar a cabo cálculos complejos en cuestión de segundos y sin los errores naturales de cualquier persona. A pesar de las ventajas, hoy en día existe cierta renuencia al uso de las computadoras, una antipatía que se debe más al miedo infundado a la tecnología que a considerarlas falibles o usurpadoras de la labor humana. Por el contrario, conviene tomar en cuenta que la posibilidad de falla del humano es mayor que la computadora cuando se trata de labores sistemáticas y de alta precisión. Sólo con el fin de mostrar los equívocos que pueden ocurrir durante el procesamiento manual de ciertas tareas, en el cuadro siguiente se encuentra un texto corto que aparece en Internet en el que se pide al lector contar el número de veces que aparece la letra “F”.

FINISHED FILES ARE THE RE-
SULT OF YEARS OF SCIENTIF-
IC STUDY COMBINED WITH THE
EXPERIENCE OF YEARS

Del ejercicio del cuadro anterior, es fácil comprobar con diferentes personas que el total de “F” va de tres a seis. Esto se debe a que el ser humano muchas veces pasa por alto las preposiciones o los artículos, en este caso las tres ocurrencias de “OF”. Este error, trasladado a análisis lingüísticos más extensos, les restaría por supuesto confiabilidad y sentido.

1.3. Características que debe cumplir un corpus

Existen corpus que se construyen en función de requerimientos específicos para ciertos proyectos particulares como puede ser el análisis estilístico de la obra de Octavio Paz o el alineamiento de las actas de una reunión bilingüe para encontrar la equivalencia de los nombres propios; por el contrario, en otras investigaciones se buscaría recolectar un estrato mayor de lengua para la realización de diversos estudios lingüísticos a partir de métodos empíricos, como sería el caso del desarrollo de un sistema de diálogo en un dominio muy concreto, en donde intervienen estudios fonéticos y acústicos, sintácticos, semánticos y pragmáticos. Algunos estudios requieren de anotaciones muy precisas, mientras que otros abarcan un panorama más amplio de códigos lingüísticos. Cualquiera que sea el estudio, los corpus comparten características dominantes que les permiten ser concebidos como tales y ser diferenciados de otras colecciones de textos que no lo son.

El concepto de corpus lingüístico ha sido discutido desde varias perspectivas, lo que ha suscitado controversias y diferencias en donde se ponderan unos criterios frente a otros, a veces sin llegar a un acuerdo. En el extremo clásico de la lingüística empírica, todo registro sobre hechos observables puede constituir un corpus, incluyendo las transcripciones en un cuaderno sobre el habla espontánea de una niña de cuatro años, en donde se anotan con círculos y cuadros los hechos que se quieren destacar. Siguiendo la misma línea, se puede argumentar que cualquier colección de textos lo es, ya sea desde aquellos en los que se quiere extraer cierta información, como sería la recopilación de la terminología de un texto de especialidad, o de aquellos que van a ser objeto de estudio, como sería una serie de oficios y memorandos a los que se les quiere identificar la estructura para elaborar modelos de uso en una compañía, por ejemplo. Como tal, estas son muestras de lo que normalmente se ha llamado corpus, es decir, material de registro para la realización de un trabajo en particular. Si bien parece haber acuerdo con esta

perspectiva, todavía el concepto está lejos de lo que se denomina corpus lingüístico, y no propiamente porque no sean textos para estudios lingüísticos, sino por los criterios con los que están construidos.

Contención de datos reales

Los datos proporcionados en los textos que conforman un corpus deben mostrar a pequeña escala cómo funciona una lengua natural. A pesar de las diferencias que se puedan suscitar entre los estudios empíricos y los intuitivos, es necesario reconocer que un corpus constituye un reflejo de la realidad y que la lengua, sea oral o escrita, puede ser modelada mediante los corpus, ya que en ellos se tienen pruebas fehacientes y constatables de actos de habla que se convierten en fuentes invaluables de datos para la investigación lingüística empírica y para las aplicaciones de la lingüística aplicada. Se puede ser tan específicos en el estudio que es válido hacer un tratado completo de la obra de un autor o incluso trabajar con sólo una de sus obras. Sin embargo, en la lingüística de corpus se pretende algo más que un análisis tan específico, muchas veces se busca que el material recopilado sirva como una referencia representativa para una variedad de lenguaje en particular. Aunque un solo acto de habla puede argumentarse como verdadero, no es suficiente para generalizarlo y extenderlo a otros hablantes. La cuantificación en la lingüística de corpus es más significativa que la cuantificación en las otras formas de lingüística empírica porque las conclusiones son válidas para una mayor población, antes que para la muestra sola, con lo que puede constituirse una hipótesis en torno a la organización y funcionamiento de una lengua natural en un determinado contexto. De esta manera se abre un abanico de posibilidades en las aplicaciones. Por ejemplo, en la enseñanza de lenguas que basa su instrucción en los modelos verdaderos de uso, es posible ahora dotar a los estudiantes con las herramientas que necesitan para producir y comprender la lengua apropiadamente en cualquier tipo de contextos.

Representatividad

A fin de poder asegurar que las conclusiones que se obtengan de un estudio puedan ser generalizadas, el primer aspecto a considerar en el diseño de un corpus lingüístico es la obtención de muestras representativas de las lenguas o lengua en cuestión, debido a que un corpus siempre es una muestra de lengua y no pretende ser la totalidad de ella. Para ello, el primer paso es definir la población que se desea estudiar, lo que tiene que estar acorde con los objetivos del proyecto. En tanto en el objetivo se defina una población específica será posible considerarla así, pero para hacer generalizaciones y presentar el estado de una lengua, debemos ser cuidadosos en escoger la población para que sea representativa, esto es, para que la variabilidad de la muestra represente lo más aproximado posible a la población en los aspectos que se desean estudiar. En investigaciones de adquisición del lenguaje materno para niños de tres años de edad

en zonas rurales, la población estará entonces específicamente restringida y habrá que buscar informantes de esta edad que cumplan con estos requerimientos, y de ninguna manera conformarse con escoger tres niños que vivan en zonas rurales. Asimismo, en casos como el análisis de frases nominales en el Siglo XVI no se puede uno restringir únicamente a la obra de Bernal Díaz del Castillo, no obstante sean las crónicas más completas sobre la conquista de México, pues el habla de un autor no es reflejo fiel del resto de sus contemporáneos. Por tanto, hay que tener cuidado en los resultados que se exponen y evitar hacer generalizaciones a partir de aspectos particulares; por ejemplo, concluir que grabaciones del habla de los estudiantes en Ciudad Universitaria, por más que exista diversidad de estudiantado, es reflejo del habla de la ciudad de México, es una conclusión desmesurada.

Definir la representatividad de un corpus no es una tarea sencilla, será necesario tener una cierta organización o estructura jerárquica de la población estudiada, de manera que sea posible identificar los puntos de interés. Esto involucra la ponderación de diversos factores, generalmente asociados a la delimitación de nuestro objeto de estudio y de los alcances del proyecto. Por ejemplo, en un estudio concreto de la anotación de las grafías del sistema de sibilantes castellanas presentes en documentos coloniales del siglo XVI, en donde el objetivo es establecer normas ortográficas de acuerdo con el origen dialectal del escribano, será pertinente restringir la población de acuerdo con parámetros como la zona geográfica que delimita los dialectos de los escribanos, las etapas que comprende el periodo de la Colonia y el tipo de documentos a buscar.

Variedad

La división del corpus con base en los estratos de la población en marcas geográficas, culturales, dialectales, étnicas, temporales, históricas, o cualesquiera que sean para los fines concretos de estudio, serán de mucho provecho para organizar el corpus, pero también servirán como criterios de clasificación y búsqueda de la información. Los criterios de organización y estratificación de los corpus dependen de cada proyecto, su utilidad rebasa las fronteras de los estudios específicos para los cuales fueron creados y es posible la apertura a otros estudios. La siguiente lista servirá para dar una idea de la diversidad de criterios utilizados en diferentes aplicaciones.

Localidad geográfica. La localidad geográfica de procedencia de los textos o de los informantes, puede dividirse por países (sin dejar de tomar en cuenta que, por ejemplo, en Estados Unidos se encuentra una población muy alta de mexicanos), ciudades (Ayutla, Tlahuitontepec, Totontepec; Ciudad de México, Managua, Caracas) o regiones (chinantecos, mixes, otomíes; Sierra Gorda, Altiplano, Península Ibérica). En este rubro hay que evitar la interferencia geográfica que puede condu-

cir a datos erróneos, en donde, por ejemplo, una persona nacida en Buenos Aires radicada por largo tiempo en México muy posiblemente siga teniendo un acento rioplatense.

Información personal. Los datos personales del informante pueden ser decisivos para algunos estudios léxicos y pragmáticos, por lo que habrá que diferenciar, entre otros datos, el género,la edad o grupo etario, el estrato sociocultural, el nivel de estudios y la preferencia sexual. En este último caso es anecdótico señalar de primera fuente que un traductor germano que ya había proporcionado un par de textos para un corpus, se retractó y negó todos los derechos de usarlos cuando se le entregó el cuestionario donde se preguntaba la preferencia sexual. Si bien este es un caso aislado, de cualquier manera hay que advertir que la información proporcionada será confidencial y para fines estadísticos.

Tópico. El tópico, entendido como el tema o asunto del que se habla, es un criterio subjetivo que siempre será cuestionable por aquellos que ven el trabajo desde fuera. Sin embargo, es necesario tomar en cuenta que los criterios con los que se haya definido el tópico obedecen a los objetivos de un estudio en particular. La tipificación conviene ser apoyada por expertos en la materia, quienes no sólo dominan la tipología del área, sino que además conocen las fuentes más representativas. La división puede ser muy general (ciencias, sociales, humanidades) o bien llegar a nivel de detalle (transportes terrestres, estructuras de concreto, instrumentación sísmica), según se requiera.

Tipo de texto. El tipo de texto, ya sea oral o escrito, es una característica que marca diferencias léxicas, semánticas y discursivas. En textos escritos se podrá hacer una división entre ficción y no ficción (texto científico, biografías, ensayos, periódicos), en tanto en otros estudios interesará llegar a niveles específicos de precisión y diferenciar entre una carta formal, un memorando o una nota manuscrita. En texto oral convendrá distinguir entre habla espontánea y la que no lo es (discursos, lecturas, habla de laboratorio).

Fuente. La fuente explícita del texto es un dato que siempre debe existir y proporcionarse de manera obligatoria en un corpus. Desde el punto de vista de la organización del material será relevante el tipo de fuente (libro, revista, Internet, manuscrito) o la forma de obtención de un “texto oral” (programas de radio y televisión, entrevistas, grabaciones telefónicas).

Tiempo. Tanto para estudios diacrónicos como para estudios sincrónicos es indispensable ubicar los textos en el tiempo. Es necesario registrar el año (normalmente referido a la fecha de publicación o de grabación), o bien el periodo (década, 1925-1945) o la época del texto (Siglo de Oro, Generación del 27, Romanticismo).

Otros. Existen otros diversos criterios utilizados en la categorización de los textos que conforman un corpus, pero dependen más de los objetivos para los que van a ser empleados. Por ejemplo, puede ser de utilidad diferenciar entre el autor, el editor o el traductor de la obra, o bien especificar la nacionalidad, el país de nacimiento y el de residencia.

Equilibrio

La representatividad del corpus requiere variedad de información, aunque en un corpus multipropósito en el que se pretende abarcar un estado general de lengua con el fin de servir a diferentes tipos de investigación es todavía insuficiente. Hay otro factor importante para tener una mejor representatividad y que puede generar mejores resultados, la cual consiste en seleccionar los textos para obtener una muestra equilibrada. El equilibrio significa que el material contenido para cada uno de los rubros sea relativamente proporcional, evitando ser tendencioso a una parte únicamente. Sin embargo, hay que considerar que la adquisición del material muchas veces está supeditado a la disponibilidad o facilidad de acceso al mismo, de manera que es comprensible que el Corpus de la Real Academia de la Lengua Española, al ser realizado en España, esté conformado en un cincuenta por ciento de textos provenientes de dicho país; o en el caso de un corpus especializado, por ejemplo, es esperable que contenga más textos de un área que de otra. Equilibrar un corpus cuando se tienen varios rubros de organización (países, años, tipos de texto y tópicos) requiere hacer un balance de todos y cada uno de los rubros, lo cual se convierte en una tarea muy complicada. Cuando se tiene el apoyo de una base de datos para el registro del material, en donde se diferencian diversos criterios, entonces es posible obtener tablas y gráficas en las cuales pueda verse la distribución de los datos según los criterios y con ello mantener un mejor equilibrio en el corpus, solicitando los textos faltantes.

Selectividad

Los textos de un corpus deben filtrarse de acuerdo con los propósitos del estudio que se quiera realizar, ya que no es posible recopilar todo lo escrito y/o hablado de una lengua. Los criterios para la selección de los textos serán siempre controversiales y cuestionables. Una recopilación de textos literarios buscará tener sin duda a escritores renombrados de la talla de García Márquez, Borges y Paz, en el caso del español, esto sin perder otros exitosos y populares como Corín Tellado, así como una selección variada de autores menores. Las preguntas irán en función de cuántos, cuáles, y qué tanto de cada uno de los textos son pertinentes para la muestra, con el riesgo de encontrar inconformidades respecto a los criterios de selección. Existen algunas propuestas sobre las medidas que deben usarse a fin de tener mayor representatividad de la lengua en la selección de

los textos y su variabilidad, pero todas ellas han respondido a las necesidades específicas del proyecto y su presupuesto. Un criterio común que se ha utilizado para corpus de referencia es la selección aleatoria de muestras, pero también se ha discutido si las características lingüísticas se encuentran distribuidas uniformemente en los textos. Ante tal diversidad de criterios, la mejor opción es hacer una reflexión profunda de lo que se pretende obtener, una delimitación de la población que se va estudiar, una estimación de costos y tiempos, un sondeo de las posibilidades de obtener el material y el apoyo de un heterogéneo equipo de trabajo familiarizado con el área, tanto de lingüística como de computación y estadística.

Tamaño finito

En la literatura es común referirse al tamaño del corpus, esto es, al número de palabras o registros que contiene, considerando palabra a la secuencia de caracteres separados por espacios vacíos, sean éstos números, letras, símbolos o combinación de los anteriores. Dicha aclaración es importante puesto que el término “palabra” es muy controversial y discutido dentro del campo de la lingüística en general, por lo que es conveniente precisar lo que se entiende por “palabra” dentro de la perspectiva de la lingüística de corpus. Existen proyectos más ambiciosos que otros, hay corpus que tienen como objetivo proporcionar una amplia representación de la lengua y normalmente están subvencionados por grandes instituciones. En algunos casos se habla de cientos de millones de palabras, como el de la Real Academia de la Lengua con más de 400 millones, o bien el del Collins Cobuild que tiene más de 520 millones, mientras existen otros corpus de menor tamaño que se conforman con alrededor de 100 millones cada uno, como sucede con el del español de Mark Davis o los corpus nacionales de Gran Bretaña o de Estados Unidos. Ante estas cantidades, erróneamente se llega a considerar que un corpus es rico cuanto más palabras contiene, destacando de esta manera los aspectos cuantitativos a los cualitativos, cuando en realidad la riqueza de un buen corpus no reside tanto en su tamaño cuanto en la variedad y representatividad del mismo. De hecho, un corpus pequeño puede estar mejor diseñado y proporcionar mejores resultados que un gran corpus que considera textos enteros y poco representativos. Como ejemplo, el corpus que se recopiló en El Colegio de México para obtener un conocimiento riguroso del uso del vocabulario en el que se basara la redacción del Diccionario del español usual en México, tan solo consta de dos millones de registros, pero es lo suficientemente rico al contener una extensa y variada recopilación de muestras, mil textos de dos mil palabras gráficas cada uno, rigurosamente seleccionadas de todo tipo de textos hablados y escritos, provenientes de todas las regiones del país, de toda clase de hablantes y de una amplia variedad de géneros.

En este sentido, resulta absurdo tratar de construir un corpus exhaustivo que cubra todos los aspectos de la lengua de manera que se consiga, por ejemplo, todo el léxico,

pues para asegurarlo se requeriría considerar todo texto, oral y escrito, que sea producido por cada individuo, labor que llevaría incontables vidas. Por el contrario, el lingüista busca la representatividad en los textos, esto es, intenta recoger una muestra de la lengua que se estudia para seleccionar los ejemplos cercanos a la realidad lingüística con el fin de analizarlos de manera pertinente. El lingüista tiene que ser selectivo con el material que escoge, tiene que ser muy cuidadosos durante el diseño para que, de acuerdo con el presupuesto y los alcances, se pueda obtener el material esperado. Si bien en un corpus más grande se podrán encontrar ocurrencias que no se localizan en corpus pequeños, vale la pena meditar si estas ocurrencias son representativas o son simplemente datos aleatorios que escapan de la norma lingüística y no son pertinentes para el trabajo de investigación.

Ahora bien, para dar una idea de lo que significa el tamaño del corpus y evitar dar una impresión falsa de los números, vale la pena advertir lo que significa un millón de palabras. Como se dice en la introducción del British National Corpus, si se quisiera leer todo el corpus de cerca de 100 millones de palabras, a razón constante de 150 palabras por minuto durante 8 horas al día, se podría terminar su lectura después de 4 años ininterrumpidos. Esta cantidad resulta impresionante y uno puede imaginarse el trabajo que costará obtener los textos. Sin embargo, si se cuentan las palabras que hay en un escrito, sea un libro, el artículo de una revista, una tesis de doctorado o el ejemplar de un periódico, se puede observar que en sólo unos cuantos textos llegamos a tener el millón de registros. Para el número 38 de la revista Estudios de Lingüística Aplicada se tiene un total de 52,605 palabras en las 150 páginas de contenido, y si consideramos esta cantidad como un promedio general de cada número, entonces se tiene un total de prácticamente dos millones de palabras en los primeros 38 números de la revista.

Cabe advertir, sin embargo, que el millón de palabras como unidad de medida no es la única que existe, ya que ésta es más utilizada cuando se tienen textos grandes. Para corpus de referencia resulta más relevante mencionar el número de fragmentos de textos (llámense muestras, registros, textos, oraciones, cláusulas, etc.) y la longitud media de los mismos (normalmente el número de palabras, ya sea el promedio o el rango). En corpus orales se da mucho peso al tiempo, por lo que es común referirse al número de horas de grabación o al número de grabaciones por el tiempo aproximado de duración de cada una.

Con todo, se define que una característica de los corpus es que su tamaño debe ser finito y aunque hablemos de pocas o muchas palabras, se trata de una muestra de lengua delimitada. La ventaja del tamaño finito es que los corpus no son estáticos, pueden agregarse nuevos textos para obtener una muestra mayor del tema que se trata. Decidir el tamaño del corpus es una cuestión difícil que involucra varios aspectos. El costo y el tiempo son dos factores relacionados para tomar una decisión, pues entre mayores sean éstos, más grande podrá ser el corpus, ya que se requiere de horas/hombre para realizar

todas las tareas de procesamiento del mismo, tareas que van desde la solicitud del material y su recopilación, hasta la digitalización, limpieza y anotación. Asimismo, hay que prever si se puede disponer del material que va a constituir el corpus, porque muchas veces es difícil conseguirlo y los proyectos se quedan truncados. A fin de conseguir variedad de textos, es necesario recurrir a distintas fuentes de información, lo cual dificulta aún más la tarea.

1.4. Corpus informatizado

La tendencia actual en la lingüística de corpus se orienta a informatizar los textos, ya que al estar en soporte electrónico pueden ser recuperados y manejados electrónicamente. Incluso, varios corpus conocidos que en su momento fueron desarrollados en forma impresa, actualmente se han pasado a un formato electrónico. Hoy en día, los estudios de lingüística de corpus ya dan por hecho que se habla de textos informatizados.

El tener textos en un soporte electrónico tampoco es suficiente para construir un corpus: se pueden convertir todos los catálogos de una biblioteca en lo que se llama biblioteca digital, pero no por ello constituir un corpus. Aunque puede ser un recurso sumamente productivo para búsquedas de información lingüística o incluso para el tratamiento de datos numéricos, no debe considerarse como un corpus lingüístico hasta no pasar por un filtro metodológico pertinente. Conviene recordar que en todo corpus se debe tener una clasificación de las fichas o de los textos, de manera que sea recuperable únicamente la información que interesa. En este sentido, una biblioteca digital tendrá una buena clasificación de los documentos, pues es trabajo que compete a la bibliotecología. Sin embargo, una vez que se obtengan los documentos recuperados, el lingüista se interesará por analizar los textos, hacer anotaciones, obtener concordancias, hacer conteos estadísticos y otras labores para las que no es suficiente acceder y manipular las imágenes del documento; en este momento es cuando se alejan las bibliotecas digitales del sentido de corpus.

Asimismo, tampoco hay que dar por sentado que la WEB o Telaraña Mundial constituye un corpus porque en él podemos buscar patrones lingüísticos determinados y obtener con ello los diferentes usos léxicos de una palabra, o bien porque proporciona material suficiente para encontrar las flexiones y formas derivativas del español. Los archivos digitales de la correspondencia de una empresa, una base de datos terminológica bien estructurada y con opción a múltiples búsquedas, o la transcripción de los segmentos de noticias en radio y televisión, todas estas colecciones electrónicas tampoco constituyen un corpus lingüístico como tal, aunque bien son el sustento para integrarlo. Es necesario dar a la colección de textos un tratamiento informático que permita la recuperación y análisis de los documentos con fines lingüísticos, pero sobre todo con base en ciertas

normas y estándares.

Cuando los textos están electrónicamente respaldados puede ser de dos maneras diferentes: uno como imágenes (como si fueran fotos) y otro como textos (por transcripciones o por un proceso de digitalización y reconocimiento óptico de caracteres). Para ver las diferencias entre ambos puede tomarse el caso de un usuario que va a una biblioteca tecnologizada y quiere consultar los periódicos de ciertas fechas o incluso un manuscrito antiguo. Será normal entonces que el bibliotecario, en lugar de traerle los documentos, lo lleve a una sala de cómputo en donde pueda verlos y pasar entre sus páginas a la velocidad que quiera y, en algunos casos, al mismo tiempo que otros lectores podrán estar consultando los mismos. Podrá ver las imágenes insertas en el periódico, los diferentes tipos de letra, los mismos encabezados con su formato correspondiente, pero todo esto no son más que imágenes, una fotografía tal cual en donde las letras no son más que la unión de puntos. Por el contrario, cuando un documento se encuentra en lo que se conoce como formato texto, las letras dejan de ser imágenes para convertirse en letras que a su vez forman palabras, de manera que se pueden manipular los textos de otro modo o buscar las ocurrencias de una palabra e, incluso, de todas las palabras que contienen una letra o una secuencia de letras. Con archivos en formato texto podremos hacer cambios y reemplazar una cadena de letras por otra, cambiar los tipos y formatos de letra, añadir texto o insertar anotaciones; asimismo, podremos traer una lista de las ocurrencias de una palabra junto con los caracteres a su alrededor. Todo esto sería imposible de llevar a cabo en archivos con formato imagen.

En resumidas cuentas, un corpus tiene que estar en soporte electrónico, debidamente clasificado para la recuperación de la información que se necesita y debe ser manejable para poder hacer análisis lingüístico.

Ventajas

Las ventajas de tener el corpus en formato electrónico sobre la forma impresa en fichas de papel son varias, a saber:

La información se puede manipular de manera más sencilla. Debido a que los datos contenidos en un corpus son accesibles dada su naturaleza digital, se puede manipular más fácilmente. Es posible guardar la información, duplicarla y respaldarla; cambiar los nombres y la ubicación de los archivos; transferir, eliminar y añadir datos; realizar las tareas cotidianas de cortar, copiar y pegar, tanto dentro de un archivo como entre archivos; ordenar, clasificar, reubicar textos; en fin, múltiples tareas para las que se requiere conocimiento básico del uso de los programas de cómputo, pues hoy en día existen varias herramientas y programas que facilitan estas tareas.

La velocidad de procesamiento es mayor. La velocidad de procesamiento y de recuperación de información es insuperable mediante la computadora. No tiene punto de comparación con la manera manual y tradicional de los lingüistas que implica tener una muy buena organización y clasificación de las fichas en papel. Generalmente el resultado de trabajar manualmente es que las fichas se vuelven muy personales y difícilmente pueden ser empleadas por otros investigadores.

La precisión de los resultados es exacta. Hay algunas tareas sistemáticas que difícilmente podrían hacerse sin apoyo de los programas de cómputo, sin contar que de manera manual son más frecuentes los errores y por tanto los resultados pueden ser inexactos. Tan solo considérese el conteo de la letra F en un pequeño texto, ejemplo mencionado al principio de este capítulo. No sólo el conteo de letras o de palabras es más productivo en soporte electrónico, su ordenamiento, agrupación, clasificación, combinación y recuperación en el contexto de aparición se vuelve labor más sencilla en un soporte electrónico manipulable por la computadora. Las tareas son múltiples y sólo algunas se describirán más adelante.

La actualización del corpus se posibilita. Al igual que las fichas en papel, se puede actualizar la información. Sin embargo, con el formato electrónico se pueden realizar cambios en todos los registros a la vez, en tanto en papel hay que hacerlo ficha por ficha. Con unos cuantos comandos se puede pedir a la computadora que reemplace una palabra por otra o que incluya una etiqueta a cierta palabra, independientemente del número de veces que ocurra dicha palabra.

El costo de acceso disminuye. La mayoría de los corpus textuales son gratis o de bajo costo, además siempre existe la posibilidad de obtener textos a través de Internet, como se verá más adelante en este capítulo. Sin embargo, cabe hacer notar que los corpus orales pueden llegar a ser muy costosos (por ejemplo, un académico actualmente tendría que pagar arriba de 150 mil euros para obtener registros telefónicos grabados), lo cual se convierte en una desventaja desde el punto de vista del investigador que quiere adquirir los textos, pero como una ventaja si se piensa en la lingüística aplicada y orientada a la comercialización de sus productos.

La posibilidad de compartir la información. Se tienen mecanismos para facilitar el acceso y compartir el material. La información puede estar disponible vía Internet, de manera que puede ser consultada por medios electrónicos desde cualquier lugar del mundo por múltiples usuarios a la vez, aunque evidentemente existen algunas restricciones de uso para garantizar los derechos de autor y evitar el empleo malintencionado del material.

La transferencia de datos. Es posible y más fácil transportar grandes cantidades de texto para leerse en prácticamente cualquier computadora, gracias a las capacidades de almacenamiento que se puede tener en un disco compacto (CD) o en

una memoria portátil USB, a la vez que también existen protocolos para transferir información a través de Internet.

Desventajas

Sin embargo, conviene ser justos y tomar en cuenta que a pesar de todas las bondades que representa contar con corpus en formato electrónico, también existen algunas desventajas:

Se requiere de programas especializados. Si bien resulta relativamente sencillo tener los textos en soporte electrónico, el objetivo final consiste en el análisis de los mismos. Existen varios programas especializados para el manejo de textos, algunos disponibles en forma gratuita y los más comerciales a precios muy accesibles. El uso de estos programas implica sujetarse a las limitaciones que cada uno de ellos impone. Por ello, otra alternativa sería crear programas a la medida para el manejo de los textos a conveniencia de cada proyecto. Esto implicaría, por supuesto, tiempo y dinero para el equipo de trabajo que haga el diseño y realice los programas adecuados a los requerimientos particulares.

Es necesario digitalizar los textos. En algunos casos se requiere digitalizar los textos, esto es, pasarlos al formato electrónico. Ya sea que se transcriban los textos manualmente o se utilice un digitalizador y un programa de reconocimiento óptico de caracteres, de cualquier forma se está sujeto a errores. Pasar los textos a formato digital es un proceso largo y caro, pues requiere una mayor inversión en horas/hombre de lo que sería el diseño y la elaboración de un programa para manejar los textos.

El equipo de cómputo tiene requerimientos particulares. Es necesario adquirir el equipo de cómputo adecuado. Conviene tener una planeación adecuada para comprar el equipo conforme a los requerimientos del proyecto. Es indispensable considerar el tipo y las dimensiones del corpus en cuestión (sea oral o escrito) y los análisis que se realizarán a partir de éste. Para un corpus pequeño de un solo investigador puede ser suficiente una computadora personal, pero para proyectos más complejos habrá que pensar en un servidor con estaciones de trabajo, impresoras, digitalizadores de alta resolución y equipo de respaldo.

Se precisa una inversión económica. Aunque es evidente, pero no por ello resulta obvio, se requiere de la computadora en todo momento. Esto implica que el lingüista debe de alguna manera enfrentarse directa o indirectamente con los equipos de cómputo, por lo menos para referir a los expertos sus necesidades y requerimientos. Si bien puede sólo ver los resultados en forma impresa y nunca acercarse a una computadora, al menos debe tener la perspicacia para ver lo que se puede obtener con un corpus lingüístico.

El equipo empleado se debe actualizar. Al cabo del tiempo y con el avance de la tecnología habrá que ir actualizando todo, desde el equipo de cómputo y los mismos programas, con relativa frecuencia, hasta los formatos, códigos y etiquetas de nuestro corpus, más ocasionalmente. El equipo de cómputo se vuelve obsoleto con una velocidad inusitada, a tal grado que de la fecha en que se hace un plan de compras hasta que se obtiene el presupuesto y se compra el equipo, ya hay en el mercado equipo más sofisticado y de menor precio. Las compañías continuamente desarrollan mejorías en sus programas y ofrecen versiones con aditamentos novedosos y prácticos en tanto surgen nuevos programas a precios más competitivos. Con los avances científicos también se irá viendo la posibilidad de incorporar códigos y etiquetas especiales a los textos para ser reconocidos por los nuevos programas, lo que puede implicar cambios internos en los corpus.

Siempre existe la posibilidad de fallas técnicas. Con todo y que se busca que los sistemas sean más robustos, es indudable que la tecnología llega a fallar, razón por la que será necesario siempre tener un respaldo y estar conscientes de las limitaciones. Solo así se puede garantizar que el trabajo de varios años no desaparezca de un día para otro y sea irrecuperable. Siempre existe la posibilidad de fallas técnicas, con todo y que se busca que los sistemas sean más robustos, y el mejor recurso es estar preparados ante las posibles eventualidades. Por ejemplo, contar con respaldos para el resguardo de la información y para la infraestructura de cómputo. Solo así se puede garantizar que el trabajo de varios años no desaparezca de un día para otro y sea irrecuperable.

Mencionar que un corpus debe ser manejable por la computadora implica que se tengan archivos textuales, no como imágenes. Conviene advertir que existe una gran variedad de formatos de texto, los cuales son manejados por los editores y procesadores de texto. Los editores de texto permiten manejar los archivos que constan de texto plano, esto es, sin formato alguno y sin imágenes, salvo los saltos de línea, la tabulación horizontal y los retornos de carro. Por el contrario, los procesadores de texto pueden incluir diferentes estilos de letra, márgenes, imágenes, comentarios, hiperenlaces y muchos otros elementos. Los programas de procesamiento de texto utilizan su propio formato con el cual codifican los distintos elementos. Actualmente los programas procuran que sus formatos sean compatibles, de manera que un archivo con cierto formato pueda ser leído por otro procesador de texto. Sin embargo, a pesar de que los datos y el contenido se convierten en este proceso, se llegan a cambiar o perder algunas características.

Cuando construimos un corpus es común pensar en conservarlo y compartirlo. Así como las fichas en papel se hacen amarillentas a lo largo de los años y para garantizar su perdurabilidad se adquieren tarjetas de cierta calidad, también es necesario asegurar que los textos electrónicos sigan siendo consultados a pesar de las innovaciones tecnológicas y equipo más sofisticado, de los nuevos procesadores de palabras, programas de

cómputo y sistemas operativos. Si se quisiera que las fichas en papel tengan un almacén central, habría que uniformizar el tamaño y formato de las mismas; de igual manera, los textos informatizados deberán seguir un formato común que pueda ser accesible para cualquiera, aun para uno mismo que los consulta en diferentes máquinas.

1.5. Derechos de autor en corpus

Los derechos de autor son tan importantes como todas las consideraciones anteriormente mencionadas para la adecuada construcción de un corpus, sobre todo cuando se trata de textos que van a ser publicados y difundidos para que el público pueda acceder fácilmente a ellos. Debido a que un corpus recopila textos (orales o escritos) de testimonios personales o privados, o bien de publicaciones ya hechas, hay que considerar las reglas pertinentes para su utilización o divulgación, esto con el fin de proteger los derechos de propiedad intelectual que corresponde a los informantes o a los autores de los textos que conforman el corpus.

Se ha discutido en varios foros si en todos los casos hay que solicitar permiso para usar parte o la totalidad de un texto. Es bien sabido el concepto y las consecuencias del plagio, razón por la que se debe citar el extracto de una obra, normalmente dando referencia a la fuente. Sin embargo, ahora que se aplican con mayor rigor leyes internacionales para respetar los derechos de autor, de manera que se considera ilegal hacer fotocopias de libros y revistas, surgen dudas naturales acerca de la validez de usar un extracto de texto y mostrarlo como ejemplo a los alumnos en una clase impartida con una red de cómputo. De similar manera, uno llega a cuestionarse la posibilidad de extraer un fragmento que servirá únicamente para los fines de un trabajo académico, en donde sólo aparecerán los resultados de la investigación, pero nunca como tal los fragmentos extraídos.

El principio básico de la propiedad intelectual es respetar y proteger la obra de los autores, de manera que cualquier forma de uso, en este caso desde la recopilación del material hasta la distribución del mismo, si se va a hacer del dominio público la explotación de los textos, requiere del consentimiento del titular de los derechos de la obra. Es de particular relevancia para un corpus, en donde justo la materia prima la constituye la obra de los autores, que se cumpla con las reglas establecidas para el acceso a la información y con los derechos de autor para evitar el mal uso.

Aunque el tema es muy controvertido y podría llevar mucho tiempo discutirlo, conviene mencionar algunas normas que aplican en la recopilación de textos, orales o escritos, para la construcción de un corpus, sea cual sea su fin:

Tener el consentimiento del poseedor de los derechos. No se deben publicar los documentos que lo forman total ni parcialmente, por ningún medio, incluyendo el electrónico, sin consentimiento del propietario del derecho de autor. En algunos casos hay que identificar quién es el poseedor de los derechos: si el autor, el editor, el publicista o el traductor, sea una persona física o moral, razón por la cual es conveniente mantener registros en bases de datos como la ejemplificada en la figura 1.1. En esta figura se muestra un ejemplo de una de las fichas del Translational English Corpus (TEC), cuya mira central está orientada a la recopilación de textos traducidos al inglés y, por tanto, involucra al menos tres datos necesarios: el idioma del texto original, el autor y el traductor. De estos dos últimos se registran datos que no se muestran en esta ficha, pero aquí se puede observar otra información usada para cuestiones administrativas, tales como el publicista, el poseedor de los derechos de autor y las fechas en que se solicitó y se obtuvo permiso para utilizar el material. La labor puede ser titánica y requerir un encargado de tiempo completo para dedicarse a cuestiones administrativas y de relaciones públicas.

Obtener cartas de autorización. Las cartas deberán ser para el uso y publicación de los textos del corpus, pues de ninguna manera debe dejarse a convenios verbales, pues no son válidos y son revocables. Las normas indican que la autorización para usar la obra debe ser previa, explícita y por escrito por el titular de los derechos. Hacerlo previamente puede ahorrar mucho trabajo, sobre todo cuando hablamos de libros que hay que digitalizar, limpiar y procesar para adecuarlos a los formatos del corpus.

Explicitar el uso. El acuerdo explícito asentado en la autorización será el tipo de uso que se le va a dar a la obra. Por ejemplo, si se va a hacer uso de los textos para una investigación particular, con acceso restringido, o si se va a permitir el acceso de los textos a todo el público. Asimismo, si sólo se permitirá el acceso a parte de los documentos y a su totalidad. O bien, si el acceso será en línea con lo que permita la interfaz de consulta o se podrán descargar los textos.

Citar las fuentes. Independiente a los permisos para explotar una obra, deben darse referencias explícitas a las fuentes de cada uno de los textos, señalando los datos bibliográficos pertinentes, incluso para los textos obtenidos de Internet que, aunque sean del dominio público, también deben ser referenciados y, en los casos que corresponda, solicitarse derechos para el uso de los textos. Según el tipo de resultados que se pretenda desplegar, a nivel de concordancias se indicará la fuente del fragmento, en tanto para cuestiones generales, como el conteo de palabras, se puede tener un apartado con el listado de todas las fuentes del corpus.

Existen excepciones o límites a las normas para las que no se requiere solicitar autorización del uso de la obra. Una de ellas es pertinente cuando se trata de trabajos

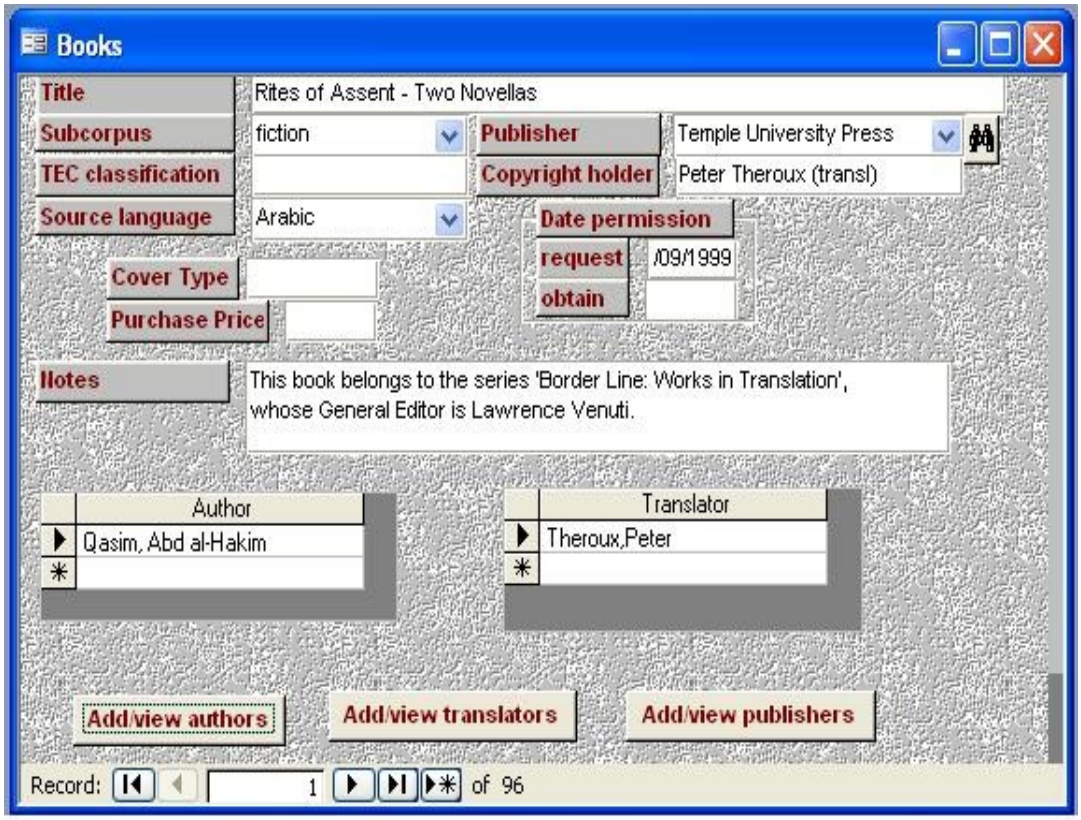


Figura 1.1: Ejemplo de la base de datos del Translational English Corpus.

de investigación o docencia sin fines de lucro, siempre y cuando no se permita el acceso más allá de fragmentos de textos y se indique claramente la procedencia de los mismos. Ante este rubro de uso honrado se han manejado varios corpus, sobretodo cuando se trata de corpus de referencia en donde se recopilan partes pero no la totalidad de múltiples obras. Otra excepción muy útil para corpus históricos es que la posesión de los derechos se conserva hasta 70 años después del fallecimiento del autor.

Para cualquiera que sea el fin de usar corpus, se busca tener datos reales. Se intenta que los hablantes, en el caso de corpus de procedencia oral, no modifiquen su estilo de habla y que, por tanto, no estén conscientes de que son observados, con el fin de que los datos obtenidos del corpus se aproximen a la lengua real y sean una fuente que proporcione información más acertada acerca del uso lingüístico. Por eso, en corpus orales, es preferible solicitar permiso al hablante una vez que ha terminado la entrevista

o grabación, a fin de no perder espontaneidad en los actos del habla. Lo mismo es aplicable en textos escritos como correos electrónicos y pláticas electrónicas o chats, por ejemplo. Cuando se trata grabaciones personales de esta naturaleza, en muchas ocasiones el autor busca proteger su privacidad, por lo que podrá solicitar que no se proporcione su nombre; además, para los fines lingüísticos lo importante será anotar los datos del informante: edad, sexo, escolaridad, etc.

Ahora bien, en cuanto al tema de derechos de autor cabe señalar que una vez construido un corpus también éste está protegido como propiedad intelectual, por lo que es importante reconocer el trabajo del equipo (diseñadores, programadores, capturistas, digitalizadores, etc.) y agradecer al patrocinador. A fin de justificar con el patrocinador el monto del dinero invertido se requiere obtener datos sobre los tipos de usuarios y tener estadísticas de uso. Se puede solicitar un registro a los usuarios y pedir que confirmen estar de acuerdo con los derechos de autor y establecer cláusulas a conveniencia; por ejemplo, si se quiere utilizar un fragmento de texto obtenido en una consulta, se debe dar la referencia al corpus, más que a la fuente misma correspondiente al texto.

1.6. Referencias

Lecturas sugeridas

Biber, Douglas (1993). ”Representativeness in Corpus Design“. Literary and Linguistic Computing 8 (4), pp. 243-257.

McEnery, Tony y Andrew Wilson (1996). Corpus Linguistics. Edinburgh: Edinburgh University Press. (Véanse capítulo 2 y sección 3.3).

MacMullen, John (2003). Requirements Definition and Design Criteria for Test Corpora in Information Science. SILS Technical Report 2003-03, School of Information and Library Science, University of North Carolina at Chapel Hill. (Véanse secciones 4.1 y 4.2).

Torruela, Joan y Joaquim Llisterri (1999). ”Diseño de corpus textuales y orales“. En J.M. Blecua et al. (Eds.), Filología e informática: Nuevas tecnologías en los estudios filológicos. Barcelona: Editorial Milenio-Universidad Autónoma de Barcelona, pp.45-77. (Véase sección 2).

Villayandre Llamazares, Milka (2008). Lingüística con corpus (I). Estudios Humanísticos. Filología 30, pp. 329-349.

Capítulo 2

Descripción de corpus existentes

La historia de la lingüística de corpus se remonta a mediados del siglo XX. A continuación se ofrece una breve cronología de sus orígenes y posteriormente, una descripción de los corpus más representativos para el español.

Index Tomisticus. El primer registro del uso de la informática en la elaboración de un corpus se tiene en Italia por quien se conoce como fundador de las Humanidades Digitales, el Padre Roberto Busa, en 1949, que transcribió la obra de Santo Tomás de Aquino y otros autores en tarjetas perforadas. Mediante un programa de cómputo obtuvo las concordancias de las más de once millones de palabras. La obra fue publicada en 1974 en 56 tomos y vertida en CD-ROM para 1992.

Survey of English Usage. En 1960 se llevó a cabo en el University College of London el Survey of English Usage (SEU) a cargo de Randolph Quirk, con el fin de proveer de recursos para la descripción de la gramática de los hablantes adultos cultos del inglés. Para ello, obtuvo 200 muestras representativas del inglés británico hablado y escrito, material que fue transcrito en cintas, impreso en papel e indizado en fichas de cartón. El corpus fue anotado con etiquetas prosódicas, gramaticales y paralingüísticas en 1964. Este corpus junto con el Survey of Spoken English (SSE), iniciado en 1975, fue la base para el London-Lund Corpus of Spoken English (LLC), creado en 1980.

Brown Corpus. En paralelo al SEU, a principio de la década de 1960, Henry Kučera y Nelson Francis trabajaron en la construcción del Brown University Standard Corpus of Present-Day American English, mejor conocido como Brown Corpus. Cuenta con 500 muestras del inglés de Estados Unidos, distribuidas en 15 categorías, dando un total aproximado de un millón de palabras. La versión anotada se hizo en 1979, con 82 etiquetas de partes de la oración.

2.1. Corpus del Español Mexicano Contemporáneo (CEMC)

El primer corpus informatizado en español, que además cumple con las características de un corpus bien construido, es el Corpus del Español Mexicano Contemporáneo (CEMC). El CEMC comenzó a elaborarse en 1973, bajo la coordinación de Luis Fernando Lara en El Colegio de México, creado para extraer los términos del Diccionario del Español de México.

El corpus está formado por un conjunto de 996 fragmentos de textos escritos y transcripciones de conversaciones grabadas, todos de autores mexicanos desde 1921 hasta 1974 procedentes del Atlas lingüístico de México y de otros acervos lingüístico-etnográficos de la biblioteca de El Colegio de México. Los fragmentos fueron escogidos aleatoriamente de las 996 obras del corpus, y tienen una extensión aproximada de 2,000 palabras cada uno.

Una peculiaridad del corpus es que se compone por fragmentos y no por obras completas, a fin de que sea equilibrado y representativo. Los fragmentos están clasificados por 14 géneros que corresponden al “lenguaje culto” y “lenguaje popular”: obras de literatura, textos científicos, académicos y técnicos, discurso político, habla coloquial, etc.

A partir del corpus se obtuvieron datos cuantitativos de cada una de las palabras, tales como la frecuencia absoluta, frecuencia relativa en cada género, frecuencia relativa entre géneros, frecuencia corregida e índice de dispersión, etc.

Actualmente el CEMC cuenta con una interfaz para la extracción de concordancias y de estadísticas, creada por el Grupo de Ingeniería Lingüística.

2.2. Corpus Diacrónico del Español (CORDE)

El CORDE comenzó a crearse por la Real Academia Española de la Lengua (RAE), en 1994, con la finalidad de facilitar el acceso a su banco de datos por medio de la aplicación de técnicas informáticas.

Su contenido abarca documentos de todas las épocas y lugares en que se habla español, desde los inicios del idioma hasta el año 1975, donde limita con el Corpus de Referencia del Español Actual. Actualmente, el corpus cuenta con 250 millones de palabras e incluye textos procedentes de diversos países de habla hispana.

El corpus pretende recoger en lo posible todas las variedades geográficas, históricas y genéricas para que el conjunto sea suficientemente representativo. Por ejemplo, se

distribuyen en prosa (85 %) y verso (15 %); procedentes de libros (97 %) y prensa (3 %); de origen español (74 %), hispanoamericano (25 %), español sefardí y otros (1 %). Dentro de cada modalidad se dividen en textos narrativos, líricos, dramáticos, científico-técnicos, históricos, jurídicos, religiosos, periodísticos, etc.

La Academia utiliza sistemáticamente el CORDE para documentar palabras, calificarlas de anticuadas o en desuso, saber el origen de algunos términos, su tradición en la lengua, determinar las primeras apariciones de palabras, entre otras. Pero uno de los cometidos fundamentales del CORDE es la confección del Diccionario histórico de la lengua española.

La interfaz en línea permite recuperar concordancias, documentos, párrafos y agrupaciones en los que aparece una consulta. Las búsquedas se pueden filtrar por autor, fecha, tema, medio de obtención del material y por país. Asimismo, permite obtener estadísticas por año, país y tema.

2.3. Corpus de Referencia del Español Actual (CREA)

El CREA fue desarrollado por la RAE, y cuenta actualmente con más de 160 millones de palabras. Tiene una gran variedad de textos escritos y transcripciones orales, producidos en todos los países donde se habla español desde 1975 hasta la actualidad.

El corpus contiene textos completos clasificados por su origen (escritos - 90 %, orales - 10 %), por la fuente (libros - 49 %, prensa - 49 %, miscelánea - 2 %), por su distribución geográfica (España - 50 %, América - 50 %) y por áreas temáticas (los textos de corpus abarcan más de 100 materias diferentes).

La interfaz en línea es idéntica a la del CORDE y las consultas se pueden filtrar de la misma manera.

2.4. Corpus del Español de Mark Davies

El Corpus del Español fue creado por Mark Davies, profesor de la Brigham Young University, en septiembre de 2002, con motivo de superar las limitaciones de búsqueda que presentaba el CORDE. Fue financiado por la National Endowment for the Humanities (NEH-USA).

Cuenta con 100 millones de palabras procedentes de textos literarios, orales, periódicos y enciclopédicos de los siglos XIII-XX. Permite realizar búsquedas avanzadas a

partir de partes de la oración, lemas, sinónimos, y frecuencias de palabras.

Algunas de las fuentes de donde se han recogido textos para el corpus se encuentran el Archivo Digital de Manuscritos y Textos Españoles (ADMYTE), la Biblioteca Virtual Miguel de Cervantes, el Electronic Texts and Concordances of the Madison Corpus of Early Spanish Manuscripts and Printings, y la Enciclopedia Encarta, entre las más importantes. Parte de la diversidad de documentos que se tienen son: literatura (novelas, cuentos, obras de drama) textos orales (transcripciones de congresos, entrevistas periodísticas) y documentos misceláneos (encicpledias y periódicos).

El número de palabras que contiene, de acuerdo con la época, se detalla en la tabla 2.1.

Periodo temporal	**Número de palabras** (redondeado)
Siglos XIII a XV	20 millones
Siglos XVI a XVIII	40 millones
Siglos XIX	20 millones
Siglos XX	20 millones

Tabla 2.1: Características del Corpus del Español de Mark Davies.

Lo particular de este corpus es el sistema de almacenamiento de datos con arquitectura abierta que permite tener varias bases de datos relacionadas, gracias a lo cual es posible realizar búsquedas por forma, lema, categoría gramatical y sus combinaciones, en el siglo que se desee. Además de que se pueden hacer búsquedas de palabras o frases exactas, se pueden utilizar los comodines tradicionales, como el asterisco y el signo de interrogación cerrada. Por otra parte, cuenta con una base de datos adicional que contiene 30,000 palabras sinónimas, lo que permite ampliar una búsqueda no solo con la palabra sino con sus sinónimos.

Los textos del corpus están almacenados en una base de datos relacional de SQL Server 7.0. La información sobre la anotación se almacena en otras bases de datos. La base de datos central contiene todos los unigramas, bigramas y trigramas de las palabras del corpus, con sus respectivas frecuencias de uso en cada siglo. Esta base de datos a su vez está relacionada con otras en las que se guarda la información lingüística como lemas y categorías gramaticales de todas las palabras del corpus.

2.5. Corpus textual especializado plurilingüe

Este corpus técnico fue diseñado en el Instituto Universitario de Lingüística Aplicada (IULA) de la Universidad Pompeu Fabra, de Barcelona. Constituye el soporte principal

de las actividades de investigación y docencia de dicho instituto. Recopila textos escritos en cinco lenguas diferentes: catalán, español, inglés, francés y alemán; de las áreas de especialidad de derecho, economía, medio ambiente, informática y medicina. Las ocurrencias de las palabras por área y lengua se pueden apreciar en la tabla 2.2.

Area	Lengua					Total
	Catalán	Español	Inglés	Francés	Alemán	
Derecho	1,463,000	2,085,000	431,000	44,000	16,000	4,039
Economía	1,776,000	1,091,000	274,000	78,000	27,000	3,246,000
M. Ambiente	1,506,000	1,062,000	599,000	230,000	429,000	3,826,000
Informática	655,000	1,227,000	338,000	194,000	83,000	2,497,000
Medicina	2,619,000	4,077,000	1,555,000	27,000	198,000	8,476,000
Total	8,019,000	9,542,000	3,197,000	573,000	753,000	22,084,000

Tabla 2.2: Características del Corpus textual especializado plurilingüe del IULA.

A través del establecimiento del corpus, se pretende inferir las reglas que rigen el comportamiento de cada lengua en las áreas mencionadas y realizar detección de neologismos y términos, estudios sobre variación lingüística, análisis sintáctico parcial, alineación de textos, extracción de datos para la enseñanza de segundas lenguas, extracción de datos para la construcción de diccionarios electrónicos, elaboración de tesauros, etc. El corpus ya ha sido utilizado para realizar numerosos estudios en terminología, neología, morfología y sintaxis, discurso y traducción.

Los textos son seleccionados por especialistas de cada área y agrupados de acuerdo con una clasificación temática y de uso propuesta por los propios especialistas; posteriormente se procesan en varias etapas: marcaje estructural, pre-procesamiento automático (identificación de fechas, números, nombres propios, abreviaturas, etc.), marcaje con los etiquetarios del IULA (lematización y etiquetado de partes de la oración) y finalmente, desambigüación lingüística y estadística.

El corpus puede consultarse a través del Internet con ayuda de la herramienta de consulta BwanaNet diseñada especialmente para este corpus. BwanaNet ofrece varias posibilidades de búsqueda. Primero, se puede realizar la búsqueda de unidades fuera de contexto. En este caso la herramienta genera una lista de formas, lemas, o etiquetas POS del subcorpus seleccionado por el usuario. Segundo, se puede realizar una búsqueda por frecuencia, la cual permite obtener una lista de frecuencias de unidades (formas, lemas o etiquetas de partes de la oración) o secuencias de unidades de todo el corpus. Tercero, BwanaNet permite buscar concordancias simple, estándar y compleja. La opción “concordancia simple” ofrece el contexto de aparición de un lema o forma en específico. En el caso de concordancia estándar se puede buscar apariciones de formas, lemas y etiquetas de partes de la oración de manera combinada, mientras que la opción “concordancia

compleja" amplía todavía más estas posibilidades, ya que permite buscar cualquier combinación de cualquier cantidad de elementos junto con su frecuencia de aparición.

2.6. Corpus del Grupo de Ingeniería Lingüística (GIL)

El Grupo de Ingeniería Lingüística, en el Instituto de Ingeniería de la UNAM, cuenta con cinco corpus propios: el Corpus Lingüístico en Ingeniería (CLI), el Corpus de las Sexualidades en México (CSMX), el Corpus Histórico del Español de México (CHEM), el Corpus de Contextos Definitorios (CORCODE) y el RST Spanish Treebank. Estos corpus fueron construidos con el patrocinio de la DGAPA-UNAM y de CONACyT.

Corpus Lingüístico en Ingeniería (CLI)

El CLI es una colección abierta de textos electrónicos en español, representativos de todas las áreas de la ingeniería, tales como la eléctrica, electrónica, civil, mecánica, etc. Cuenta con medio millón de palabras, aproximadamente, y es el primer corpus lingüístico en ingeniería para el español de México. En términos generales, este corpus proporciona la información necesaria sobre el uso del lenguaje en esta área de especialidad y puede servir como recurso para la extracción terminológica y construcción de diccionarios especializados.

En la construcción de este corpus se prestó especial atención a su representatividad y equilibrio. Se desarrolló una base de datos bibliográfica que contiene, por un lado, la información sobre el origen de los documentos y, por otro, el número de palabras o tokens y el número de palabras únicas o types de cada texto. Esto permite controlar la cantidad y el tamaño de documentos por el tipo de texto y área temática, para garantizar, de esta manera, el equilibrio y la representatividad del corpus.

El corpus está anotado en formato XML. Los textos pasaron por una etapa de pre-procesamiento, que consistió en marcar unidades como nombres propios, fechas, números, abreviaturas, siglas, etc., con el fin de facilitar el posterior procesamiento automático del corpus. Asimismo, fue etiquetada la estructura, las partes de la oración (con el estándar EAGLES) y las marcas tipográficas y de estilo de los textos.

Para este corpus se están creando herramientas de análisis, algunas de las cuales están disponibles a través del Internet. Estas herramientas están diseñadas para dar información estadística sobre el corpus (frecuencias absoluta y relativa y entropía de unigramas y bigramas), generar concordancias y medidas de asociación entre palabras, así como determinar las colocaciones típicas por documento, por área temática o para todo el corpus.

Corpus de las Sexualidades en México (CSMX)

La construcción del CSMX se hizo en el GIL, en miras a un proyecto en colaboración con El Colegio de México, Acciones Voluntarias sobre la Educación en México y el Grupo Interdisciplinario de Sexología, A.C. El objetivo principal de este proyecto es desarrollar un modelo de extracción de conocimiento lexicográfico de áreas de especialidad a partir del Internet, a fin de servir como base para la extracción automática del vocabulario básico de las sexualidades y la creación de varios diccionarios especializados: Diccionario Básico de las Sexualidades en México, Diccionario de las Sexualidades en México, Diccionario Enciclopédico de las Sexualidades en México y un Diccionario Multimedia de las Sexualidades en México.

El diseño del corpus está basado en la metodología de compilación y explotación de otros corpus desarrollados en el GIL: el CLI y el CHEM. Se tomaron en cuenta los tres criterios establecidos para la construcción del corpus: variedad, representatividad y equilibrio.

Los textos del corpus fueron extraídos de la web, y se dividen, por su origen, en cuatro niveles diastráticos que abarcan artículos científicos, documentos de asociaciones sobre el tema y foros de discusión. Asimismo, están clasificados por ocho subáreas, definidas de acuerdo con la clasificación de la biblioteca del Instituto Kinsey de Sexualidad; estas subáreas son:

- Fundamentos biológicos de la sexualidad.
- Respuesta y la expresión sexual.
- Comportamiento sexual.
- Identidad sexual.
- Enfermedades de transmisión sexual.
- Sexualidad variante.
- Atracción sexual y relaciones entre individuos.
- Educación sexual, cultura y sexualidad social.

Así, el corpus no solamente presenta variedad temática y refleja la organización del conocimiento en el campo de la sexualidad, sino que también es representativo en cuanto al uso del lenguaje en distintos géneros, registros y tipos de textos.

Corpus Histórico del Español en México (CHEM)

El CHEM está constituido por documentos diacrónicos del español, de los siglos XVI-XIX, procedentes del territorio mexicano desde que era la Nueva España.

Para la construcción del CHEM se usaron bases de datos relacionales como componentes fundamentales de la arquitectura. A partir de los textos del CHEM se modeló y creó una base de datos relacional que tiene dos componentes principales: tablas con información bibliográfica y un conjunto de tablas de unigramas con la información sobre su ubicación en el corpus, su frecuencia de aparición en cada siglo y sus características lingüísticas.

En su primera versión el corpus tiene un total de 320 textos tomados de los Documentos Lingüísticos de la Nueva España, editados por Concepción Company, del Instituto de Investigaciones Filológicas de la UNAM.

Los textos están etiquetados en formato XML. Cada documento tiene un encabezado que indica la fuente del documento y proporciona información sobre el autor (origen, género, grupo social), tipo de texto (carta, testimonio, etc.) y sobre el origen del documento (oral o escrito).

Debido a que se trata de un corpus diacrónico y a que la ortografía ha variado a lo largo del tiempo, se hizo un etiquetado fonológico que asocia palabras con distintas grafías a una misma representación fonológica. Dicho de otro modo, al hacer una consulta, el corpus presentará el resultado de un mismo registro (palabra) escrito con las diferentes ortografías que ha tenido a lo largo de la historia.

La interfaz en línea permite realizar búsquedas por lemas y patrones gramaticales. Por ejemplo, al buscar “lema (dar) + artículo + sustantivo”, el usuario obtiene resultados como “dar la mano”. Para facilitar el proceso de etiquetado se están desarrollando un lematizador y un parser automático adaptados para el español mexicano del siglo XVI.

Algunas ventajas del CHEM sobre otros corpus similares son que en éste, para identificar los types, se toman en cuenta el nivel fonológico, es decir, se evitan las dificultades que pueden ocasionar las diferencias ortográficas de las diferentes épocas.

Corpus de Contextos Definitorios (CORCODE)

El CORCODE fue desarrollado en el marco del proyecto Extracción automática de definiciones en textos de especialidad. Se compone de 127 contextos definitorios, esto es, fragmentos textuales que unen términos con definiciones por medio de patrones

verbales y aportan información que permite comprender el significado de un término, su función y sus relaciones con otros términos.

Uno de los objetivos de este tipo de corpus es estudiar los diferentes tipos de estructuras sintácticas y marcadores pragmáticos que caracterizan la introducción de nuevos términos, para la extracción automática de información terminológica y conceptual.

Los documentos de los que fueron extraídos los contextos definitorios son especializados en los ámbitos científico, académico y técnico, y fueron tomados de otros corpus ya existentes: el CLI y el corpus textual especializado plurilingüe del IULA, entre otros.

El corpus permite consultas con base en el tipo de término, de definición (analítica, sinonímica, funcional, extensional), de patrones definitorios (tipográficos, sintácticos, pragmáticos) y de patrones pragmáticos (autoría, patrones temporales o instruccionales).

Es un corpus etiquetado en XML. Los encabezados de los documentos contienen información sobre la fuente de extracción, la fecha en la que el documento fue etiquetado e integrado al corpus, el verbo definitorio y el tipo de definición. La anotación de los mismos contextos definitorios tiene etiquetas XML que ayudan a identificar los patrones tipográficos, sintácticos y pragmáticos de las definiciones.

RST Spanish Treebank

El RST Spanish Treebank fue creado en colaboración con la Universidad Pompeu Fabra. Fue dirigido por la Dra. Iria da Cunha Fanego. Se trata de un corpus anotado con la Rhetorical Structure Theory (RST), una teoría que permite describir la estructura discursiva de un texto de manera jerárquica, es decir en forma de árbol. Este corpus fue diseñado para el desarrollo de un parser (analizador) discursivo automático para el español, así como otras aplicaciones del análisis del discurso en lingüística computacional (generación de texto, resumen automático, extracción de información, traducción automática, etc.).

El corpus abarca diversas áreas como psicología, matemáticas, lingüística, medicina, sexualidad, etc. y tanto las áreas temáticas como el número de documentos que conforman el corpus están en constante evolución, ya que cualquier usuario puede subir sus propios árboles discursivos a la web.

El RST Spanish Treebank tiene una metodología de anotación discursiva clara, explícita y sistemática. Es el primer corpus anotado con relaciones discursivas para el español.

La anotación del corpus fue realizada de manera manual (con ayuda de la interfaz gráfica RSTtool) por 10 personas previamente entrenadas para la tarea. Primero, los textos fueron segmentados en unidades discursivas mínimas, después se detectaron las relaciones discursivas, y finalmente se construyeron representaciones arbóreas que reflejan la estructura discursiva de los textos.

El corpus se compone de textos cortos y especializados en los ámbitos mencionados. Tiene un total de 267 textos y 52,746 palabras. Se procuró que tuviera suficientes ejemplos de cada relación discursiva (aproximadamente 20 ejemplos para cada relación).

El corpus está disponible a través del Internet y tiene una interfaz para consultas que permite descargar los textos en formato txt (título, referencias e hiperligas) y los árboles discursivos en formato rs3 y jpg. Asimismo, el usuario puede seleccionar un subcorpus por áreas de especialidad y obtener estadísticas de relaciones discursivas a partir del subcorpus seleccionado.

2.7. Corpus apoyados por el Grupo de Ingeniería Lingüística

Dada la experiencia en los corpus anteriores del Grupo de Ingeniería Lingüística, éste ha apoyado además en el desarrollo de otros corpus: el Corpus del Español Mexicano Contemporáneo (ya descrito anteriormente), el Corpus Electrónico para la Enseñanza de la Lengua Escrita (CEELE), el Corpus Científico del Español de México (COCIEM) y el Corpus de ad-hocracia (CAC).

Corpus Electrónico para la Enseñanza de la Lengua Escrita (CEELE)

El CEELE fue desarrollado en el marco del proyecto Aprender mientras se enseña: una experiencia de acompañamiento en la enseñanza de la lengua escrita, organizado por la Coordinación Estatal del Programa Nacional de Lectura, dependiente de la Dirección de Educación Básica de los Servicios de Educación Pública del Estado de Nayarit (SEPEN), en colaboración con investigadores de la Universidad Nacional Autónoma de México (UNAM), de la Universidad Autónoma Metropolitana Xochimilco (UAMX) y de la Escuela Nacional de Antropología e Historia (ENAH).

El objetivo general del proyecto es mejorar la calidad de la educación básica. Una parte fundamental del proyecto fue la capacitación de docentes. La capacitación se llevó a cabo con base en los recientes avances de investigación psicológica,ación Primaria del Estado de Nayarit. Para evaluar el impacto de esta capacitación se realizó un ejercicio con los alumnos: se leyó un cuento sobre una niña africana llamada Fátima.

Después, se pidió a cada uno de los niños que escribiera una pequeña narración sobre la vida de la niña. El ejercicio también se llevó a cabo en escuelas donde los docentes no tuvieron capacitación, a fin de poder comparar los textos y evaluar el impacto de la capacitación de los maestros en la adquisición de lectoescritura en los niños. De este experimento se obtuvieron los textos que constituyen el corpus: un total de 300 textos escritos por niños de 6 a 8 años, cada texto de una a seis cuartillas.

Los textos del corpus fueron digitalizados en formato txt. La digitalización se realizó pensando por una parte, en facilitar la lectura y el análisis y, por otra parte, procurar que la transliteración sea fiel a los textos "originales". Se preservó la estructura que los niños dieron a los textos (espacios entre letras, palabras, renglones y párrafos), el uso de minúsculas y mayúsculas, de signos de puntuación, y las autocorrecciones que hicieron al escribir. Este tipo de información es muy valiosa para la investigación de corte lingüístico, ya que evidencia el concepto de escritura que tienen los niños y su reflexión acerca de ella.

Además de simplemente transliterar los textos, se hicieron las versiones normalizadas de éstos. En dichas versiones se corrigieron las faltas de ortografía y las omisiones. La normalización permitió que el corpus fuera más legible. La comparación de los textos transliterados y sus versiones normalizadas permite observar fenómenos de interés para lingüística, pedagógica y didáctica de la lengua. Se capacitaron 180 docentes en 100 grupos de primer grado y 80 grupos de segundo grado de los 7 Sectores Escolares de Educun análisis lingüístico.

El corpus está etiquetado en XML. El etiquetado proporciona información sobre cada niño (nombre, sexo, edad, escuela, etc.) en el encabezado. Se decidió incluir esta información porque demuestra la influencia del entorno social en la adquisición de lectoescritura. En el cuerpo de los textos se etiquetaron los límites de palabras, ya que uno de los temas de interés para el análisis del discurso infantil escrito es la hipersegmentación (palabras segmentadas de más) y la hiposegmentación (varias palabras pegadas formando una sola). Se utilizaron etiquetas especiales para el uso de mayúsculas apropiado/no apropiado, ya que reflejan la idea que se forman los niños sobre este tipo de marcas y su función en el texto; asimismo, se introdujeron etiquetas para marcar autocorrecciones, signos de puntuación, dibujos, comentarios, etc.

Corpus Científico del Español de México (COCIEM)

El COCIEM es un corpus creado en El Colegio de México, en colaboración con el GIL. La directora del corpus es la Dra. María Pozzi. Este corpus fue diseñado para estudiar el aspecto léxico del lenguaje científico básico del español de México, es decir, qué parte del lenguaje científico (conjunto de lenguajes especializados que representan y

transmiten el conocimiento científico) idealmente debería conocer un hablante promedio.

El objetivo principal de este corpus es proporcionar una base empírica para la descripción y el análisis de las características morfológicas, sintácticas y semánticas de este lenguaje (por ejemplo, modelos más productivos de formación de términos en diferentes ciencias y a diferentes niveles de la lengua, patrones sintácticos en términos multipalabra, variación conceptual y denominativa en el uso de los términos). Asimismo, otro de sus objetivos es el estudio sistemático de estas características para construir un modelo lingüístico completo del vocabulario científico básico en español de México.

El corpus incluye una amplia variedad de textos científicos mexicanos de todas las áreas de la ciencia y a diferentes niveles de especialización. Los textos del corpus se clasificaron, primero, por áreas y subáreas temáticas, y segundo, por niveles académicos (primaria, secundaria, preparatoria) y de especialización (divulgación, educación universitaria, textos científicos especializados).

El etiquetado del corpus proporciona información sobre el origen de los documentos, la estructura de los textos, y la información morfosintáctica (lematización y y etiquetado de partes de la oración).

En la etapa actual el corpus cuenta con 89 libros de texto (al menos, dos libros por cada ciencia) clasificados por niveles académicos (primaria, secundaria y preparatoria) y por áreas científicas. Así, el COCIEM contiene el vocabulario científico básico que dominan los alumnos al terminar la educación media superior.

Corpus de ad-hocracia (CAC)

El CAC fue desarrollado por la Dra. Margarita Palacios, con el propósito de estudiar la construcción discursiva del concepto 'democracia' a través de su uso en el Congreso de la Unión de México. En principio, el corpus se creó en el marco de una investigación de corte pragmático y discursivo, pero puede ser de gran utilidad para indagar en otros aspectos de la lengua. El término democracia tiene un uso cotidiano y una definición científica; se refiere, por tanto, a un concepto amplio y difuso, lo cual hace posible su aparición en contextos diferentes con valoraciones y matices muy distintas que reflejan las representaciones sociales de los hablantes sobre dicho término. Esta variedad de usos y contextos resulta en una "Ad hoc-cracia" o una democracia para la ocasión (de ahí el nombre del corpus). El CAC tiene dos partes. La primera está conformada por los documentos curriculares de los 115 senadores y 500 diputados de la LX Legislatura (septiembre 2006-enero 2009) y la segunda, por las transcripciones de los debates sostenidos en las sesiones del 1 de febrero al 13 de agosto (senadores) y del 4 de enero al 20 de junio (diputados) del 2008 (38 sesiones de la Cámara de Diputados y 45 de la

Cámara de Senadores). Los archivos correspondientes a las sesiones en la Cámara de Diputados tienen una extensión de 982,834 palabras y las sesiones de la Cámara de Senadores 920,756 palabras; el corpus tiene un total de 1'903,590 palabras. Todos los textos del corpus fueron extraídos de la página web oficial del Congreso de la Unión.

La organización del corpus tiene ciertas ventajas de cara a las necesidades de la investigación en pragmática y análisis del discurso. Primero, el corpus contiene textos escritos (documentos curriculares) y muestras de la lengua oral (transcripciones de las sesiones de debates) con las intervenciones tanto premeditadas como espontáneas de los legisladores, lo cual le da una riqueza en géneros y registros. Segundo, la inclusión de los documentos curriculares de los diputados y senadores permite relacionar, de manera sistemática, el uso de la lengua con la información sobre los hablantes.

El corpus tiene etiquetas en Access y en XML que proporcionan información sobre los diputados y senadores (nombre, partido, fecha de nacimiento, tipo de elección, comisiones a las que pertenece, grado de escolaridad, trayectoria política, experiencia legislativa, etc.); y sobre la estructura de las sesiones de debates. Las transcripciones de las sesiones de debates fueron segmentadas por unidades discursivas (intervención, turno de los actores). Se identificaron los segmentos en los que aparece el lexema democr. Estos segmentos fueron etiquetados con categorías gramaticales y con el tipo de oración activa/pasiva, esto último a fin de capturar el tema de interés.

2.8. Corpus Diálogos Inteligentes Multimodales en Español (DIME) y DIMEx100

Estos dos corpus fueron diseñados dentro del marco del proyecto DIME (Diálogos Inteligentes Multimodales en Español) que se lleva a cabo en el Instituto de Investigaciones Matemáticas de la UNAM y tiene por objetivo el desarrollo de sistemas conversacionales para el español hablado en México en contextos lingüísticos especializados.

DIME

Uno de los propósitos de la Inteligencia Artificial es modelar la interacción humana. Las mayores dificultades que se presentan a la hora de realizar esta tarea son la ambigüedad de las lenguas humanas y el conocimiento del mundo que se necesita para la interacción lingüística. De ahí que para la construcción de un sistema automático capaz de dialogar con un ser humano existan restricciones en cuanto al tema y al propósito de la interacción que deben ser específicos y establecidos de antemano. Para modelar la interacción en estas condiciones es necesario disponer de

material empírico tanto sobre las particularidades de la tarea que se va a cumplir a través del diálogo como sobre el lenguaje que se utiliza en el ámbito. Así, en el marco del proyecto DIME surgió la necesidad de desarrollar un corpus con estas características.

Para la recolección de datos se realizaron experimentos Mago de Oz, en los que una persona juega el papel del sistema computacional, y otras personas, el papel del usuario. Los participantes colaboran a través del diálogo y de una interfaz gráfica, con un propósito determinado. En el caso del DIME este propósito fue diseñar una cocina. La información obtenida durante los experimentos fue registrada en formato de video, audio y transcripción.

El corpus cuenta con 27,459 palabras, 5,779 enunciados y 3,606 turnos, de un total de siete horas con diez minutos grabados. Las transcripciones de los diálogos fueron segmentadas por turnos y enunciados. Para establecer los límites de los enunciados se tomaron en cuenta criterios prosódicos y sintácticos. Asimismo, las transcripciones fueron etiquetadas a niveles ortográfico, fónico, morfosintáctico y pragmático. En cuanto al nivel fónico, cabe señalar que se marcaron los alófonos, las sílabas fonéticas, y la entonación (índices de ruptura y curvas de entonación), teniendo en cuenta las propiedades específicas del español hablado en México. A nivel pragmático, se etiquetaron actos de habla y expresiones referenciales.

El corpus DIME es oral transcrito, no premeditado, con anotación fónica (fonética y prosódica), textual (ortográfica), morfosintáctica (POS tagging), pragmática, general, de autoría variada, sincrónico contemporáneo, monolingüe, pequeño, para fines específicos, desequilibrado, público, documentado y oportunista.

DIMEx100

Este corpus fue desarrollado con miras a la construcción de modelos acústicos y diccionarios de pronunciación para la variante mexicana del español que constituirían una base empírica tanto para la investigación en fonética y fonología como para diversas aplicaciones prácticas de las tecnologías del lenguaje. Para la construcción de los sistemas de reconocimiento y síntesis del habla debe tomarse en cuenta la variante de la lengua en específico, además de diccionarios de pronunciación. Dado que no se había realizado mucho trabajo en esta dirección para el español mexicano, surgió la necesidad de tener un corpus fonético para esta variante del español. Para el corpus DIMEx100 se investigaron las características del habla de la parte central de México.

Para construir este corpus, primero, se recolectó material de la web, lo que resultó en el Corpus230 (15 millones de palabras), del cual se seleccionaron 15 mil frases con

una extensión de 5 a 15 palabras. Se calculó el índice de entropía para cada frase y se seleccionaron 7 mil frases con el índice más bajo. De éstas se eliminaron números, acrónimos y palabras extranjeras para acercar la representación gráfica a la fonética y así facilitar la lectura. Ello resultó en un total de 5,010 frases diferentes, que corresponden a 50 frases individuales leídas por cada uno de cien hablantes de la ciudad de México, además de 10 frases comunes leídas por todos, lo cual dio un total de 6,000 frases. Se tomaron en cuenta las características sociales de los hablantes (procedencia, edad y nivel de estudios).

Para la transcripción se utilizó un conjunto de 37 alófonos del español mexicano representados con el alfabeto Mexbet. A partir del corpus se desarrollaron varias herramientas computacionales. Por ejemplo, el Fonetizer que sirve para pasar el texto escrito a su representación en fonemas y alófonos.

2.9. Corpus de Investigación en Español de México del Posgrado de Ingeniería Eléctrica y Servicio Social (CIEMPIESS)

El CIEMPIESS fue creado en el Laboratorio de Procesamiento de Voz del Posgrado de Ingeniería Eléctrica de la UNAM por prestadores de servicio social provenientes de la Facultad de Ingeniería y de la Facultad de Filosofía y Letras de la UNAM, dirigidos por un estudiante de doctorado de dicho Laboratorio. Se trata de un corpus oral del español hablado en el centro de México, extraído de programas de radio transmitidos por Radio-UNAM y diseñado para la creación de modelos acústicos para uso en reconocimiento automático de voz, con extensión de 17 horas y aproximadamente 12 mil palabras distintas.

El corpus está disponible en la red desde junio de 2015 y es de uso libre y gratuito bajo una licencia del Linguistic Data Consortium.

Este corpus está anotado a nivel palabra y cuenta además con los archivos necesarios para configurar un sistema de reconocimiento automático de voz en español de México. Los más importantes son:

Archivos de audio. El corpus se conforma por un total de 16,717 archivos en el formato SPHERE de la National Institute of Standards and Technology (NIST), de 16 bits, monoaural y con una frecuencia de muestreo igual a 16 kHz.

Archivo de transcripción. El archivo de transcripción está dado en el formato que utiliza el software de reconocimiento de voz CMU-SPHINX. Esto es, se transcribe a nivel ortográfico el contenido de un solo archivo de audio por cada línea y además

se especifica entre paréntesis una clave con la que se puede ligar la línea en cuestión con el archivo de audio correspondiente. El archivo está escrito en minúsculas y se especifica la vocal tónica de cada palabra con una mayúscula (por ejemplo, pErro, gAto, acciOn). También se toman en cuenta los diferentes sonidos de la letras; por ejemplo, la letra "x" tiene un símbolo especial para cada uno (así, mEJico en vez de mExico, $ilOfono en vez de xilOfono, SicotEncatl en vez de xicotEncatl, eKStraEr en vez de extraEr).

Diccionarios de pronunciación. En todos los diccionarios de pronunciación se muestra la palabra a reconocer y su transcripción fonética en alfabeto Mexbet. Se proveen cuatro diferentes tipos de diccionarios de pronunciación, uno de ellos tomando en cuenta solo los fonemas del español de México y las vocales tónicas, otro tomando en cuenta los fonemas y alófonos del español mexicano, además de las vocales tónicas, y finalmente las versiones de estos dos diccionarios sin que se especifique ninguna vocal tónica. Cada diccionario cuenta con alrededor de 50,000 palabras diferentes.

Modelo del lenguaje. El modelo de lenguaje fue extraído de Boletines UNAM y cuenta con 1.5 millones de palabras. Se entrega en formato ARPA y en dos versiones, una de ellas tomando en cuenta las vocales tónicas de cada palabra y la otra con todas las palabras en minúsculas y sin acentos.

Etiquetas. Las etiquetas son archivos con la extensión "textgrid" que fueron generados con el software PRAAT. Cada una de estas etiquetas está ligada a un archivo de audio y cuentan con la información de dónde inicia y termina cada palabra en ese archivo.

Conjunto de pruebas y de entrenamiento. En reconocimiento de voz se necesita un corpus para la creación de modelos acústicos y a este se le conoce como conjunto de entrenamiento, pero también es conveniente tener un pequeño subconjunto de ese corpus para poder evaluar la calidad del reconocimiento y a ese subconjunto se le conoce como conjunto de pruebas. Ahora bien, de los 16,717 archivos de audio que conforman el corpus CIEMPIESS se tomaron 700 para conformar un conjunto de pruebas y el resto se dejó para el conjunto de entrenamiento. A estos 700 archivos se le incluyeron otros 300 archivos de audio tomados de un pequeño corpus que ya existía en el laboratorio de procesamiento de voz. Debe especificarse que ninguno de los archivos del conjunto de pruebas está incluido en el conjunto de entrenamiento y que ni el conjunto de pruebas ni el conjunto de entrenamiento se encuentran incluidos en el modelo de lenguaje. Esto es para garantizar que las pruebas de reconocimiento estén sesgadas indebidamente.

2.10. Corpus Electrónico del Español Colonial Mexicano (COREECOM)

Como parte de un proyecto del Instituto de Investigaciones Filológicas de la UNAM, se construyó el COREECOM para estudiar el fenómeno de variación y cambio lingüístico del español colonial mexicano. Con el fin de mostrar esta variedad, se consideraron cuatro niveles: diacrónico, diatópico, diastrático y diafásico.

El nivel diacrónico marca las diferencias temporales desde 1475 hasta 1821, en periodos de 25 años. Esta división con la base de que para poder registrar variaciones léxicas se requiere al menos tener cortes generacionales de 25 años.

El nivel diatópico señala las variaciones geográficas, tomando en cuenta también los aspectos sociales, históricos y lingüísticos. Por ello, diferencia tres zonas: la zona de raíces o de la Península Ibérica; la zona de tránsito, que incluye las Canarias, las Antillas y la Capitanía General de las Filipinas, estas dos últimas que formaron parte del Virreinato; la zona del territorio novohispano, que a su vez lo dividen en zona norte, zona central, zona peninsular de Yucatán, zona de intercambio comercial y Capitanía General de Guatemala.

El nivel diastrático señala las diferencias sociales, lo que permite conocer si estas existen y de qué tipo. Por un lado, en cuanto a los estratos sociales, considera seis grupos: españoles, criollos, mestizos, indios, negros y mulatos. Por otro lado, diferencia entre los textos escritos por mujeres o por hombres. Asimismo, señala cuando los documentos son de portugueses, italianos o judíos.

El nivel diafásico marca los diferentes registros lingüísticos, esto es, las modalidades comunicativas que se eligen en una situación comunicativa. Así, en el corpus se señala la variedad textual según el propósito de los documentos, tales como cartas privadas, testamentos, denuncias, solicitudes, etc. Asimismo, se diferencia el tipo de documento, entre textos informales, semiformales y formales.

Los documentos se pueden recuperar en cualquiera de las tres presentaciones tradicionales para documentos históricos: el facsímil o copia del documento antiguo, la transcripción paleográfica que plasma en caracteres actuales respetando en lo posible la grafía original, y la transcripción crítica que interpreta los distintos niveles lingüísticos pero manteniéndose fiel al original.

2.11. Archivo de textos hispánicos de la Universidad de Santiago de Compostela (ARTHUS)

ARTHUS forma parte de la Base de datos sintácticos del español actual (BDS), un proyecto iniciado en 1988 en la Universidad de Santiago. El corpus se construyó dada la necesidad de un recurso que contuviera datos representativos para el análisis sintáctico del español.

ARTHUS contiene en la actualidad textos pertenecientes a diferentes etapas de la historia del español. La parte contemporánea abarca treinta y cuatro textos narrativos, teatrales, ensayísticos, periodísticos y orales, procedentes de España e Hispanoamérica; tiene 160 mil cláusulas y 1,450,000 palabras. Todos los textos fueron digitalizados mediante escáner y programas de reconocimiento óptico de caracteres, están en formato ASCII y tienen una codificación mínima en formato COCOA que permite, con los programas de recuperación adecuados, conocer texto, página y línea en que se encuentran los ejemplos buscados.

El análisis sintáctico se realizó de manera manual y presenta una descripción detallada y a profundidad de importantes fenómenos sintácticos, prestando especial atención a la estructura de cláusula y al régimen verbal. En particular, el análisis proporciona la siguiente información: a) Localización de los predicados en los textos. b) Tipo y función de la cláusula, su voz, modalidad, polaridad, forma verbal, orden de los elementos, etc. c) Funciones sintácticas dentro de las cláusulas. d) Casos de predicativos no argumentales, ejemplos especialmente interesantes de cláusulas complejas, ausencia de marcas, etc.

Cabe mencionar que ARTHUS no es un corpus anotado sintácticamente en el sentido tradicional del término, ya que consiste en una codificación del análisis realizado manualmente. Los textos analizados para ARTHUS no están lematizados ni anotados con categorías gramaticales. El análisis se realizó tomando la cláusula como unidad central e identificando los elementos funcionales que la componen y la clase a que pertenecen, pero sin proceder al análisis interno de cada uno de ellos.

La interfaz de consulta permite obtener los análisis sintácticos correspondientes a diversos tipos de cláusulas u oraciones, así como las frecuencias de un verbo o de una construcción sintáctica en el corpus.

2.12. Archivo Gramatical de la Lengua Española (AGLE)

AGLE está constituido por más de 100 mil fichas de texto escrito y oral. Para su construcción se digitalizaron los fragmentos de texto escritos durante más de 50 años por

el gramático español Salvador Fernández Ramírez (1896-1983), que tenía el propósito de escribir una gramática de la cual al final sólo se publicó un volumen. El AGLE ilustra una serie de construcciones sintácticas del español, y los fragmentos o fichas que lo conforman están ordenados con base en un criterio gramatical y no lexicográfico.

Actualmente se está editando y anotando en el Instituto Cervantes para ser consultado electrónicamente de manera ágil, gracias a una base de datos en la que se clasificó la información. La intención de este archivo es respetar el orden establecido por el autor, organizar las partes menos articuladas, clasificar las fichas que el autor no llegó a ordenar, y completar, sin añadir ni una sola ficha, los bloques temáticos existentes tomando siempre como guía el criterio del autor.

AGLE, en su primera entrega, constaba de unos 75 ficheros, cada uno de los cuales contiene alrededor de 1,500 fichas a las que el autor se refiere siempre como cédulas. No todos los ficheros poseen el mismo grado de ordenación interna ni todos poseen una articulación similar. Los ficheros seguían aproximadamente el orden que el autor tenía previsto para su gramática, pero aun así eran muy numerosas las fichas que se agrupaban en apartados como “varios” o “sin clasificar”. Actualmente el archivo cuenta con cerca de 116 mil fichas organizadas en cinco grupos: partículas (18 mil fichas aproximadamente), verbos (58 mil fichas), nombres (5 mil fichas), determinantes y pronombres (26 mil fichas), y oraciones (9 mil fichas). Cada célula contiene el número de referencia para su identificación en el archivo, la cita en el que aparece un fenómeno gramatical dado, el autor y la obra.

AGLE ofrece la posibilidad de obtener información detallada sobre los fenómenos gramaticales con ejemplos concretos de uso. La interfaz en Internet permite obtener concordancias o citas por alguno de los cinco grupos en los que está ordenado.

2.13. Proyecto CRATER

El proyecto europeo Corpus Resources and Terminology Extraction (CRATER) es un corpus trilingüe (inglés, francés, español) paralelo de textos técnicos de la International Telecommunications Union (ITU). El corpus está etiquetado morfológicamente e incluye el alineamiento de frases en español con sus equivalentes en francés e inglés. Fue elaborado en conjunto por la Universidad Autónoma de Madrid y la Universidad de Lancaster.

Para este proyecto fue creado un etiquetador gramatical (POS tagging) en español, con el que se rectificaron los errores de marcado gramatical tanto en la versión inglesa como en la francesa.

La extensión del corpus es de un millón de palabras y ha sido de beneficio para proyectos de traducción automática, lingüística computacional y para el estudio de corpus en general. Asimismo se desarrolló un conjunto de herramientas para la recuperación de datos del corpus, que permite examinar las alineaciones de términos o palabras entre los distintos idiomas que lo conforman.

2.14. CHILDES

El corpus Child Language Data Exchange System (CHILDES) es en sí un repositorio para corpus de investigadores en todo el mundo en el tema de adquisición del lenguaje infantil. Cuenta con transcripciones de interacciones conversacionales de niños de diferentes edades agrupados en más de 130 corpus para más de 20 idiomas. CHILDES integra tres componentes básicos: CHAT, para la anotación y codificación discursiva; CLAN, para buscar y manipular la base de datos; y la base de datos misma, que contiene un conjunto de archivos de texto transcrito.

2.15. Base de datos de Energía ETDEWEB

La base de datos ETDEWEB contiene la colección más grande del mundo de la literatura sobre energía. Cuenta con más de 3.8 millones de archivos, en los que se incluyen referencias bibliográficas y artículos de periódicos, informes, conferencias, libros, etc., y cubre varios aspectos medioambientales del uso, producción, políticas y planeación de energía, así como las ciencias básicas que apoyan su investigación y desarrollo.

ETDEWEB contiene citas publicadas mundialmente de las áreas nuclear, del carbón y la información de cambio de clima global, entre otras. Los usuarios de esta base de datos son tan diversos como los temas que abarca: científicos, ingenieros, bibliotecarios, líderes de industria, y estudiantes.

El banco de datos está disponible vía Internet para cualquier usuario, organización, biblioteca o institución de los países miembros de ETDEWEB (México, Estados Unidos, Japón, entre otros); asimismo, se encuentra en productos CD-ROM.

2.16. Referencias

Páginas de los corpus

- Index Tomisticus: www.corpusthomisticum.org

- SEU: www.ucl.ac.uk/english-usage/
- Brown Corpus: clu.uni.no/icame/
- CEMC: www.corpus.unam.mx/cemc
- CORDE: corpus.rae.es/cordenet.html
- CREA: corpus.rae.es/creanet.html
- Corpus del Español: www.corpusdelespanol.org
- BwanaNet: bwananet.iula.upf.edu
- CLI: www.corpus.unam.mx/cli
- CSMX: www.corpus.unam.mx/csmx
- CHEM: www.corpus.unam.mx/chem
- CORCODE: www.corpus.unam.mx/corcode
- RST: www.corpus.unam.mx/rst
- CEELE: www.corpus.unam.mx/ceele
- COCIEM: www.corpus.unam.mx/cociem
- DIME: leibniz.iimas.unam.mx/~luis/DIME/CORPUS-DIME.html
- CIEMPIESS: http://www.ciempiess.org
- COREECOM: www.iifl.unam.mx/coreecom/presentacion.html
- ARTHUS: adesse.uvigo.es/data/corpus.php
- AGLE: cvc.cervantes.es/lengua/agle
- CHILDES: childes.psy.cmu.edu
- ETDEWEB: www.etde.org

Lecturas sobre los corpus

Index Tomisticus. Busa, Roberto (1980). “The annals of humanities computing: The index thomisticus”. Computers and the Humanities 14 (2), pp. 83-90.

SEU. Quirk, Randolph (1960). “Towards a description of English usage”. Transactions of the philological society 59 (1), pp. 40-61.

Brown Corpus. Francis, Winthrop Nelson y Henry Kučera (1989). “Manual of Information to accompany a Standard Corpus of Present-Day Edited American English, for use with Digital Computers”. Rhode Island: Department of Linguistics, Brown University.

CEMC. Lara, Luis Fernando, Roberto Ham Chande y M. Isabel García Hidalgo (1979). Investigaciones lingüísticas en lexicografía. México: El Colegio de México.

CORDE y CREA. Sánchez Sánchez, Mercedes y Carlos Domínguez Cintas (2007). “El banco de datos de la RAE: CREA y CORDE”. Per Abbat, Boletín filológico de actualización académica y didáctica 2, pp. 137-148.

Corpus del español. Davies, Mark (2002). “Un corpus anotado de $100{,}000{,}000$ palabras del español histórico y moderno”. Sociedad Española para el Procesamiento del Lenguaje Natural, pp. 21-27.

BwanaNet. Bach, Carme, Roser Saurí, Jordi Vivaldi y Teresa Cabré (1997). “El corpus de l’IULA: descripció”. Papers de l’IULA, Serie informes 17. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada.

CLI. Garduño, Gabriel, Gerardo Sierra y Alfonso Medina (2004). “Herramientas de análisis para el Corpus Lingüístico en Ingeniería”. En Miguel Arias Estrada y Alexander Gelbukh (Eds.), Avances en la Ciencia de la Computación. Colima: Sociedad Mexicana de Ciencias de la Computación, pp. 219-226.

CSMX. ReyesCareaga, Teresita, Alfonso Medina y Gerardo Sierra (2011). “Un corpus para la investigación en la extracción de términos y contextos definitorios: hacia un diccionario de las sexualidades en México”. Debate Terminológico 7, pp. 24-35.

CHEM. Medina, Alfonso y Carlos Méndez Cruz (2006). “Arquitectura del Corpus Histórico del Español en México (CHEM)”. En A. Hernández y J. L. Zechinelli (Eds.), Avances en la ciencia de la computación. México: Sociedad Mexicana de Ciencia de la Computación, pp. 248-253.

CORCODE. Sierra, Gerardo, Rodrigo Alarcón, César Aguilar, Alberto Barrón, Valeria Benítez y Itzia Baca (2008). “Corpus de contextos definitorios: una herramienta para la lexicografía y la terminología”. En Estrada, Z. y A. Munguia (Eds), IX Encuentro Internacional de Lingüística en el Noroeste, Hermosillo: Universidad de Sonora.

RST. da Cunha, Iria, Juan Manuel Torres-Moreno y Gerardo Sierra (2011). “On the Development of the RST Spanish Treebank”. Proc. Fifth Law Workshop, Association for Computational Linguistics, Portland, Oregon, pp. 1-10.

DIME. Villaseñor, Luis, Antonio Massé y Luis Pineda (2001). “The DIME Corpus”. En C. Zozaya, M. Mejía, P. Noriega y A. Sánchez (Eds.), Memorias 3^{o} Encuentro Internacional de Ciencias de la Computación ENC01, Tomo II, Aguascalientes, México.

CIEMPIESS. Hernández Mena, Carlos Daniel y Abel Herrera Camacho (2014). “CIEMPIESS: A New Open-Sourced Mexican Spanish Radio Corpus”. Proc. 9th International Conference on Language Resources and Evaluation, Reykjavik, Islandia, pp. 371-375.

COREECOM. Arias, Beatriz (2009). “Confección de un corpus para conocer el origen, la evolución y la consolidación del español en la Nueva España”. En A. Enrique-Arias (Ed.), Diacronía de las lenguas iberorrománicas: Nuevas aportaciones desde la lingüística de corpus. Madrid: Iberoamericana, pp. 55-78.

ARTHUS. Rojo, Guillermo (2001). “La explotación de la Base de Datos Sintácticos del español actual”. En J. de Kock (Ed.), Lingüística con corpus. Salamanca: Universidad de Salamanca.

AGLE. Leonetti Jungl, Manuel, Silvia Gumiel Molina, Pilar Pérez Ocón y Juan Romero Morales (2006). “El proyecto AGLE (Archivo Gramatical de la Lengua Española): desarrollo y perspectivas”. Actas del XXXV Simposio Internacional de la Sociedad Española de Lingüística, León, México, pp. 1048-1059.

CRATER. Sánchez León, Fernando y Amalio Nieto Serrano (1995). “Desarrollo de un etiquetador morfosintáctico para el español”. Procesamiento de Lenguaje Natural 17, pp. 14-28.

CHILDES. Macwhinney, Brian (2000). The CHILDES Project: The database. Psychology Press.

Capítulo 3

Clasificación de corpus

Con el fin de sentar bases sólidas que sirvan como guía para el diseño, construcción y caracterización de diferentes corpus, conviene establecer una tipología. Los corpus se clasifican de acuerdo con el origen de sus elementos y las características que poseen. En la siguiente tabla 3.1 se muestra un cuadro para facilitar la distinción de los tipos de corpus.

3.1. Según el origen de los datos

Se puede hablar de dos diferentes tipos de corpus según el origen: corpus textuales y corpus orales. Los corpus textuales consisten, como su nombre lo indica, en la recopilación de textos procedentes de la lengua escrita, mientras que los orales se constituyen por grabaciones o transcripciones de la lengua hablada. En este sentido, los orales pueden estar formados por grabaciones o por transcripciones del habla, es decir que pueden dividirse en orales sonoros, que están formados por grabaciones, y orales transcritos, que contienen transcripciones de lengua oral y pueden incluir marcas directas de oralidad o señalamientos sobre la misma.

Entre los ejemplos de corpus orales se encuentran el corpus DIME, en español, y el London-Lund corpus, para el inglés. De corpus textuales, se encuentran el Lancaster Oslo/Bergen corpus (LOB), en inglés. Asimismo, es posible encontrar corpus que contienen tanto textos orales como escritos, como es el caso del CREA, en español, y del British National Corpus (BNC), en inglés.

3.2. Según la espontaneidad del habla

Un corpus oral se compone de grabaciones o transcripciones del habla, mientras que un corpus textual consta de elementos en lengua escrita. Sin embargo, en algunas ocasiones resulta difícil clasificar ciertos tipos de corpus, por ejemplo, la grabación de una muestra de lengua que en sus orígenes estuvo escrita, o bien, la transcripción de habla que originalmente fue una conversación. Dada la naturaleza de los corpus descritos con anterioridad, debemos clasificar el primero como un corpus oral, y el segundo como textual. Sin embargo, para considerar estos matices en los que textualidad y oralidad parecieran mezclarse, un criterio de clasificación de corpus es el de espontaneidad. De acuerdo con esto, una muestra verbal espontánea, como una conversación entre dos personas, constituirá un corpus no premeditado; mientras que muestras de habla no espontánea, como la lectura de un texto en voz alta, compondrán un corpus premeditado.

Diversos corpus incluyen muestras de habla no premeditadas, como el Corpus de ad-hocracia, que contiene transcripciones de los debates sostenidos entre diputados y senadores. Por otro lado, otros corpus sólo contienen muestras de habla premeditadas, como es el caso del Corpus textual especializado plurilingüe, del IULA.

3.3. Según la codificación y anotación

También existe una clasificación de corpus según la codificación y anotación, donde encontramos el corpus simple y el corpus codificado o anotado. El corpus simple es, en el caso de los corpus textuales y de los orales transcritos, el que ha sido guardado en un formato ASCII y no tiene una codificación especial. Mientras que el corpus codificado o anotado es aquél que está formado por elementos de la lengua a los cuales se ha añadido, electrónica o manualmente, etiquetas para reconocer algunos de sus elementos o para facilitar su análisis y procesamiento. Al ser una muestra de lengua, los corpus pueden utilizarse para analizar cualquiera de los niveles de ésta: textual, fónico, morfológico y morfosintáctico, sintáctico, semántico, discursivo y pragmático. Todos estos niveles, por su complejidad e importancia se detallarán en el Capítulo 7: Bases para la anotación de corpus en este libro.

3.4. Según la especificidad de los elementos

Existen también clasificaciones de los corpus según la especificidad de los textos. Así tendremos corpus generales y corpus especializados o también llamados específicos. Los generales se encargan de recoger todo tipo de géneros y tipologías textuales; son útiles para describir la lengua común de una colectividad. Los corpus especializados,

por el contrario, recogen material que puede aportar datos para la descripción de un área o tema en particular. Dentro de los específicos, podemos hablar de un nivel más de clasificación en el caso de los corpus textuales; los que contienen textos literarios y los que se conforman de textos informativos. Dentro de los primeros se incluyen ensayo, narrativa, poesía y teatro. En lo que concierne a los textos informativos, existen los periodísticos, científicos, académicos y técnicos.

El banco de datos de la RAE, conformado por el CREA y el CORDE, contiene información de diversos géneros y tipologías, por lo que sus corpus son generales. Por su parte, el Corpus textual especializado plurilingüe, del IULA, es un ejemplo de corpus específico.

3.5. Según la autoría de los elementos

De acuerdo con la autoría de sus elementos, un corpus puede ser genérico, canónico o de autoría variada. En el primer caso, se trata de un corpus que recoge documentos de un solo un género textual, por ejemplo, artículos de revistas científicas especializadas o textos poéticos. El segundo tipo de corpus, el canónico, recoge la obra completa de un autor, independientemente del género textual, es decir, no importa si se trata de poemas, novelas, cuentos u obras de teatro. Por último, si los textos no comparten alguna de las características anteriores, se trata de un corpus de autoría variada.

El CORCODE, compuesto por contextos definitorios, es un corpus genérico, mientras que el Open Source Shakespeare es canónico; la Biblioteca Digital del Pensamiento Novohispano puede considerarse un corpus de autoría variada.

3.6. Según la temporalidad de los elementos

Los corpus pueden ser sincrónicos o diacrónicos, de acuerdo con la temporalidad. Los sincrónicos contienen elementos del lenguaje de un momento específico en el tiempo, y pueden ser de dos tipos: contemporáneos e históricos. Los corpus sincrónicos contemporáneos se componen por documentos actuales y los históricos por documentos de un periodo de tiempo pasado. Por otro lado, los corpus diacrónicos confrontan o relacionan muestras lingüísticas a través de varios periodos de tiempo, y pueden ser cronológicos o periódicos. Los corpus diacrónicos cronológicos estudian alguna lengua o variedad de lengua a través del tiempo, mientras los periódicos estudian la lengua en lapsos definidos de tiempo. En función del tiempo también tenemos a los corpus monitor. Estos buscan mostrar un estado actual de la lengua, de manera que contienen textos recientes, por decir algo, pertenecientes a los últimos 25 años. El corpus contiene un volumen textual o periodo de tiempo constantes, por lo que su contenido se actualiza

con frecuencia. Esta acción ofrece la posibilidad de tener un corpus dinámico, en los que se van incluyendo nuevos materiales al mismo tiempo que se eliminan los más antiguos.

El CREA es un ejemplo de corpus sincrónico contemporáneo, y el CORDE, por su parte, es un ejemplo de corpus diacrónico cronológico. A la vez, el CREA es un tipo de corpus monitor, pues al menos pretende tener el español contemporáneo y deja en el CORDE textos más antiguos.

3.7. Según el propósito de estudio

En general, los corpus se construyen con un propósito determinado para realizar investigaciones concretas, sin embargo, la lingüística de corpus busca crear corpus utilizables para más de un objetivo. De esta manera, podemos hablar de corpus para propósitos específicos en oposición a corpus multipropósito. Los corpus de propósito específico son los más comunes y pueden ser de tres tipos: de estudio, de entrenamiento y de prueba. El corpus de estudio se utiliza para identificar y describir algún aspecto del lenguaje; por ejemplo, los fragmentos de una obra literaria que contienen una palabra o una construcción sintáctica determinada. El corpus de entrenamiento y el de prueba comparten la característica de ser etiquetados manualmente, pero su función es diferente. El de entrenamiento se toma como base para que un sistema computacional “aprenda” determinada tarea (como sería la extracción de información o el resumen automático, por ejemplo). Por su parte, el de prueba se compara con un corpus procesado automáticamente para evaluar la efectividad de los algoritmos y procesos de los sistemas computacionales. Por otro lado, el corpus multipropósito por antonomasia es el de referencia, que no persigue el objetivo de proporcionar un tipo específico de información pues se construye, como su nombre lo indica, para proporcionar información diversa aplicable a estudios de diferente naturaleza.

El CEMC es un ejemplo de corpus de estudio. Fue construido con el único propósito de confeccionar el Diccionario del Español de México, no obstante ha sido utilizado en numerosas investigaciones más allá del fin original. Por su parte, el CREA es un corpus multipropósito de referencia que se ha puesto a disposición del público para distintos fines.

3.8. Según la lengua

Según la lengua de los elementos, existen los corpus monolingües y los multilingües. Un corpus monolingüe utiliza un solo idioma, por ejemplo, español o inglés, sin importar si sus elementos son originales del idioma o si son traducciones. A su vez, estos corpus

pueden ser monolingües según la variedad dialectal, cuando en sus elementos se diferencian dialectos o variedades lingüísticas, o bien, pueden ser comparables, si se componen por textos originales de una lengua y traducciones de otros textos semejantes en la misma lengua.

Por otro lado, los corpus multilingües pueden ser de textos en distintos idiomas o paralelos. Los primeros se conforman por colecciones de textos en varios idiomas, recopilados con criterios de selección muy diversos, desde la simple disponibilidad de los textos hasta la selección según géneros y tipos similares. Por su parte, los corpus paralelos contienen la misma colección de textos en más de una lengua, es decir, las versiones originales acompañadas por sus traducciones. El caso óptimo de paralelismo se produce cuando las traducciones son un reflejo simétrico de la versión original.

Ejemplo de corpus monolingüe es el CREA. El corpus comparable por excelencia lo constituye el Translational English Corpus, que contiene más de seis millones de palabras de textos traducidos al inglés. Ejemplos de corpus multilingüe son el ARTHUS, compuesto por textos de derecho contractual en danés, francés e inglés.

3.9. Según la cantidad de texto

De acuerdo con la cantidad de texto que se recoge, podemos dividir entre corpus grandes y corpus pequeños. Se considera corpus grande al que contiene una cantidad considerable de documentos, en oposición a corpus cuantitativamente más pequeños. Un corpus de diez millones de palabras podría considerarse grande, aunque, existen corpus de cien millones. Ahora bien, un corpus pequeño es aquél que no satisface necesidades informáticas y estadísticas por la pequeña cantidad de texto recogido en él, pero que puede sernos muy útil para fines lingüísticos determinados.

El CORDE es un ejemplo de corpus grande, con más de 250 millones de palabras, en tanto el CEMC, con dos millones de palabras, que en su momento era un corpus grande, actualmente es un corpus pequeño.

3.10. Según la distribución del tipo de texto

En esta clasificación se toma en cuenta la distribución y el porcentaje recogido de los diferentes tipos de texto que componen un corpus. Éstos pueden ser corpus equilibrados en oposición a los corpus no equilibrados; también pueden ser corpus piramidales. Un corpus equilibrado contiene una variedad de documentos que se encuentran distribuidos en proporciones parecidas para cada uno de los tipos de documentos. De esta variedad

se puede tener la zona geográfica, el tipo de documento, el año, etc. En oposición al anterior, el corpus desequilibrado contiene tipos de documentos cuyas cantidades no son proporcionales entre sí. Ahora bien, los corpus piramidales contienen textos que están distribuidos en diferentes niveles: el primer nivel recoge pocas variedades temáticas pero con muchos textos en cada variedad; el segundo nivel contiene más variedades temáticas, pero menos cantidad de textos en cada una de ellas; el tercer nivel está compuesto por muchas variedades con pocos textos en cada una y así sucesivamente hasta un número opcional de estratos. Cabe aclarar que, por su organización en cuanto a las variedades temáticas, el corpus equilibrado por excelencia es el piramidal.

El CORDE, que recoge muestras de diversas fuentes y países, es un corpus desequilibrado tanto en tamaño como en el tipo de documentos o la distribución geográfica; por otro lado, el CSMX es del tipo piramidal.

3.11. Según la accesibilidad

Una clasificación que puede hacerse sobre corpus es en función de la accesibilidad o disponibilidad para usarlo. Esta clasificación involucra dos tipos de corpus: los de dominio público y los de uso privado o restringido. Los corpus de dominio público pueden, a su vez, contar con otra clasificación, en corpus comerciales, que exigen pagar una cuota para su utilización, o no comerciales. A estos últimos puede tenerse acceso restringido o acceso libre. La restricción a ciertos corpus se debe a que son construidos en instituciones públicas de investigación y buscan asegurarse que el material será utilizado sin fines de lucro. Para ello, es necesario establecer convenios específicos y comprometerse a usar el corpus para fines de investigación. Estos convenios, además, permiten asegurar a la institución que creó el corpus, justificar su utilización y conseguir recursos. La accesibilidad a un corpus depende del soporte electrónico para el que fueron diseñados. Puede tenerse un corpus disponible para su uso en línea a través de una dirección URL en Internet, o puede bajarse a través de servidores ftp. Asimismo, puede estar disponible en discos flexibles o en CD-ROM. Para corpus orales también se cuenta con videos y grabaciones electrónicas.

El CREA es un corpus de dominio público, mientras que muchos grupos de investigación mantienen sus corpus privados, como fue durante muchos años en el caso del CEMC.

3.12. Según la documentación

Esta clasificación depende de si se tiene documentación de los textos que componen el corpus. Se habla de corpus documentados cuando se tiene registro de la documentación de los textos y, además, es posible usar dicha documentación, ya sea para hacer una búsqueda específica o para conocer de dónde provienen los textos. Por el contrario, un corpus no documentado carece de registros documentales de los textos.

Ejemplo de corpus documentado es el ARTHUS, mientras que el CLI es un corpus no documentado.

3.13. Según la representatividad

Por último, aunque podría darse por hecho que un corpus debe ser representativo, existe una clasificación según la representatividad. En este caso, la primera tipología obligada será la de representativo, aunque, además, existen los corpus oportunistas. Éstos no son necesariamente representativos de toda una lengua, pero pueden serlo de un fenómeno en específico. En estos corpus se recogen muestras que presenten el fenómeno a estudiar, según el recopilador las lea, las escuche o las encuentre de diversas maneras, de ahí el nombre de oportunista. Si bien este tipo de corpus más bien puede ser considerado una colección miscelánea de ejemplos verbales, puede servir de base para la construcción de otros corpus. El CEMC es un ejemplo de corpus representativo del español de México, en tanto el CREA tiende más a ser un corpus oportunista por la cantidad desproporcionada entre el español peninsular y el americano.

3.14. Referencias

Lecturas sugeridas

Berber Sardinha, Tony (2000), “Lingüística de Corpus: Histórico e Problemática”. DELTA 16 (2), pp. 323-367.

Torruela, Joan y Joaquim Llisterri (1999). “Diseño de corpus textuales y orales”. En J.M. Blecua et al. (Eds.), Filología e informática: Nuevas tecnologías en los estudios filológicos. Barcelona: Editorial Milenio-Universidad Autónoma de Barcelona. (Véase sección 3).

Páginas de otros corpus mencionados

- Lancaster-Oslo/Bergen Corpus (LOB): www.helsinki.fi/varieng/CoRD/corpora/LOB
- British National Corpus (BNC): www.natcorp.ox.ac.uk
- Open Source Shakespeare: www.opensourceshakespeare.org
- Biblioteca Digital del Pensamiento Novohispano: www.bdpn.unam.mx
- Translational English Corpus: www.llc.manchester.ac.uk/ctis/research/english-corpus

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 autoria 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	
		Piramidal	
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		

Tabla 3.1: Clasificación de corpus.

Capítulo 4

Internet como corpus

Se ha visto ya que la constitución de corpus representativos permite realizar análisis específicos. Aunque el Internet no es un corpus como tal, pues sus documentos no cuentan con una clasificación general ni con una codificación estandarizada, es una fuente inagotable de textos en diferentes idiomas, accesibles, en su mayoría, de manera gratuita. Así, se puede considerar el Internet como un corpus dinámico, una entidad viva que se modifica día a día. Lo que hay en una página hoy, mañana puede estar en otra página o cambiar su contenido drásticamente. Por ello, a la hora de documentar una página, es conveniente señalar el día en que se obtuvo.

Debe quedar claro que el Internet no es un corpus como tal, pues ni es una selección de textos, ni cumple con los requisitos de representatividad, variedad y equilibrio. Sin embargo, es posible usar el Internet como corpus con una adecuada selección de documentos según sean los objetivos de análisis. Así, se puede convertir un corpus vivo y dinámico en otro estático, debidamente variado y equilibrado, bajado en el disco duro de un servidor.

4.1. Formatos electrónicos

En la web se pueden encontrar documentos guardados en distintas clases de formatos. Entre los más comunes se encuentran:

txt (Text). Es el formato más sencillo y puede ser leído por la mayoría de los procesadores. Se trata de archivos de texto sin formato, por lo que no se marcan los diferentes tipos y estilos de letra.

rtf (Rich Text Format). Es un archivo de texto que puede ser leído por casi todos los procesadores de texto. Conserva el tipo y estilo de las letras, entre negritas y cursivas, por ejemplo.

doc (Document). Es una extensión de archivo de Microsoft Word, que antes de la versión 2007 tenían los documentos por defecto.

docx. Es la extensión actualizada del formato doc, escrito en XML (eXtensible Markup Language).

html (HyperText Markup Language). Es el formato que por defecto tienen las páginas web creadas en el lenguaje de programación que lleva el mismo nombre.

pdf (Portable Document Format). Es el formato creado por Adobe que sólo permite lectura. Su uso es ideal para proteger texto que no queremos que sea modificado por otros usuarios. Debido a que solo es para lectura, no resulta trivial pasarlo a un formato texto, pues las dobles columnas las considera como una sola columna, el fin de línea lo marca como fin de párrafo, etc. Además, en algunos casos se llega a digitalizar un documento como imagen, de manera que no puede leerse el texto, lo cual implica usar un reconocedor óptico de caracteres.

A fin de poder trabajar con los documentos será conveniente transformar todos a un mismo formato, normalmente en archivos de texto.

4.2. Los buscadores como herramientas

Los buscadores de Internet son una herramienta de suma utilidad para compilar corpus. Éstos funcionan mediante la recuperación de información en bases de datos, y vinculan páginas web que contienen información similar. De esta manera, al hacer una búsqueda, arrojan direcciones relacionadas con el tema consultado, ordenadas de acuerdo con su relevancia.

Existen, de acuerdo con su funcionamiento, cuatro tipos de buscadores: directorios o índices de búsqueda, motores de búsqueda, metabuscadores y buscadores semánticos.

Directorios o índices de búsqueda

Los directorios o índices de búsqueda tienen una base de datos o un directorio de páginas web organizado manualmente por categorías y subcategorías temáticas en función de su contenido, de manera que el usuario, para hacer una consulta, debe elegir entre una serie de subcategorías hasta llegar a la información deseada. Por ejemplo, para llegar a la consulta 'poesía de Sor Juana' habría que elegir entre las siguientes categorías: arte> literatura> poesía> período barroco> Sor Juana. Un ejemplo de estos directorios actuales es Open Directory Project (ver figura 4.1).

Figura 4.1: Ejemplo de índices de búsqueda en Open Directory Project.

Motores de búsqueda

Los motores de búsqueda funcionan de manera similar a la de un directorio, excepto porque la indexación u organización por categorías y su actualización se hace de manera automática, por medio de programas informáticos llamados arañas, robots, agentes robot o web crawlers, que relacionan la información de diversos servidores web. Así, al realizar una búsqueda, arrojan un listado de direcciones web relacionadas con la consulta. Los motores de búsqueda más empleados son:

AltaVista. Buscador de la empresa Overture Services Inc., comprada por Yahoo! en 2003. AltaVista ofreció el primer índice de la web de Internet y fue el primer motor de búsqueda de Internet que permitió hacer búsquedas multilingües, de imágenes, audio y video. Actualmente emplea la tecnología del motor de búsqueda de Yahoo! y funciona para los idiomas inglés y español.

Yahoo! Search. Motor de búsqueda propiedad de Yahoo! Inc. Comenzó como un directorio web, organizado de manera jerárquica. En 2009 Microsoft y Yahoo! anunciaron que el motor de búsquedas de Bing sería adoptado por Yahoo! Search. Actualmente este buscador permite hacer búsquedas de páginas web, de imágenes, de noticias y videos, y cuenta con un buscador social que ofrece un servicio de preguntas y respuestas realizadas en la web por cualquier usuario registrado. Asi-

mismo, tiene un buscador de compras por Internet, que funciona con un directorio de búsqueda.

Bing. Buscador de Microsoft, presentado en el 2009. En este mismo año se anunció que Bing reemplazaría a Yahoo! Search. Las versiones anteriores de este buscador se llamaron: Live Search, Windows Live Search y MSN Search, la primera de las cuales se lanzó en 1998. Actualmente, este buscador funciona para varios idiomas y permite realizar búsquedas de páginas web, imágenes, noticias, deportes, finanzas, etc., así como hacer cálculos matemáticos. Cuenta con una función para hallar definiciones de palabras desde el diccionario de Encarta.

Google. Buscador de Google Inc., lanzado en 1998. Este buscador permite realizar búsquedas de páginas web, imágenes, videos y noticias. Asimismo, cuenta con la modalidad académica que encuentra resultados de artículos científicos, y con la modalidad de búsquedas en libros, que permite visualizarlos parcialmente. En él se tienen filtros para distintos formatos, de manera que pueden traerse, por ejemplo, la palabra o frase buscada en archivos doc, html, pdf o rtf, entre otros, lo cual resulta útil para traer únicamente documentos y no sólo páginas electrónicas.

Metabuscadores

Los metabuscadores, por su parte, no tienen una base de datos propia. Emplean las bases de varios buscadores para ofrecer resultados, es decir, los metabuscadores buscan en los buscadores. Un ejemplo es:

WebCrawler. Pertenece a InfoSpace Inc. Se trata de un metabuscador que emplea Google, Yahoo! y Bing, y permite hacer búsquedas en periódicos, páginas blancas y amarillas. Asimismo, proporciona resultados multimedia, incluyendo imágenes, video, noticias e información local. La forma en la que muestra los resultados permite distinguirlos entre los que son orgánicos (expuestos al medio de la página) y los que son patrocinados (aparecen en la parte superior e inferior de la página).

Buscadores semánticos

Los buscadores semánticos efectúan consultas atendiendo al significado del grupo de palabras que se introducen, sin importar las etiquetas que tengan las páginas web. Por lo general, estos buscadores hacen una desambiguación de la entrada que se consulta y arrojan los resultados en tablas que atienden a la anotación semántica del buscador.

WolframAlfa. Fue creado por Wolfram Research y está en funcionamiento desde el 2009. Es un sistema de búsqueda de respuestas; esto permite que el usuario introduzca una pregunta en lenguaje natural y que el sistema categorice la respuesta

en tablas, en las que organiza la información. Las categorías que hace se basan en criterios semánticos, es decir, que los atributos de las tablas varían de acuerdo con tipo de información que requiera el usuario. Por ejemplo, los resultados de la búsqueda que tenga que ver con una persona estarán ordenados en una tabla que tenga como atributos fecha de nacimiento, nombre completo, etc. mientras que para una consulta que tenga que ver con un país los atributos podrían ser población, clima, ciudades cercanas, etc. Este sistema permite, asimismo, cargar archivos y analizarlos con sus herramientas.

4.3. Técnicas para usar los buscadores

Los motores de búsqueda permiten distintas posibilidades para encontrar las páginas que tienen indexadas. Así, algunos buscadores distinguen entre mayúsculas y minúsculas y el orden en el que se introducen las palabras en la caja de búsqueda.

Dependiendo del motor que se emplee para realizar una búsqueda, se pueden usar diversos tipos de símbolos para precisar los resultados de la información que queremos que nos arrojen. Estos símbolos pueden ser operadores de diversos tipos, que denotan operaciones lógicas, de orden, relación o truncamiento, y tienen la función de unir, separar, excluir y precisar las palabras o frases necesarias para filtrar los resultados de una búsqueda en la web o en un sistema de datos. Los más comunes se explican a continuación.

Operadores booleanos

AND. También se usa el signo + (más) en algunos buscadores. Traerá todas las páginas que contengan ambas palabras. Por ejemplo, casa AND campo traerá las páginas que contengan casa y campo.

OR. Traerá las páginas que contengan una u otra palabra. Por ejemplo, casa OR campo traerá las páginas que contengan ya sea casa o campo. Este operador se usa en el caso de que el usuario no utilice ningún operador en dos o más palabras.

NOT. Este operador se utiliza para excluir de los resultados búsqueda las páginas que contengan una palabra en particular. Por ejemplo, casa NOT campo traerá las páginas que contengan la palabra casa pero en dicha página no debe aparecer la palabra campo. En algunos buscadores se usa el signo - (menos) o AND NOT.

Operadores de posición

ADJ. Permite buscar dos palabras contiguas. En algunas ocasiones se simboliza con comillas.

BEFORE. Funciona de manera similar al AND, con la diferencia de que las palabras incluidas en la búsqueda aparecerán en el orden especificado, sin importar la distancia a la que se encuentren en un mismo documento.

FAR. Permite encontrar resultados en los que se incluyen dos términos en un mismo documento, pero que se encuentran alejados el uno del otro.

FOLLOWED BY. Indica que uno de los términos debe estar inmediatamente seguido por otro.

NEAR. Se emplea cuando se quiere buscar una palabra que esté cerca de otra, en una misma página. Se puede incluso especificar la distancia entre las palabras. Actualmente funciona en Altavista y resulta de mucha utilidad para varias aplicaciones de lingüística de corpus.

W/N. Donde N es el número de lugares junto a los cuales queremos encontrar una palabra. Por ejemplo, la búsqueda de rana W/3 lago, arrojará resultados en los que la palabra lago se encuentre a tres posiciones de la palabra rana.

Operadores de truncamiento (comodines)

$*$ Posibilita escribir sólo una parte del término deseado. Por lo general sólo se puede aplicar al final de la palabra y es útil para encontrar formas no lematizadas (como verbos conjugados, sustantivos en diminutivo, etc.).

? Suele sustituir un solo caracter en una búsqueda. Por ejemplo, amig?s arrojará resultados para el sustantivo amigo en ambos géneros.

$+$ Encuentra las distintas formas que tiene una palabra. Por ejemplo, dar$+$ arroja como resultado daría, daba, etc.

$\#$ Excluye de la búsqueda las formas de una palabra que no sean como se escribió. Por ejemplo, actúan$\#$ arroja resultados que contienen “actúan” pero excluye actuar, actor, actuaba, etc.

Otros

Frase exacta. Se utiliza para buscar cadenas de palabras y no palabras aisladas. Para ello, generalmente se emplean las dobles comillas. Por ejemplo, çasa de campo”.

Paréntesis. Se usan para combinar operadores booleanos. Por ejemplo, casa AND (verano OR vacaciones) traerá las casas que sean de verano o de vacaciones. Nótese que el lugar del paréntesis puede afectar la búsqueda.

4.4. Documentos disponibles en Internet

Aunque en todo texto se pueden hacer análisis de cualquier nivel de la lengua, de acuerdo con la finalidad de la investigación habrá documentos con información más precisa para nuestros estudios. Los siguientes son ejemplos de documentos que se encuentran en la web, y algunas aplicaciones que pueden arrojar sus análisis.

Periódicos. Existe una diversidad de periódicos de diferentes países, capturados diariamente a su forma electrónica, de tal forma que es posible ver la información clasificada por temas y de todas las fechas desde su primera publicación electrónica. También hay información periodística de grandes agencias de noticias, como CNN, tanto en inglés como en español, que tienen formatos más adecuados para Internet y con mayor contenido de información.

Libros digitales. La mayoría de los publicistas mantienen digitalizados un cada vez más creciente número de libros. En Internet, algunos autores independientes permiten la lectura libre y completa de sus libros que ya han sido ampliamente difundidos, o de los artículos que quieren dar a conocer, y de los cuales aún poseen derechos de autor.

Bibliotecas digitales. Son colecciones de libros análogos o semejantes entre sí, en las que la información se encuentra en algún formato digital. En estas bibliotecas las fuentes de información están disponibles y su acceso es ubicuo. Algunos ejemplos de bibliotecas virtuales son la Biblioteca Virtual Miguel de Cervantes y la Biblioteca Digital del Pensamiento Novohispano.

Colecciones temáticas. Son bases de datos que contienen artículos, libros, guías técnicas, manuales, etc. sobre algún tema en particular. Pueden facilitar el acceso a documentos completos o a resúmenes de los mismos. Ejemplos de estas colecciones son MedLine y ETDEWEB.

Actas. Varias organizaciones mantienen en Internet las actas de sus conferencias y reuniones. En algunos casos éstas se comercializan por medio de CD-ROM y su costo sirve para cubrir los gastos de edición. El CD-ROM contiene los textos completos de las conferencias en cuestión.

Páginas en línea. Para el registro de los principales textos en línea y su ágil consulta conviene tener una base de datos especial. Hay que considerar que las fichas

de páginas electrónicas en Internet o de documentos en línea difieren a las tradicionales bibliográficas, ya que se deben incluir campos específicos, tales como: dirección electrónica o URL, fecha más reciente en la que se verificó la página, contenido temático, formato y extensión del documento (pdf, doc, ps, html).

Chats. Las pláticas a través de los servicios de mensajería instantánea pueden ser salvadas y utilizadas para análisis de errores en teclado, ortográficos, de diálogos, etc. Dado que en ellas no se tiene especial cuidado en la redacción, sino en el contenido, se aproximan más a un registro coloquial de habla espontánea que al habla natural.

Correo electrónico. Los correos electrónicos pueden salvarse y emplearse como entradas de un corpus. Los análisis que pueden realizarse en éstos son de registro de habla (familiar o formal); de las temáticas que abarcan; de la finalidad que tienen: si son correo publicitario, informativos, personales, etc. lo que permite, por ejemplo, detectar correo no deseado.

Grupos de discusión. Los comentarios que los usuarios dejan en los foros en línea pueden emplearse para hacer estudios de minería de opiniones, esto es, detectar cómo percibe una persona una marca y sus productos, los fenómenos sociales del momento, la tendencia de las candidaturas y debates de los personajes públicos, etc.

4.5. Programas para análisis lingüísticos

Por su parte, existen motores de búsqueda que permiten hacer consultas de manera mucho más específica que los buscadores convencionales. Aquí se presentan tres ejemplos.

WebCorp

Fue desarrollado por la Research and Development Unit for English Studies (RDUES) en la Escuela de Inglés, de la Universidad de Birmingham. Este buscador permite obtener concordancias para una palabra o una frase, en una ventana definida por el usuario. Asimismo, se pueden ver las colocaciones y hacer búsquedas especificando la posición en la que se desea encontrar una palabra. Cuenta también con una opción para filtrar la búsqueda por fecha y omitir las palabras funcionales. Además, tiene una herramienta que genera listas de palabras de las páginas web, al introducir la URL; estas listas no sólo son unipalabra sino de hasta cinco n-gramas, según especifique el usuario. Como resultados adicionales, se puede traer la página (en el formato electrónico con la que fue creada) donde aparece la palabra o frase, resaltada en color amarillo; la página en archivo texto plano y la lista de las palabras que aparecen en el texto, ordenadas

por frecuencia o por orden alfabético. Su buscador permite filtrar las búsquedas por periódicos (del Reino Unido, Estados Unidos o Francia) y por temas. Los resultados también se pueden solicitar para que lleguen por correo electrónico.

Actualmente, WebCorp emplea buscadores como Google, Altavista y Bing, entre otros, sin embargo, se contempla una nueva versión que funcionará con un motor de búsqueda propio, que poseerá un rastreador web, un analizador, tokenizador, indexador y otros componentes que le permitirán procesar grandes sectores de la web.

Diatopix

Diatopix es una herramienta diseñada en el Observatoire de linguistique Sens-Texte, de la Universidad de Montreal, que permite observar la distribución geográfica del uso de palabras.

La herramienta usa el motor de búsqueda de Yahoo! y divide los resultados entre los principales países donde se usa la lengua en la que se realizó la búsqueda; por ejemplo, España, México, Chile, Argentina, Venezuela, Colombia y Cuba, para el idioma español. La interfaz permite elegir el dominio al que pertenecen los términos que se quiere comparar (agricultura, arquitectura, arte, biología, lingüística, pesca, química, etc.).

Diatopix permite realizar búsquedas, ya sean de una sola palabra o binarias (ejemplo: 'piscina' y 'alberca'). Los resultados son mostrados en gráficas de barras y en una tabla que indica sus ocurrencias por país (ver tabla 4.1), y pueden usarse para confirmar intuiciones sobre el uso y distribución geográfica de ciertos términos.

	piscina	*alberca*
Spain	1489	167
Mexico	497042	1317
Chile	4689	591
Argentina	1489	167
Venezuela	1596274	674
Colombia	2822	3857
Cuba	583	0

Tabla 4.1: Ejemplo de una tabla de resultados para una búqueda en la herramienta Diatopix.

Sketch Engine

El motor Sketch Engine (SKE) es un producto de Lexical Computing Ltd. Esta herramienta cuenta con un corpus muy extenso, disponible para 42 idiomas diferentes, aunque

también permite que el usuario suba e instale su propio corpus, por medio de una herramienta llamada WebBootCat. El corpus cargado para español tiene cerca de dos mil quinientos millones de palabras etiquetadas con partes de la oración y lematizadas. El SKE es de uso diario para la lexicografía en Oxford University Press, Cambridge University Press, Collins, Robert y Cornelsen Verlag, entre otros.

Para formar un corpus propio se puede proporcionar una serie de palabras semilla para que su buscador obtenga textos de Internet con dichas palabras semilla, o bien proporcionar las direcciones URL de las páginas que se deseen. SKE permite etiquetar y lematizar los corpus obtenidos. Para los corpus se ofrecen distintas herramientas de análisis, entre otras:

Concordancias. Obtiene las concordancias de una palabra como lema, de manera que se despliega el contexto en que aparece una palabra con todas sus formas. Esto es, en el caso de sustantivos, en plural o singular, femenino o masculino; en el caso de verbos, sus formas verbales. Asimismo, proporciona el enlace al archivo en donde aparece la concordancia.

Lista de palabras. Presenta una lista ordenada por frecuencias absolutas de palabras, lemas o partes de la oración. En el ejemplo mostrado en la figura 4.2, puede apreciarse que el lema más frecuente es “el”, que incluye las formas “el, la, los, las”, en tanto que la parte de la oración más frecuente es la preposición (etiqueta SPS00), seguido del sustantivo común, femenino, singular (etiqueta NCFS000). Asimismo, permite obtener n-gramas de las palabras, lemas o partes de la oración, por ejemplo, la frecuencia de ocurrencia del conjunto de tres palabras seguidas en el corpus.

Colocaciones. Identifica en el corpus las palabras que coocurren sintácticamente ante un lema determinado. De esta manera, se pueden observar colocaciones como “meter la pata”, “el tío mete”, o “mete y saca”. Como todo proceso automático, hay que tomar en cuenta que el etiquetado que realiza tiene errores, pues en el caso de “meter”, considera que una de sus formas es “meta”, y por tanto nos trae como colocación “tiene como meta contribuir”. En la figura 4.3 aparece solo un concentrado de las colocaciones, pero el SKE despliega también las concordancias de cada uno de los casos.

Tesauro. Obtiene las palabras que tienen alguna relación semántica con otra determinada, tomando en cuenta que si ambas tienen un contexto similar serán entonces similares. Por ello, se basa en analizar los contextos similares en que aparecen dos palabras, según sus colocaciones. Así, para “sacar” y “meter”, en tanto pueden tener colocaciones únicas y diferentes, como en “sacar provecho” y “meter miedo”, tienen contextos similares, como “meter/sacar la cabeza/el pie”.

lemma	Freq
el	233736566
de	146589256
y	62120815
que	61575709
en	56954418
a	47460083
uno	41919681
ser	32801043
se	25363371

tag	Freq
SPS00	340089967
NCFS000	148792590
NCMS000	144893686
Fc	122632290
NP00000	121028220
DA0MS0	92583311
CC	82330274
Fp	78823181
DA0FS0	77388091

Figura 4.2: Ejemplo de una lista de palabras generada por el motor SKE.

4.6. Ligas de interés

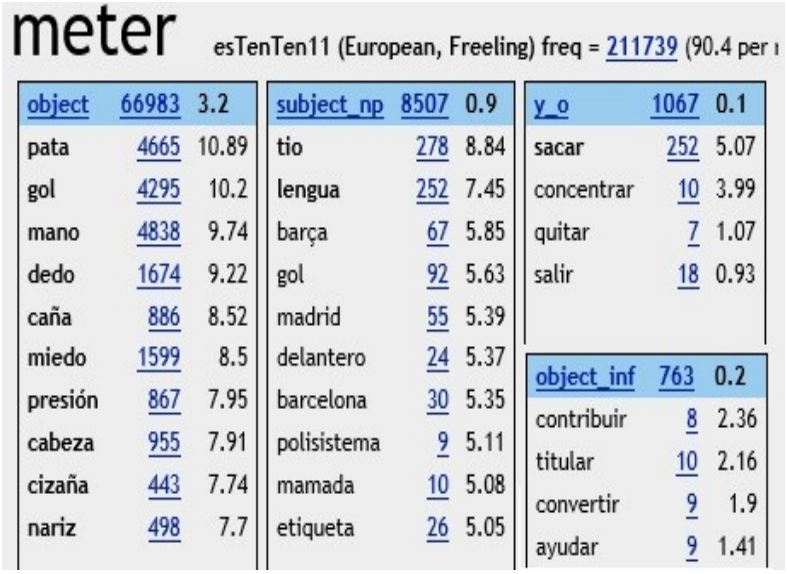
Periódicos de México

Periódicos diarios:

- prensaescrita.com: www.prensaescrita.com/america/mexico.php
- Kiosko.net: www.kiosko.net/mx

Bibliotecas digitales en español en Internet:

- Bibliotecas virtuales: www.bibliotecasvirtuales.com
- Artnovela: www.artnovela.com.ar
- Biblioteca Virtual Miguel de Cervantes: www.cervantesvirtual.com
- Bibliotheka: http://ebiblioteca.org/
- Ciudad Seva: www.ciudadseva.com
- Mundo Aliat: http://www.aliatuniversidades.com.mx/bibliotecasdigitales/index.php/en/



meter esTenTen11 (European, Freeling) freq = 211739 (90.4 per

object	66983	3.2
pata	4665	10.89
gol	4295	10.2
mano	4838	9.74
dedo	1674	9.22
caña	886	8.52
miedo	1599	8.5
presión	867	7.95
cabeza	955	7.91
cizaña	443	7.74
nariz	498	7.7

subject_np	8507	0.9
tio	278	8.84
lengua	252	7.45
barça	67	5.85
gol	92	5.63
madrid	55	5.39
delantero	24	5.37
barcelona	30	5.35
polisistema	9	5.11
mamada	10	5.08
etiqueta	26	5.05

y_o	1067	0.1
sacar	252	5.07
concentrar	10	3.99
quitar	7	1.07
salir	18	0.93

object_inf	763	0.2
contribuir	8	2.36
titular	10	2.16
convertir	9	1.9
ayudar	9	1.41

Figura 4.3: Ejemplo de una lista de colocaciones generada por SKE.

Programas para análisis lingüísticos a partir de Internet

- WebCorp: www.webcorp.org.uk/live
- Diatopix: http://olst.ling.umontreal.ca/~drouinp/diatopix/
- Sketch Engine: www.sketchengine.co.uk

Lecturas de web como corpus

Kilgarrif, Adam y Gregory Grefenstete (2003). Introduction to the special issue on the web as corpus. Computational linguistics 29 (3), pp. 333-347.

Volk, Martin (2002). “Using the Web as Corpus for Linguistic Research”. En R. Pajusalu et T. Hennoste (Eds.), Tähendusepüüdja. Catcher of the Meaning. A Festschrift for Professor Haldur Õim. Publications of the Department of General Linguistics 3, University of Tartu.

Parte II

Compilación de corpus

69

Capítulo 5

Compilación de corpus textuales

Los corpus textuales provienen de documentos escritos. Ante la vastedad de información es necesario considerar criterios mínimos para seleccionar el material que los conforma. Por ello, es importante identificar el objetivo del estudio, seleccionar y obtener los documentos que, en muchos casos, hay que digitalizar, o que bien, se pueden obtener de la web. Si se trata de un proyecto grande habrá que conformar un equipo de trabajo en el que se incluya un administrador.

5.1. Identificación del objetivo

Es importante considerar el propósito que tiene el corpus para que la selección de materiales sea la más adecuada. Algunas cuestiones que se deben tener en cuenta son las siguientes:

a) ¿Cuál será la accesibilidad, es decir, se tendrá en Internet, CD-ROM, ftp, Internet u otro medio?

b) ¿Será necesario diseñar herramientas de análisis y explotación? Y, en caso afirmativo, ¿qué tan amigables o complejas serán?

c) ¿De qué manera se realizará la anotación?, ¿se puede seguir o no un estándar de etiquetado?

d) ¿Cuál es el propósito del corpus? Los propósitos pueden ser generales o específicos; los siguientes son algunos ejemplos:

   - Análisis lingüístico (fonético, morfológico, sintáctico, semántico, pragmático, discursivo).
   - Enseñanza de idiomas.

- Lexicografía y terminología.
- Investigación en procesamiento de lenguaje natural.
- Aplicaciones de ingeniería lingüística (traducción automática, corrección ortográfica o de estilo, generación de documentos, etc.)

e) ¿Qué límites tendrá el corpus? El propósito, el costo y el tiempo para hacer el proyecto definen los límites, que pueden ser:

- Límites temporales: sincrónico vs. diacrónico.
- 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.

5.2. Selección de textos

Un corpus para propósitos generales está diseñado para usarse en varios proyectos, por lo que suele ser grande y, al emplearlo, hay que seleccionar los textos que representan el tema que se va a analizar.

Por esta razón, el corpus debe estar clasificado por temas de especialización y contener los datos lingüísticos relevantes. Se debe contar con un buen diseño de base de datos, en donde se tenga una estructura adecuada con campos bien definidos, que permita la búsqueda mediante descriptores claros y precisos, así como encontrar sólo los textos o documentos pertinentes, clasificados jerárquicamente en orden de su relevancia.

Hay que tomar decisiones, considerando los propósitos del estudio, en cuanto a la selección de los documentos a incorporar en función del balance y representatividad. Asimismo, hacer un muestreo de los fenómenos que se quieran estudiar, a partir de una población determinada; es decir, de un conjunto de todos los elementos que se quieren estudiar (esto es, la población), habrá que elegir una pequeña parte, que sea representativa del fenómeno que analizamos (muestra).

También deberán seleccionarse las partes del texto que se utilizará. En un corpus léxico se toman fragmentos de diferentes documentos, normalmente escogidos aleatoriamente. Un ejemplo prototípico lo constituye el CEMC, que está conformado por 996

fragmentos de aproximadamente 2000 palabras cada uno, recogidos de 14 géneros diferentes. En un corpus de referencia habrá que definir los elementos a guardar en función de los objetivos del proyecto. Es común que sólo se guarde el texto del cuerpo de los documentos, incluso sin hacer distinción para los títulos y subtítulos. Sin embargo, para ciertos fines puede resultar de interés obtener toda la información del documento, incluyendo portadas, índices, pies de página, recuadros, figuras, tablas, etc.

5.3. Obtención de textos

Para obtener los textos de un corpus es necesario, ya bien, buscar los documentos que lo conformarán, o solicitarlos a editoriales, editores y académicos. Asimismo, hay que pedir las cartas de autor y llevar un seguimiento del estado de documentación de cada texto que conforma el corpus.

El registro de los documentos se tendrá que hacer en una base de datos, en la que se indique la procedencia de los textos y el estado de los documentos, es decir, si se tiene el derecho de autor, quién lo está digitalizando, en qué proceso de la digitalización se encuentra, etc.

5.4. Digitalización de documentos

De ser posible, conviene contar con textos en formato electrónico que sean fácilmente convertidos a archivos texto. Cuando no es posible, existen dos procesos alternos para pasar los textos a formato electrónico. El primero es mediante la lectura óptica automática de documentos impresos a través de un dispositivo (escáner) y un programa de cómputo. La calidad y confiabilidad de este proceso depende tanto del software como del hardware, esto es, del programa y del dispositivo. Si bien hoy en día se ha avanzado en ambos, siempre hay un margen de error, por lo que es necesaria una etapa de edición, la cual tiene que ser rigurosa y muy cuidadosa para ser fieles al texto digitalizado. Como alternativa a este proceso semiautomático, el segundo proceso es mediante el tecleado directo de los textos en el archivo electrónico o base de datos a través de un capturista, lo cual hace más lento el proceso, pero puede aumentar su confiabilidad.

El digitalizador de imágenes o escáner

El digitalizador de imágenes o escáner es un equipo que transforma una imagen analógica en una digital. Utiliza un foto-sensor que recibe la luz que es enviada desde la imagen a través de un juego de espejos y la convierte en señales eléctricas controladas por la intensidad y el color de la imagen, de la misma manera en que funciona un ojo. Estas

señales eléctricas son recibidas por un convertidor analógico-digital que las convierte en bites, los cuales forman nuevamente la imagen digitalizada en la computadora.

La resolución óptica o real es el número de puntos individuales de una imagen que el foto-sensor del escáner es capaz de captar y se expresa en puntos por pulgada (ppp). Para textos deben tener una resolución de 300 a 600 ppp.

El escáner es un periférico que se conecta a la computadora mediante un puerto paralelo, un conector SCSI o un puerto USB.

Existen tres principales tipos: de sobremesa o planos (de gran tamaño), de mano (portátiles) y de rodillo (para hojas sueltas).

Reconocedores de textos

El escáner reconoce los puntos que forman un texto, como si fuera una fotografía, no reconoce como tal las letras, palabras o frases. El reconocedor de caracteres es un programa que lee las imágenes digitales del texto y busca conjuntos de puntos que se asemejen a letras; para hacerlo, existen dos reconocedores:

Optical Caracter Recognition (OCR). Convierte archivos que tienen formato imagen en archivos de formato texto. Los OCR reconocen los caracteres tipográficos de un documento escrito por medio de una máquina de escribir o una impresora y convierte automáticamente la información para ser legible por la computadora.

Intelligent Caracter Recognition (ICR). Aplica pruebas de inteligencia lógica a los caracteres escaneados (como los manuscritos) para convertirlos de manera más confiable en información más legible para la computadora. Para interpretar de manera más correcta la información, aplican reglas de ortografía, gramática y contexto.

Entre los criterios para seleccionar el software están: el método de reconocimiento, la velocidad de reconocimiento, los idiomas aceptados, los tipos de formatos de salida, el grado de especialización de los diccionarios, las características especiales (como los correctores ortográficos integrados) y el precio.

Algunos programas para lectura de caracteres son el OmniPage, que reconoce texto en 119 idiomas, terminología legal y médica. Asimismo, está el ABBYY, que reconoce textos con letra manuscrita.

La digitalización conlleva tres procesos principales: la digitalización del documento, el reconocimiento de los caracteres y el guardado del documento. La digitalización consiste en la conversión del documento físico o analógico a una imagen digital, como si fuera

una fotocopia. En este sentido, las letras no son más que imágenes, las cuales hay que interpretarlas como letras, justo en la segunda etapa a través del reconocimiento de caracteres. En ésta, primeramente hay que marcar cada una de las zonas que componen una página y quitar las que no interesen salvar. Por ejemplo, en una hoja de doble columna habrá que marcar la primera columna y después la segunda, pues de lo contrario considerará ambas columnas como si fuera una sola. Dado que no se deben tratar de la misma forma el texto, las tablas, las figuras y los esquemas, hay que marcar las zonas dependiendo de su género; no hay que olvidar que las notas al pie de página y las referencias bibliográficas son texto. Sin embargo, las figuras y los números de página normalmente se omiten.

El reconocimiento de caracteres es dependiente del idioma, pues no resulta igual digitalizar un texto en inglés que en francés o español, ya que, para empezar, se tienen diferentes caracteres, lo que puede ocasionar confusión de letras cuando no corresponde al idioma seleccionado. Así, puede reconocer un 6 en caso de “á” y de “ó”, o bien, la vocal sin acento; l, t, f o un uno, en el caso de “í”; para la letra ñ, el OCR lo llega a cambiar por fl, f, li, n, $\sim$. La selección del idioma permite también utilizar un diccionario, el cual se activa una vez que se han reconocido los caracteres. De esta manera, cuando una palabra no se encuentra en el diccionario, el OCR lo marca para que la persona que digitaliza la señale como válida o haga el cambio correspondiente.

Una vez que ya se ha digitalizado y revisado todo o una parte, se puede ir guardando el documento ya digitalizado, así como la parte que falta por revisar. Para salvar el documento existen varios formatos disponibles: txt, rtf, doc, html o xml.

Problemas de la digitalización

El proceso de digitalización parece una tarea sencilla y a menudo es subestimada en el diseño del proyecto. Hay que considerar que es un trabajo que requiere pasar varias horas atrás del escáner y varía en función de la experiencia del digitalizador, del equipo de cómputo y del reconocedor óptico de caracteres. Incluso puede también depender del tipo y la calidad del papel, pues el papel rugoso o de fax puede afectar seriamente la calidad del reconocimiento de caracteres, razón por la que en tal caso es necesario sacar previamente fotocopias al documento original para tener otro tipo de papel. Así, puede darse el caso que un libro de 300 páginas se digitalice, limpie y etiquete en 10 horas de trabajo, en tanto para otros tipo de papel pueda requerirse más de 50 horas.

No obstante que el Diccionario OCR reconoce un 90 % de los errores, el 10 % restante se trata de cambio de letras dentro de la palabra (a$\rightarrow$o, i$\rightarrow$e, l$\rightarrow$t, entre otras) que gramaticalmente son aceptables como: las $>$ los, los $>$ tos, piso $>$ peso; de tal manera que será necesaria una rápida pero eficaz revisión ocular.

5.5. Obtención de textos de Internet

Tomando en cuenta que Internet es una fuente de donde pueden extraerse textos para crear un corpus, se han diseñado a la fecha varias herramientas. Si bien resultan útiles para ciertos proyectos y se obtienen los textos en corto plazo, conviene considerar que no cumplen con los criterios que hemos observado para la construcción de un corpus representativo, ya que no podemos asegurar la procedencia de los textos, por lo que pueden provenir de distintas regiones geográficas, épocas y dialectos, a la vez que no se puede precisar la variedad y el equilibrio de los textos obtenidos.

BootCaT

En las universidades de Bologna, Trento y Zagreb fue desarrollado el ahora software libre BootCaT, que permite compilar un corpus a partir de los textos en Internet. El programa se descarga para trabajar en Windows, Mac, Linux o UNIX. El usuario obtiene los textos a partir de palabras semilla que considera correspondan con el dominio de su corpus. Existe también una interfaz en línea, denominada WebBootCaT, que corre a través del Sketch Engine, descrito anteriormente.

Obtención de textos para el CSMX

Como caso concreto de aplicación sobre la obtención de corpus representativo a partir de Internet, se tiene el Corpus de las Sexualidades en México (CSMX). Para la selección de los textos se tomaron en cuenta los siguientes criterios de clasificación:

Origen de los elementos. En principio se trata de un corpus textual obtenido de documentos en Internet al que se eliminaron las imágenes.

Espontaneidad. Una gran parte son textos premeditados, pues fueron concebidos para ser publicados, aunque también hay textos no premeditados que fueron obtenidos de foros de discusión y de diálogos escritos entre debatientes en algunas páginas.

Especificidad de los elementos. Es un corpus específico en el área de las sexualidades, en el cual se buscaron textos científicos, periodísticos, académicos y técnicos.

Autoría. Se seleccionaron textos de autoría variada, pero tratando de asegurar que los autores fueran de México.

Tiempo. El corpus es de naturaleza sincrónica, con textos contemporáneos.

Lenguaje. Es un corpus monolingüe y monodialectal, ya que trata de la variante mexicana del español.

Distribución del texto. Se trata de un corpus piramidal, ya que de los 160 archivos fueron divididos en ocho subáreas temáticas y éstas a su vez se subdividieron en cinco niveles.

En relación con las características fundamentales de corpus, el CSMX cumple con los criterios de representatividad, variedad y equilibrio.

Representatividad. Se incorporó el mayor número posible de registros de hablantes, diferenciados en cinco niveles:

- **Google Académico.** Este nivel corresponde a textos escritos por especialistas en el tema de sexualidad y se dio preferencia a los textos con mayor número de referencias de acuerdo con las palabras introducidas en el motor de búsqueda.
- **Asociaciones.** Se consideraron las asociaciones públicas o privadas, ya sea instituciones educativas, de salud y de servicios, que tienen reconocimiento en el área.
- **Artículos PDF.** Debido a que los artículos salvados en PDF están protegidos para no ser modificados, se encuentran en este nivel artículos que no están muy referenciados como los anteriores, pero representan autores con experiencia, como comunicólogos y difusores de la ciencia.
- **HTML y Word.** En los documentos salvados en formato Word o de páginas web, codificación HTML, se encuentran textos que no cuentan con un respaldo institucional.
- **Foros.** Se tomaron en cuenta tanto archivos de foros de discusión reconocidos, como aquellos completamente libres. En este nivel se encuentra un lenguaje más coloquial y más cercano a la oralidad, debido a que se escribe de manera espontánea y próximo a los actos de habla.

Variedad. Se tomó como punto de partida la división del Instituto Kinsey, teniendo finalmente ocho apartados:

- **Fundamentos biológicos de la sexualidad.** Incluye temas de anatomía humana, el cuerpo visto como objeto de estudio, el control de la natalidad y medicina general orientada a sexualidad.
- **La respuesta y la expresión sexual.** Abarca temas sobre el orgasmo, fantasías sexuales y afrodisiacos.
- **Comportamiento sexual.** Con temas sobre la estimulación, las posturas, el placer y la actividad sexual.

- **Identidad sexual.** Sobre los roles, la identidad y orientación sexual, los estereotipos y las anomalías en la identificación sexual.
- **Enfermedades de transmisión sexual.** Considerando la sintomatología, el control y los mecanismos de prevención.
- **Sexualidad variante.** Sobre los trastornos y diferencias psicosociales.
- **Atracción sexual.** Sobre las relaciones de pareja, la fidelidad y el abuso u ofensa sexual.
- **Educación sexual.** Diferentes temas sobre los procesos de educación, cultura y sociedad de la sexualidad.

Equilibrio. Primero se buscó guardar equilibrio entre los registros (representatividad) y las temáticas (variedad), pues de cada una de las ocho temáticas y cinco niveles se obtuvieron cuatro archivos, quedando un total de 160 archivos. Además, cada archivo debía tener un tamaño uniforme, por lo que se cortaron los libros y se juntaron varios textos cortos (como los foros de discusión) en un solo archivo.

5.6. Estandarización de formatos

Cuando los documentos son recopilados de diferentes fuentes, como es el caso de Internet o que se recopiló la información de diversos investigadores en donde cada uno tiene su propio formato, es indispensable estandarizar el formato. En el caso de archivos de audio, los formatos pueden ser wav, mid, mp3 y ogg, entre otros. Para archivos texto, los documentos pueden estar en pdf, doc, html, etc., que conviene unificar a un solo formato. Según el tipo de información que se requiera anotar, los archivos pueden ser de texto plano, aunque hay que tomar en cuenta que aún en texto plano existen diferentes codificaciones, por lo que también hay que uniformizar estos archivos. Entre estas codificaciones se encuentran ASCII, UNICODE, UTF-8 e ISO-9959-1, las cuales pueden ocasionar serios problemas en la lectura de ciertos símbolos.

Si bien uniformizar las codificaciones es una tarea sencilla y en muchos casos solo requiere tener cuidado a la hora de salvar los documentos, en el caso de los archivos pdf se requiere usar programas especiales.

5.7. Administración del proyecto

En caso de la elaboración de corpus grandes y en donde se vayan a desarrollar sistemas de búsqueda de información, como en todo proyecto de ingeniería, se requiere llevar a cabo una administración adecuada. Entre otros, es necesario contar con:

- Líder de proyecto para organizar el equipo.
- Administrador financiero para la consecución y manejo de recursos.
- Diseñador del corpus.
- Relaciones públicas para búsqueda de documentos y adquisición de derechos.
- Revisor de seguimiento del proyecto.
- Digitalizadores.
- Desarrolladores del sistema.
- Informáticos de apoyo.

5.8. Referencias

Lecturas sugeridas

Barnbrook, Geoff (1996). Language and Computers: A Practical Introduction to the Computer Analysis of Language. Edinburgh: Edinburgh University Press. (Véase capítulo 2).

Cole, Ronald A., Joseph Mariani, Hans Uszkoreit, Annie Zaenen y Victor Zue (Eds.) (1996). Survey of the State of the Art in Human Language Technology. Cambridge: Cambridge University Press. (Véase capítulo 2).

Sinclair, John (1991). “Creación de corpus”. En J. Vidal Beneyto (Ed.), Las industrias de la lengua. Madrid: Pirámide.

Torruela, Joan y Joaquim Llisterri (1999). “Diseño de corpus textuales y orales”. En J.M. Blecua et al. (Eds.), Filología e informática: Nuevas tecnologías en los estudios filológicos. Barcelona: Editorial Milenio-Universidad Autónoma de Barcelona. (Véase sección 4.1).

Zampolli, Antonio (1992). “Corpora de referencia”. En J. Vidal Beneyto (Ed.), Las industrias de la lengua. Madrid: Pirámide.

Sistemas de reconocimiento óptico de caracteres

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

Capítulo 6

Compilación de corpus orales

Las muestras orales pueden constituir corpus orales, que provienen de grabaciones de la señal sonora (también conocidos como speech corpora o speech databases) o pueden ser corpus de lengua hablada, que se forman por transcripciones ortográficas de la lengua hablada (también llamados spoken language corpora).

6.1. Diseño de corpus orales

Para diseñar un corpus oral o uno de lengua oral, es necesario tomar en cuenta los mismos aspectos de un corpus textual, esto es, identificar el objetivo del corpus, seleccionar y obtener el material y tener un administrador del proyecto.

Sin embargo, también hay que tener en cuenta otros aspectos; por ejemplo, en la parte de documentación, las características de los hablantes que se van a grabar (también llamados informantes); en la parte del diseño de las bases de datos, los fonemas que se van a considerar y sus distintas realizaciones; asimismo, hay que tomar en cuenta las tecnologías que tenemos a nuestra disposición para trabajar con habla y los estándares que existen para representar los sonidos.

6.2. Características de los hablantes

Mientras que en el caso de los corpus textuales hay que hacer una documentación bibliográfica del material que los constituye, en los corpus orales hay que tomar en cuenta las características de los hablantes o informantes. Algunos aspectos a considerar son: género, debe hacerse una nivelación en el corpus; edad, que según distintos autores se puede dividir en grupos de 20, 40 y 60 años o bien, 16-32, 33-55, 56-en adelante años; lugar de residencia, colonia, delegación o municipio; ocupación; nivel sociocultural,

educación, profesión y situación económica; lugar de origen; grupo étnico; y otros datos pertinentes, como tener la dentadura completa y salud mental.

6.3. Características de la grabación

Actualmente, no se puede concebir un corpus oral que no tenga soporte electrónico. Las primeras grabadoras que se emplearon para trabajar con corpus orales fueron las grabadoras magnetofónicas Wollensack y Usher, que usaban cintas magnetofónicas de acetato Scotch a una velocidad de $3\frac{3}{4}$ pulgadas por segundo (ips). El tipo de grabación debe hacerse atendiendo a los propósitos de la investigación; puede ser necesaria, por ejemplo, una grabación hecha en un ambiente controlado como un laboratorio, o en un espacio público si se necesita que tenga ruido.

6.4. Herramientas para grabar, editar y anotar

Algunos de los programas más empleados para hacer la transcripción de corpus orales son: Speech Viewer, Praat, Sound Forge, Speech Tools y WaveLab.

Asimismo, cabe mecionar el CSLU ToolKit, del Center of Spoken Language Understanding (CSLU), que incluye varias herramientas para etiquetado fónico, entre ellas Speech Viewer.

Speech Viewer

Es un programa creado por la IBM cuya finalidad es la mejora del habla en pacientes que lo requieren. Se ocupa, básicamente, de la corrección fonética y el control de la fonación. Incluye trece módulos de trabajo agrupados en cuatro apartados:

Módulo de conocimiento. Su función es mejorar la autopercepción de la voz para el mejor autocontrol de su propia voz. Entre otras cosas trabaja el conocimiento del sonido, sonoridad, intensidad, tono, etc.

Módulo de desarrollo de técnicas. Está destinado a la mejora de la fonación.

Módulo de técnicas vocálicas. Trabaja la precisión, el contraste y la creación de modelos vocálicos.

Módulo de estructuración. Analiza aspectos como la sonoridad, intensidad, tono y entonación.

Praat

Es un programa diseñado por Paul Boersma y David Weenink, de la Universidad de Amsterdam, para estudios fonéticos y de tecnologías de voz. Praat es un programa de libre distribución, gratuito, de código abierto y multiplataforma (opera en Mac, Windows y Linux). Permite analizar, sintetizar y manipular corpus orales para hacer análisis acústico, síntesis articulatoria, procesamiento estadístico de los datos, edición y manipulación de señales de audio, entre muchas otras. El usuario puede crear sus propias rutinas e incluso añadirlas a los menús del programa. En este programa se pueden hacer grabaciones, edición de señales, etiquetas, análisis de formantes, de duración y de frecuencia fundamental, y se pueden visualizar espectrogramas.

Sound Forge

Es un programa comercial que permite grabar, editar y masterizar audio, con una gran variedad de opciones de procesamiento. Cuenta con diversas herramientas que posibilitan la aplicación de efectos, la edición, grabación y codificación en casi cualquier forma de audio digital con una calidad óptima. Además, el programa soporta video, lo que permite sincronizar audio y video con la precisión de un fotograma.

Speech Tools

Se trata de un programa creado por la organización SIL International. Cuenta con diversas aplicaciones independientes que se pueden descargar aisladamente dependiendo del uso que se le quiera dar al programa. Estas herramientas permiten grabar, transcribir y analizar archivos de sonido, así como gestionar datos transcritos sin necesidad de contar con los archivos de sonido. El programa también permite escuchar, transcribir y producir los sonidos del alfabeto fonético internacional.

WaveLab

WaveLab es una herramienta de la compañía Steinberg, que permite el tratamiento del sonido en general y de la música en particular, desde los primeros acordes hasta la grabación de las maquetas. Entre sus funciones destaca la aplicación de hasta ocho efectos simultáneamente y la visualización de espectros de sonido por medio de sus analizadores. Otras de sus funciones permiten integrar videos e imágenes.

6.5. Tipos de transcripción

La transcripción consiste en pasar a la escritura una señal sonora, que puede provenir de una grabación o directamente de la oralidad. Una buena transcripción debe ser fiel a los segmentos de la señal sonora en la que se basa. El tipo de transcripción depende de los propósitos del proyecto y puede atender a los distintos niveles de análisis de la lengua. En particular, cabe destacar dos tipos: ortográfica y fónica.

Transcripción ortográfica

Transcripción ortográfica. Este tipo de transcripción se hace empleando las normas ortográficas convencionales, es decir, no usa un alfabeto fonético para representar los segmentos, sino el que se usaría normalmente en la escritura de la lengua que se trate.

Transcripción de formas canónicas. En este caso, el transcriptor apunta las formas de prestigio o formas normalizadas de la lengua. Cuando el informante usa formas contrarias al canon, por ejemplo, 'haiga', el transcriptor registrará la forma 'haya', con el alfabeto que se emplea corrientemente en la lengua con la que esté trabajando.

Transcripción fónica

Transcripción fonética. Un fono es un sonido que no produce diferencia de significado en una lengua dada. Un conjunto de fonos que no produce diferencia de significado en un mismo contexto, constituye un fonema. A cada uno de los fonos que constituyen un fonema (o dicho de otro modo, a las distintas realizaciones de un fonema), se le llama alófono. Una transcripción fonética consiste en representar, con un alfabeto fonético, los fonos de la señal sonora. De esta manera, la transcripción fonética busca describir la producción y percepción de los sonidos, tomando como punto principal sus manifestaciones físicas, es decir, a diferencia de la fonológica, no busca agrupar sonidos para delimitar un sistema funcional, sino prepondera la descripción de los mismos para fines variados.

Transcripción fonológica. Una transcripción fonológica consiste en escribir los fonemas, sin tomar en cuenta sus diferentes realizaciones; por ejemplo, las palabras 'barra' y 'haba' se transcribirían con la misma /b/, pese a que son fonos distintos: en el segundo caso el modo de producción es diferente (la primera es oclusiva y la segunda fricativa). En la transcripción fonológica se busca marcar los sonidos desde el punto de vista de su función en la lengua y establecer valores distintivos dentro del conjunto de sonidos que la componen.

6.6. Alfabetos fonéticos tradicionales

Un alfabeto fonético consiste en un conjunto de símbolos que posibilitan la transcripción fónica; esto es, la representación de los sonidos de una lengua mediante un conjunto de convenciones.

Los alfabetos fonéticos tradicionales cuentan con caracteres que no son necesariamente compatibles con los programas computacionales para transcripciones, mediciones y contabilizaciones, es decir, que no necesariamente son ASCII (American Standard Code for Information Interchange).

Alfabeto de la Asociación Fonética Internacional (AFI)

El Alfabeto Fonético Internacional fue creado por la AFI, un grupo de investigación que operaba bajo la guía de Paul Passy. Se realizó por sugerencia del fonetista danés Otto Jespersen, que tenía consciencia de la necesidad de un único alfabeto, internacional, en lugar de alfabetos particulares para las distintas lenguas. Este alfabeto posibilita representar gráficamente cualquier lengua, sin importar si cuenta o no con escritura y constituye un estándar de transcripción. Muchos de los símbolos que la AFI ha creado para los sonidos específicos de ciertas lenguas han sido incorporados al sistema ortográfico de ellas, principalmente para algunas lenguas africanas. La AFI declara que el alfabeto no está necesariamente completo, pues está en constante mejora.

Alfabeto de la Revista de Filología Española (RFE)

La Revista de Filología Española, del Centro de Estudios Históricos en Madrid, realizó su propio alfabeto fonético, que tenía la finalidad de estandarizar los signos fonéticos usados en sus publicaciones y en los estudios que realizaba el Centro. Uno de sus creadores fue Navarro Tomás. Este alfabeto consiste en una adaptación de los métodos de transcripción más empleados en la época. Fue muy usado en el mundo hispánico y ha servido de base para la creación de nuevos alfabetos.

6.7. Alfabetos fonéticos computacionales

Los alfabetos computacionales están diseñados para ser tratados mediante computadoras, por lo que sólo cuentan con caracteres ASCII, esto es, un código de caracteres basado en el alfabeto latino, que facilita que cualquier computadora lea las transcripciones.

Alfabeto computacional Speech Assessment Methodology Phonetic Alphabet (SAMPA)

SAMPA es un alfabeto básicamente fonológico que funciona para el danés, francés, inglés, holandés, italiano, noruego, sueco, español, griego y portugués. Fue desarrollado como parte del proyecto ESPRIT 1541, por un grupo internacional de fonetistas. Uno de sus objetivos fue facilitar el proceso de transcripción, por lo que su uso es sencillo. Este alfabeto ha tenido gran difusión entre los desarrollos de tecnologías del habla realizados en Europa. La versión española de SAMPA ha tratado de expandirse a América, en el marco del proyecto SpeechDat Across Latin America (SALA), y se ha establecido ya el inventario de los fonemas y alófonos de 6 dialectos americanos (incluido el español de México), con los sonidos de procedencia indígena formalizados.

Alfabeto computacional del Oregon Graduate Institute of Science and Technology (OGIbet)

El Oregon Graduate Institute Alphabet (OGIbet) fue creado para etiquetar, a nivel de fonema y alófono, los corpus orales en inglés del Center for Spoken Language Understanding (CSLU). Por su parte, Tlatoa, el Grupo de Investigación en Tecnologías del Habla, de la Universidad de las Américas de Puebla, realizó una adaptación de OGIbet para el español de México. Las convenciones que propone Tlatoa están basadas en la experiencia del etiquetado, es decir, en la práctica de transcribir corpus orales.

Alfabeto computacional Worldbet

Worldbet fue creado por James L. Hieronymus, en el marco de los desarrollos de reconocimiento y síntesis de habla de los laboratorios Bell, en Estados Unidos. Fue creado porque, de acuerdo con su inventor, algunas versiones de la AFI en códigos ASCII no incluían sonidos de otras lenguas que no fueran las europeas y algunos símbolos se utilizaban incorrectamente. Worldbet goza de mucho prestigio en los Estados Unidos, donde se le considera un alfabeto robusto, capaz de captar los sonidos de todas las lenguas y de detallar distinciones fonéticas.

Alfabeto computacional Mexbet

Mexbet es una adaptación de OGIbet y de Worldbet, para el español de México, y ha servido como base para etiquetar el corpus DIME. En este alfabeto se distinguen claramente el carácter fonético y fonémico de los segmentos, así como las variantes alofónicas, basados en los caracteres del código ASCII. Se incluyen únicamente fonemas del español de México, es decir, 17 fonemas consonánticos y 5 vocálicos (tabla 6.1).

Consonantes	Labiales	Labiodental	Dentales	Alveolares	Palatales	Velares
Oclusivos sordos	p		t			k
Oclusivos sonoros	b		d			g
Africado sordo					tS	
Fricativos sordos		f		s		x
Fricativo sonoro					Z	
Nasales	m			n	n∼	
Vibrantes				r(/r		
Laterales				l		
Vocales				**Anteriores**	**Media**	**Posteriores**
Cerradas				i		u
Medias				e		o
Abiertas					a	

Tabla 6.1: Alfabeto computacional Mexbet.

6.8. Referencias

Lecturas sugeridas

Carré, René (1991). “Los bancos de sonidos”. En J. Vidal Beneyto (Ed.), Las industrias de la lengua. Madrid: Pirámide, pp. 95-107.

Estruch, Mónica, Juan Ma. Garrido, Joaquim Llisterri y Montserrat Riera (2007). “Técnicas y procedimientos para la representación de las curvas melódicas”. Revista de lingüística teórica y aplicada 45(2), pp. 59-87.

Llisterri, Joaquim (1999). “Transcripción, etiquetado y codificación de corpus orales”. Revista española de lingüística aplicada 1, pp. 53-82.

Torruela, Joan y Joaquim Llisterri (1999). “Diseño de corpus textuales y orales”. En J.M. Blecua et al. (Eds.), Filología e informática: Nuevas tecnologías en los estudios filológicos. Barcelona: Editorial Milenio-Universidad Autónoma de Barcelona. (Véase sección 4.2).

Villaseñor-Pineda, Luis, Manuel Montes-y-Gómez, Dominique Vaufreydaz y Jean-Francois Serignat (2003). “Elaboración de un Corpus Balanceado para el Cálculo de Modelos Acústicos usando la Web”. En J. Díaz de León, G. González y J. Figueroa (Eds.), Avances en Ciencias de la Computación. México:IPN.

Programas para transcripción de corpus orales

- Praat: www.fon.hum.uva.nl/praat

- Sound Forge: www.sonycreativesoftware.com/soundforgesoftware
- Speech Tools: http://www-01.sil.org/computing/speechtools/
- WaveLab: www.steinberg.fr

Parte III

Anotación de corpus

Capítulo 7

Bases para la anotación de corpus

Una vez que se tiene un texto en formato electrónico, ya no como imagen sino en formato textual, de manera que con un procesador de palabras o con un buscador podemos encontrar cualquier palabra o parte del texto, pareciera que ya sería posible analizarlo. Sin embargo, un mismo texto puede servir a distintos tipos de análisis, de forma que resulta necesario primero identificar los elementos del texto que son de interés y, segundo, marcar los segmentos con las anotaciones que sean pertinentes. Por ejemplo, tómese cualquier libro escogido al azar y ábrase una página cualquiera. En tanto que cualquier ser humano distinguiría entre el texto del libro del encabezado (con el nombre del libro o de los autores) y de los números de página, o de las figuras y las tablas, una máquina no sabe reconocer tales diferencias. Por ello, el primer aspecto será enseñarle a la máquina cuáles son los elementos de interés, qué se va a seleccionar, esto es, si solo el cuerpo del libro, sin encabezados ni ilustraciones, o incluso esta información. Después de ello, el análisis a realizarse puede ser múltiple. A continuación unos ejemplos.

Enseñanza de lenguas. Como parte de los estudios sobre aprendizaje de una segunda lengua, interesa tener datos confiables del uso de las expresiones usadas por los nativo-hablantes. Por ejemplo, para revisar los tiempos verbales usados en oraciones subordinadas en relación con el verbo nominal, será necesario contar con textos que contengan este tipo de oraciones. Además, será necesario conocer los elementos que introducen las oraciones subordinadas e identificar claramente los verbos y, de ser posible, las características de los mismos. Todo ello requiere tener una anotación adecuada de las partes de la oración a nivel detallado. Gracias a esta anotación sería fácil pedir a un programa recuperar un listado de todas las oraciones subordinadas, ordenadas en función del verbo.

Metadatos. Con el fin detectar y capturar automáticamente los metadatos de artículos provenientes de diferentes revistas en una base de datos, tales como título del artículo, nombre de los autores, nombre de la revista y otros datos editoriales, se

podría diseñar un sistema que se base en información lingüística y metalingüística. En la información lingüística se tendrían los patrones para títulos, autores, nombres de revistas o fechas. Esto es, los títulos normalmente son frases entre 5 y 10 palabras, en tanto los autores son nombres propios de dos a 4 palabras, separados por comas, etc. En la información metalingüística se tendrían la ubicación y el tamaño o tipo de letra de cada uno de estos datos. Los títulos aparecerían con letras más grandes que el cuerpo del texto, centrado en la parte superior de la hoja, en tanto los autores aparecerían con letras del mismo tamaño que el cuerpo del artículo, pero con tipografía diferente, centrados y abajo del título. De esta manera, para diseñar el sistema será necesario tener un corpus de entrenamiento en donde se marquen los metadatos de interés con sus características tipográficas y editoriales.

Edición. Si se quisiera construir un asistente para la edición de diferentes publicaciones, ya sea libros, normas, artículos, informes técnicos o folletos de difusión, con base en un cierto número de publicaciones existentes, será conveniente conocer los diferentes formatos de edición, si a renglón seguido, a doble columna, los tipos de letra para títulos, las sangrías, las marcas para elementos que se quieren resaltar, etc. Todos esos elementos metalingüísticos deben ser identificados para poderlos replicar.

Redes sociales. El análisis de los mensajes en redes sociales permite identificar problemas de interés para detección de riesgos, para conocer las opiniones sobre un producto o un programa, o para identificar posibles actos delictivos. Por ejemplo, en twitter no solo importa el mensaje sino quién lo dijo, a quién y con qué frecuencia ha escrito recientemente, así como la categoría en la que podemos clasificar el mensaje. Estos datos requieren ser etiquetados, de forma que se pueda hacer minería de los textos, tal como hacer un seguimiento de una noticia, de un autor determinado, de la periodicidad, etc.

Tomando en cuenta el objetivo del etiquetado, se etiqueta o anota un corpus para hacer énfasis en los aspectos lingüísticos que se quieren estudiar en él y para posteriormente recuperar la información necesaria según los patrones que se deseen. Etiquetar un corpus consiste en marcar (generalmente con las etiquetas de un lenguaje de marcaje, como XML) categorías léxicas, fonológicas, discursivas, sintácticas, etc.

Dependiendo del estudio, variará el etiquetado; así, si se requiere un estudio discursivo, el etiquetado adecuado será el que corresponde a este nivel; si el estudio es léxico o dialectológico, convendrá un etiquetado morfosintáctico. Además, hay que tomar en cuenta que en la mayoría de los casos se requiere tener información sobre el autor de los textos o del hablante, información que debe ir anotada en nuestro corpus.

7.1. Lenguajes de etiquetado

Si la anotación de los textos puede ser múltiple y variable por cada uno de los requerimientos, entonces las posibilidades de intercambio de información se acotarían a la documentación que habría que preparar. Existirían diferentes anotaciones como programas para codificar y decodificar los textos. Sin embargo, existen diversos lenguajes de etiquetado que a lo largo del tiempo se han modificado para ajustarse a las nuevas necesidades de marcado. En la siguiente cronología se explican algunos de los más usados y sus características.

GML. En 1969 nace el Generalized Markup Language (GML) como un lenguaje de IBM para resolver la necesidad de almacenar, procesar y clasificar las grandes cantidades de información de temas diversos que se habían estado generando en varias empresas y organizaciones. Este lenguaje separa la presentación o metadatos del contenido, de manera que facilita la clasificación de la información y la búsqueda de datos específicos. Las etiquetas describen el formato, la estructura y el contenido de los documentos, y diferencian entre párrafos, cabeceras, tablas, títulos, listas, etc.

SGML. Debido al éxito de GML y a que venía siendo ampliamente utilizado por cada vez más empresas, por el año 1986 se crea el estándar ISO 8879 de GML, capaz de adaptarse a un gran abanico de problemas. Sin embargo, tiene una sintaxis compleja y diferente según los modelos e instancias. El Standard Generalized Markup Language (SGML) permite anotar documentos de una forma altamente estructurada, mediante un conjunto de etiquetas definibles por el propio usuario.

TEI. En 1987 se crea la Iniciativa de Codificación de Textos o Text Encoding Initiative (TEI), bajo el patrocinio de la Association for Computers and the Humanities, la Association for Computational Linguistics y la Association for Literary and Linguistics Computing, con el fin de establecer recomendaciones comunes para el etiquetado de textos, pensando en el intercambio y reutilización de recursos en el ámbito de humanidades, ciencias sociales y lingüística, incluyendo imágenes y sonido. La estructura de árbol seguida en SGML resulta irrelevante en varios documentos en este ámbito, dada la organización de los textos en libros, capítulos, líneas o versos, fojas, artículos, etc.

HTML. Por el año 1989, para el ámbito de la red Internet, Tim Berners-Lee creó el HyperText Markup Language (HTML), o Lenguaje de Marcado de HiperTextos, que se basa en el estándar ISO SGML. El hipertexto señala una referencia cruzada en el mismo documento o en otro documento. Así, HTML incorpora los hiperenlaces o vínculos, a fin de navegar entre distintos documentos con información relacionada. Este lenguaje de marcado específico para compartir documentos en Internet

contiene un conjunto de elementos y atributos fijos a los cuales hay que sujetarse. Por ejemplo, se definen etiquetas específicas para marcar la tipografía, tales como las negritas o las cursivas. Con ello, no hay que ir reinventando las etiquetas, sino utilizar las que ya vienen dadas en una guía. Gracias a sus diferentes ventajas y facilidad de uso, fue adoptado rápidamente por la comunidad. Sin embargo, varias organizaciones comerciales crearon sus propios visores de HTML y riñeron entre ellos para hacer el visor más avanzado, inventándose etiquetas como su propia voluntad les decía. Por ello, el HTML creció de una manera descontrolada y no cubrió todos los problemas que planteaba la sociedad global de Internet.

XML. Desde 1996, XML (eXtensible Markup Language) es un metalenguaje desarrollado por el World Wide Web Consortium que define las reglas para la creación de lenguajes de marcas para codificar documentos particulares o tipos de mensajes. Pone en orden el HTML y establecer sus reglas y etiquetas para que sea un estándar. Este lenguaje tiene el poder de SGML pero de manera simplificada y venciendo las limitaciones de HTML. Por su importancia y por ser el de mayor uso para el etiquetado de corpus, en el siguiente capítulo se hablará de este lenguaje a detalle.

Por la importancia para corpus, conviene remarcar una diferencia básica entre el lenguaje HTML y el XML. Las etiquetas HTML en su principio tienen el objetivo de servir para la presentación de páginas web, y permiten señalar su estructura, su presentación y los hipertextos. En cuanto a estructura, se definen las funciones que tienen los diferentes elementos, como serían los títulos y subtítulos, los párrafos, o las listas, tablas y figuras. Las etiquetas de presentación establecen la apariencia del texto, como sería el tipo o forma de letra (negritas, cursivas, subrayado, color, etc.), el espaciado o la ubicación (centrado, justificado, etc.). El etiquetado hipertextual permite enlazar una parte de documento con otras partes dentro del mismo o con otros documentos. Para cualquiera que sea la etiqueta, ésta ya se encuentra definida previamente, de manera que cualquier navegador conoce de antemano el significado de cada etiqueta.

Por su parte, con XML se tiene un etiquetado descriptivo en el que las etiquetas no son únicas, sino que son definidas según las necesidades y especificaciones del proyecto. Por tanto, es necesario especificar el significado de cada etiqueta en lo que se conoce como DTD y la relación que existe entre las mismas. Más que identificar los elementos para darles formato, con XML se describen los datos para poder procesarlos.

7.2. Hacia la estandarización en la anotación

Se han hecho varios intentos para compartir información y regular los diferentes tipos de anotación. Entre los distintos grupos y asociaciones de ingeniería lingüística que han

buscado normalizar la codificación de corpus, cabe destacar:

LDC. El Linguistic Data Consortium (LDC) es un consorcio de varias universidades, centros de investigación, bibliotecas y compañías en Estados Unidos, creado en 1992. Entre otras actividades, promueve y mantiene una serie de recursos lingüísticos como lexicones, archivos de audio y sus transcripciones, y archivos de texto escrito. La Universidad de Pensilvania, que ha llevado el proyecto del Penn Treebank, ha sido la institución receptora del consorcio.

EAGLES. El Expert Advisory Group on Language Engineering Standards (EAGLES) es una iniciativa de la Comisión Europea creada en 1993 que comprende varios grupos trabajando en estandarizar:

- Recursos lingüísticos de gran escala (por ejemplo, corpus textuales, lexicones computacionales y corpus orales).
- Sistemas de tecnología lingüística, como los medios para manipular conocimiento lingüístico, etiquetamiento de los lenguajes y varias herramientas computacionales.
- Medios de evaluar los recursos, herramientas y productos.

Para llevarlo a cabo, se cuenta con cinco grupos de trabajo:

- Corpus textuales.
- Lexicones computacionales.
- Cualquier tipo de diccionario (flexiones, terminológico, etc.).
- Formalismos gramaticales.
- Evaluación de sistemas de procesamiento de lenguaje natural.
- Sistemas de lenguaje hablado.

TEI. Después de cuatro arduos años de trabajo coordinados por un Comité Directivo, se publican y difunden a mediados de 1990 las Normas TEI bajo el título de Guidelines for the Encoding and Interchange of Machine-Readable Texts. Estas Normas, que tienen como principio servir para el intercambio de información almacenada en cualquier formato electrónico para múltiples aplicaciones, son hoy en día la base para las Humanidades Digitales y, en principio, para la mayoría de los proyectos de corpus lingüísticos. Las Normas persiguen tres funciones:

- Guiar las prácticas locales o individuales en la creación de textos y captura de datos.
- Apoyar el intercambio de datos.
- Apoyar el procesamiento local independiente de la aplicación.

7.3. Principios sobre la anotación en corpus

Cualquiera que sea el tipo de anotación que se haga, existen lineamientos generales que deben cumplirse en cuanto a la anotación de corpus. Varios de estos lineamientos se han basado en las Normas TEI y en los diferentes grupos de normalización, entre ellos se encuentran los siguientes:

Inteligibilidad. La anotación que se realice debe ser distinguida del texto mediante etiquetas bien diferenciadas y únicas, de manera que sea relativamente fácil leíble por un humano el texto. Asimismo, las etiquetas deben tener un cierto nivel de lectura y entendimiento, por lo que debe evitarse tener nomenclaturas crípticas, sino por el contrario ser lo suficientemente claras y acordes con lo que se está etiquetando. Por ejemplo, para etiquetar sustantivo común singular femenino, en lugar de la etiqueta 3142, conviene scsf o sus_c_ sin_f.

Extracción. Debe ser posible remover la anotación de un corpus y convertirlo en corpus no anotado, en texto simple. Asimismo, también será posible extraer solo las anotaciones de un corpus y ser salvadas de manera independiente. Así, en caso de requerirse un análisis exclusivo de las categorías gramaticales de un texto, con el fin de identificar el estilo de un autor, por ejemplo, conviene tener solo las distintas categorías, sin tener el texto presente.

Intercambio. Con el fin de reutilizar los textos etiquetados de un proyecto determinado, se buscará que las etiquetas sean fácilmente reemplazables por las etiquetas utilizadas en otro proyecto.

Documentación. Las etiquetas deben estar basadas en documentación disponible para el usuario, que incluya el esquema de anotación, e incluso, se debe dar información sobre la confiabilidad y consistencia de la anotación seguida. El usuario final debe estar consciente de que la anotación de un corpus no es infalible, sino simplemente una herramienta poderosa. Asimismo, debe dejarse claro cómo y por quién fue realizada la anotación.

Estandarización. En el diseño de un corpus en particular, los esquemas de anotación deben estar basados en principios ampliamente definidos, de preferencia en consenso, en donde se eviten, en lo posible, cualquier tipo de interpretación subjetiva. Según el tipo de anotación que se haga, un corpus puede ser de utilidad para otros investigadores. Por ello es importante tratar de que las etiquetas sean compatibles o fácilmente reemplazables. Sin embargo, cabe mencionar que no deben considerarse los esquemas de anotación como una norma, ya que tienden a variar por razones prácticas. En tanto para un proyecto resultan útiles ciertos tipos de anotación, para otros pueden ser innecesarios o bien se pueden requerir otras etiquetas.

7.4. Conceptos básicos de etiquetado

Antes de comenzar a procesar un corpus, es necesario considerar los elementos que se desean etiquetar, el nombre que se les va a poner a cada una de las etiquetas y las características que tendrán los distintos elementos.

Entidad de marcaje

Lo primero a definir son los diferentes elementos que interesan etiquetar, pues solo estos serán los que se puedan en su momento procesar. Esto es, si se etiqueta un título, entonces se podrá tener un hipertexto para llegar a ese título; si la etiqueta identifica los distintos autores de los documentos de un corpus, entonces será factible recuperar información específica sobre determinados autores; si se cuenta con una etiqueta que describe a un término y el área temática al que pertenece se podra obtener un listado de todos los términos que existen clasificados por las diferentes áreas temáticas.

Se define una entidad de marcaje a cualquier objeto concreto del texto que sea de interés para marcar. Cada entidad tiene un nombre o posee una referencia. Las entidades pueden estar compartidas por distintos documentos y se encuentran organizadas en el texto mediante una estructura lógica y jerarquizada. Algunos ejemplos de entidad podrían ser:

- Un caracter o conjunto de caracteres.
- Una palabra o una serie de palabras.
- Una línea.
- Un párrafo.
- Un dibujo, gráfica, tabla, etc.
- Una nota de pie de página o a fin de documento.
- Un capítulo entero.
- El documento completo.
- Referencia de entidad.

Elemento de marcaje

Un elemento de marcaje se constituye por los elementos del documento que interesa anotar para su posterior procesamiento. Los documentos están conformados, en general, por los metadatos o información referencial que describe y precisa el origen del documento, y por el cuerpo del documento. Si se hace analogía con una ficha filológica, los metadatos informan la fuente del documento o los datos del entrevistado, así como cualquier dato pertinente que permita delimitar las variables del estudio. El cuerpo vendría a ser la información obtenida, el texto o la transcripción de la entrevista. En una ficha, tanto los metadatos vendrían referenciados ya sea por el título o por la ubicación o por un tipo o color de letra diferente; como el cuerpo del texto vendrían anotadas todas las características que se desean analizar, tales como los verbos conjugados, construcciones perifrásticas y uso de pronominales, en el caso de corpus textuales, o las pausas y los cambios de voz, en el caso de corpus orales. Todos estos constituyen los elementos de marcaje.

Los elementos de marcaje están normalmente delimitados por una etiqueta de apertura `<elemento>` y una etiqueta de cierre `</elemento>`. El nombre del elemento debe ir sin espacios vacíos y sin caracteres especiales. En los nombres de los elementos se distingue entre mayúsculas y minúsculas, de manera que será diferente `<nombre>` a `<Nombre>`. A continuación se ejemplifican algunas etiquetas XML posibles.

```
<titulo>El cañón de largo alcance</titulo>
<etiquetador>Teresita Reyes</etiquetador>
<italicas>ad hoc</italicas>
<SintagmaPreposicional>de la ciudad</SintagmaPreposicional>
```

Atributo de marcaje

El atributo de un elemento de marcaje proporciona información adicional, dependiente de un elemento de marcaje. Los atributos tienen un nombre y un valor. Por ejemplo, para el caso de la etiqueta fecha, podemos tener como atributos fecha de captura, fecha de registro, fecha de inicio y de término, fecha de revisión, etc. El valor será la fecha correspondiente, que tendrá definido un formato, ya sea DD/MM/AA, AAAA, o día mes y año para el caso de fechas, pero puede ser numérico, alfanumérico, etc. En el caso de XML, los nombres de los atributos se separan del elemento de marcaje con un espacio en blanco; los valores de los atributos se introducen con un signo igual y van entre comillas. Además, los atributos sólo se escriben en la etiqueta de apertura, no en la de cierre. Algunos ejemplos de etiquetas XML con atributos y sus valores se describen a continuación.

```
<hablante edad="24" sexo="masc">Francisco García</hablante>
```

```
<token lema= "saber" POS="verbo conjugado">sabría</token>
<obra generoLit="poesía">Primero sueño</obra>
<CD num="34" idioma="esp">El perro es un cánido</CD>
```

Referencias de entidad

Un texto está conformado por una secuencia de caracteres: letras, dígitos, signos, espacios, símbolos; de igual manera, las etiquetas del corpus utilizan secuencias de caracteres.En general, los lenguaje de etiquetado utilizan los caracteres con base en el estándar ASCII (American Standard Coding for the Interchange of Information). Los caracteres especiales como letras diferentes al inglés (á, ü, ñ), símbolos matemáticos, signos ($, &, '), letras griegas, etc., se transforman en representaciones ASCII y se llaman referencias de entidad. Una referencia de entidad sirve como nombre único para una pieza de datos y está compuesta por un ampersand (&), el nombre de la entidad y un punto y coma (;). El nombre de la entidad es normalmente un conjunto de letras relacionada con el carácter que se hace referencia, pero también puede ser substituida por una referencia numérica (con &#) o una hexadecimal (con &#x). A continuación (tabla 7.1) un ejemplo de equivalencias de algunas referencias de entidad.

Nombre	Número	Hexadecimal	Símbolo
Ú	Ú	Ú	Ú
ú	ú	ú	ú
ü	ü	ü	ü
ñ	ñ	ñ	ñ
β	β	β	β
≠	≠	≠	$\neq$
¡	¡	¡	¡
Σ	Σ	Σ	$\sum$
♠	♠	♠	♠
"	"	"	"
'	'	&#x27	'
<	<	<	$<$
>	>	>	$>$
&	&	&	&

Tabla 7.1: Referencias de entidad de algunos caracteres especiales.

Para el etiquetado de corpus, algunos caracteres son elementos de marcaje, por lo que a la hora de transcribir un texto que incluye uno de estos caracteres es necesario utilizar la referencia de entidad para evitar confusiones de interpretación. Para el caso de XML, los siguientes caracteres son elementos de marcaje: comillas, apóstrofe, menor que, mayor que y ampersand.

Comentarios

Los comentarios se usan en un documento SGML/XML para presentar información que técnicamente no forma parte del contenido de ese documento. Se usan para proporcionar descripciones de datos de documentos para provecho del usuario, y pueden mostrarse en cualquier parte del documento en la que aparezcan datos de caracteres analizados sintácticamente. Los comentarios empiezan con <!— y terminan con —>. La única limitación que tienen es que no se pueden incluir guiones altos (–), ya que entrarían en conflicto de sintaxis.

7.5. Referencias

Lecturas sugeridas

Arrarte, Gerardo (1999). “Normas y estándares para la codificación de textos y para la ingeniería lingüística”. En J.M. Blecua et al. (Eds.), Filología e informática: Nuevas tecnologías en los estudios filológicos. Barcelona: Editorial Milenio-Universidad Autónoma de Barcelona, pp. 17-44.

Barcala, F. Mario, Cristina Blanco y Victor Manuel Darriba (2006). “Metodología para la construcción de córpora textuales estructurados basados en XML”. Procesamiento del lenguaje natural 36, pp. 9-16.

Kahrel, Peter, Ruthanna Barnett y Geoffrey Leech (1997). “Towards cross-linguistic standards or guidelines for the annotation of corpora”. En R. Garside et al. (Eds), Corpus Annotation: Linguistic Information from Computer Text Corpora. London; Longman, pp. 231-242.

Leech, Geoffrey (1993). “Corpus annotation schemes”. Literary and Linguistic Computing 8 (4), pp. 275-281.

Normas

- LDC: www.ldc.upenn.edu
- Eagles: www.ilc.cnr.it/EAGLES/intro.html
- TEI: www.tei-c.org/Vault/P4/doc/html/

Capítulo 8

XML

El uso de XML para la construcción y manejo de corpus lingüísticos ha prevalecido hoy en día, gracias a sus distintas funcionalidades y ventajas. Entre ellas, cabe mencionar las siguientes:

Metalenguaje. Es un lenguaje que permite la organización y etiquetado de los documentos, con el que se pueden definir otros lenguajes de etiquetado.

Etiquetas personalizadas. A diferencia de HTML, el usuario puede definir sus propias etiquetas según convenga al proyecto, las cuales solo tienen que estar debidamente declaradas. Las etiquetas son claras y concisas, y definen semánticamente la información.

Extensible. Se pueden incorporar nuevas etiquetas en cualquier momento, con lo que se hace extensible a diferentes campos de conocimiento y aplicaciones.

Compatibilidad. Es compatible con SGML, se puede comunicar con otras aplicaciones en diversas plataformas, a la vez que existe una gran variedad de programas que procesan documentos XML.

Fácilmente interpretable. Es interpretable fácilmente por las máquinas, pero también inteligible para los humanos, ya que por su estructura y lenguaje es sencillo de leer y entender, incluso para terceras personas.

Aplicaciones. Admite una gran variedad de aplicaciones para procesar los documentos, extraer sus datos y, en general, para manipularlos. Gracias a que la información se encuentra etiquetada en forma precisa y semántica es posible tener buscadores inteligentes.

Facilidad. Además de que el diseño de XML se prepara rápidamente, los documentos en este lenguaje son fáciles de generar, implantar y de procesar. Asimismo, su programación es muy sencilla.

8.1. Conformación de un documento XML

Mediante el etiquetado en XML se crea una estructura bien definida de los textos y presentada en forma de árbol, de manera que se tiene una base principal o raíz de la que se desprenden ramas y de cada una de estas se desprenden otras más. Un corpus puede estar creado por un solo documento, como sería por ejemplo el CORCODE, en donde en un solo documento se tiene el listado de todos los contextos definitorios, o como sería una obra de un autor. Sin embargo, en general los corpus están conformados por una colección de documentos, en donde cada documento integra una y solo una obra.

En cualquier caso, cada documento XML está conformado por un prólogo (incluyendo el encabezado) y un cuerpo.

Prólogo

El prólogo es una o más etiquetas al principio de un documento XML en donde se especifica que se trata de un documento XML, la versión del mismo, el tipo de documento y la especificación sobre la descripción de las etiquetas. La primera línea, mostrada abajo, indica la declaración XML que especifica la versión de XML y la codificación de caracteres usada. En general, los procesadores XML como los navegadores HTML utilizan el conjunto de caracteres Unicode. De las distintas codificaciones Unicode, la más ampliamente usada y recomendada es UTF-8, ya que soporta varios idiomas.

```
<?xml version="1.0" encoding="UTF-8"?>
```

Encabezado

El encabezado aparece inmediatamente después del prólogo y contiene la información extralingüística sobre el documento. En el caso de corpus de texto escrito, el encabezado puede describir los siguientes rubros:

Descripción bibliográfica del documento. Esta información permite ubicar el documento. Aquí se incluye la información bibliográfica del documento y el tamaño aproximado del texto.

Metodología de la codificación. Contiene información relevante sobre la relación que existe entre el texto anotado y las fuentes originales, además de los métodos y principios editoriales que se siguieron durante la transcripción del corpus. TEI distingue seis componentes:

- Descripción del proyecto y del propósito por el que fue codificado el texto.
- Descripción narrativa de los métodos usados en la creación del corpus.

- Descripción detallada de los principios y prácticas editoriales aplicados durante la codificación.
- Información detallada sobre las etiquetas aplicadas al corpus.
- Especificación de las referencias canónicas construidas para el texto.
- Definición de los códigos de clasificación para el texto dentro del corpus.

Caracterización bibliográfica del texto (perfil). Proporciona información sobre los diferentes aspectos que describen a un texto, como: datos de la creación del texto, los idiomas que aparecen en el texto y el tema del texto conforme a un tesauro o clasificación estándar.

Descripción de responsables. Se anotan los datos de los encargados de la transcripción, anotación y revisión del documento, así como las fechas y los cambios realizados en el texto después de una revisión.

En el caso de corpus orales, algunos de los datos que se pueden describir son:

Descripción de responsables. Al igual que en corpus textuales, conviene registrar los responsables en las diferentes etapas del procesamiento de los datos: entrevista, transcripción, codificación, etiquetado, alineación y revisión, así como las fechas en que se realizaron.

Descripción del hablante. Para los análisis lingüísticos en el caso de corpus orales es de gran importancia los datos del hablante, como serían la edad, género, nivel educativo y rasgos sociolingüísticos. Por ejemplo, el nivel de fluidez y claridad articulatoria podrían depender del uso de una dentadura postiza o de la falta de un diente.

Datos de la fuente. En el caso de ser una entrevista, el procedimiento de consulta, si fue espontánea, el soporte físico, la calidad acústica, etc. En caso de provenir de otro medio, los datos del mismo.

Además de estos datos, se pueden tener otros registros útiles y que ya se han mencionado. Tal es el caso de los aspectos legales y de derechos de autor. Para corpus orales convendrá anotar si se tuvo consentimiento del locutor antes o después de la entrevista, o si se pidió permiso de grabación. En texto, los derechos existentes de reproducción.

Cuerpo

El cuerpo del documento contiene la transcripción del propio texto con sus etiquetas. El cuerpo empieza con un elemento raíz, al que por supuesto debe añadirse al final el cierre del mismo. En las normas TEI se sugieren tres partes del cuerpo:

Body. Es el cuerpo del documento y es la información básica y obligatoria.

Front. La información que precede al cuerpo del documento, como portada, índice, dedicatorias, etc.

Back. Información posterior al body, como apéndices, bibliografía, índices temáticos, etc.

8.2. Elementos

Conviene ir declarando de los elementos superiores a los inferiores (superordinados a subordinados). Los elementos también pueden ir precedidos y seguidos por otros elementos de marcaje.

Contenido de los elementos

El contenido de un elemento XML puede ser de diferentes tipos:

Vacío. XML permite elementos que no contienen nada, como sería en imágenes o saltos de línea. Normalmente sería una etiqueta sencilla de solo cierre que termina con "$/>$", o bien una etiqueta que abre y cierra, las cuales pueden tener o no atributos. A continuación se muestra un ejemplo de los dos casos, respectivamente:

```
<línea_continua/>
<Documento tipo="novela"></Documento>
```

Simple o de texto. El elemento sólo contiene texto.

```
<titulo>Una hoja verde dentro del cajón</titulo>
```

Elementos. El elemento contiene uno o varios elementos, los cuales pueden ser vacíos o simples.

```
<CD>
<ID num="36"></ID>
<termino>contexto definitorio</termino>
</CD>
```

Mixto. El elemento contiene texto y además uno o varios elementos.

```
<CD>Usa<italica>XML</ italica >como lenguaje de marcado</CD>
```

Anidamiento

Los documentos XML deben guardar una estructura estrictamente jerárquica respecto a las etiquetas que delimitan sus elementos, es decir, los elementos deben estar correctamente anidados y no se pueden solapar entre ellos. Debe haber un elemento raíz o elemento documento, el cual es único y no aparece en el contenido de ningún otro elemento, salvo hasta el final del documento. Los elementos después de la raíz deben anidarse adecuadamente. Para ello, una etiqueta debe estar incluida en su totalidad dentro de otra, de manera que si la etiqueta de comienzo está en el contenido de otro elemento, la etiqueta de fin también debe estar contenida en el mismo elemento.

Un ejemplo de buen anidamiento es el siguiente:

```
<elem1><elem2>   </elem2> <elem3>   </elem3> </elem1>
```

Un ejemplo de elementos mal anidados se da a continuación:

```
<elem1><elem2>    </elem1></elem2>
```

Atributos

Los atributos incorporan las características o propiedades a los elementos de un documento XML, por lo que los atributos no pueden contener otros elementos. Los valores de los atributos deben estar encerrados entre comillas, ya sea simples (') o dobles ("). La comilla simple puede utilizarse si el valor contiene caracteres comillas dobles, y viceversa, por ejemplo:

```
<titulo nombre='El señor de "la rosa" '>
```

8.3. Definición de tipo de documento (DTD)

La gramática para los documentos XML se conoce principalmente como DTD, Document Type Definition (término más usado) y Document Type Declaration (término usado en la ISO), o en español Definición (o Declaración) del Tipo de Documento. La DTD se escribe en SGML y se representa como un simple archivo en el sistema. Entre sus funciones se encuentra la de definir las reglas correspondientes a las etiquetas que se han creado para un corpus y precisar los nombres de las etiquetas y el modelo del contenido. Por ejemplo, el orden de las ocurrencias y las reglas de anidación para una implementación SGML particular.

La DTD consta, generalmente, de tres partes: una etiqueta inicial, el contenido y una etiqueta final. El nombre del elemento aparece en las etiquetas inicial y final. Todas las declaraciones de la DTD están delimitadas por los caracteres `<! ...>`, por ejemplo:

```
<!ELEMENT seccion (#PCDATA | Nombre)*>
```

Es importante notar que un archivo con la DTD (extensión *.dtd) no es un documento XML, sino un archivo texto, compuesto de declaraciones. Por tanto, no tiene por qué empezar con la declaración de documento XML, puede empezar con un comentario o directamente con la declaración de los elementos.

Declaraciones

Una DTD especifica los elementos y entidades que aparecen en el corpus con sus atributos, sus valores y tipo de valor posible, así como sus relaciones jerárquicas. Existen cuatro tipos de declaraciones:

De entidades. La entidad es una referencia o abreviatura de un bloque de datos o caracteres y está conformada por un nombre y un valor.

De notaciones. Las notaciones proporcionan información adicional para detallar los atributos y las entidades. Definen las entidades externas que no van a ser analizadas por un procesador XML. Para la notación se utilizan directamente el nombre de la entidad.

De elementos. Se declaran los elementos permitidos, su tipo y los valores permitidos.

De atributos. Se señalan los atributos de cada elemento, su tipo y los valores permitidos.

Ubicación de la DTD

La DTD puede estar definida en dos formas básicas:

DTD externa. La DTD puede venir en un archivo aparte, lo que resulta de utilidad para varios archivos semejantes en estructura. Al ser un documento externo, puede entonces ser compartido por múltiples documentos y se simplifica al no estar repitiendo la misma DTD en todos los documentos, sino solo en uno. Después de la declaración de documento XML (en la línea 1) tiene que declararse el tipo de documento con el nombre del archivo con la DTD. Al tipo de declaración (DOCTYPE) le sigue la descripción del tipo de documento, luego SYSTEM y el nombre del archivo con la DTD entre comillas.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE nombreDTD SYSTEM "declaracion.dtd">
```

DTD interna. Una DTD puede venir en el mismo documento XML, y en tal caso, la DTD irá dentro del prólogo de cada uno de los documentos. Si bien esto resulta útil para el caso de textos que tienen diferente estructura, es poco frecuente. Después de la declaración de documento XML (en la línea 1) tiene que declararse el tipo de documento y a continuación la DTD entre corchetes.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE nombreDTD [
--aquí va la definición de la DTD--
]>
```

Asimismo, se da el caso en que se tiene una mezcla de las dos formas, esto es, parte de la declaración viene en el propio documento y el resto en una DTD externa, para la parte de archivos diferentes con cosas comunes.

8.4. Esquemas

Otra manera de definir la estructura de un documento XML es a través de esquemas. Los esquemas indican los elementos que se permiten en un documento y las combinaciones permitidas mediante una especificación formal. Gracias a los esquemas se define exactamente los nombres de los elementos permitidos en un documento, sus subelementos, atributos y relaciones.

8.5. Hoja de estilo

Una alternativa al DTD para presentar la información de un documento XML con un formato determinado en una pantalla son las hojas de estilo.

CSS. Las hojas de estilo en cascada (Cascading Style Sheets) describen el formato como aparecerán las entidades definidas en un documento.

XSL. Lenguaje de hojas de estilo (Extensible Stylesheet Language) diseñado para ser utilizado en la web.

8.6. Validación de XML

Se acostumbra usar un editor estructurado con un analizador sintáctico para ir validando e indicando las inconsistencias de las etiquetas que se van anotando en el documento con relación a la DTD.

Documento bien formado. Un documento bien formado es aquel que es sintácticamente correcto, esto es, que sus etiquetas están escritas conforme a las normas XML. Por ejemplo, estará mal formado si una etiqueta que abre no está debidamente cerrada, si los valores de un atributo no van entre doble comillas, etc.

Documento válido. Un documento es válido si cumple con la estructura predefinida en el DTD, esto es, si los datos entrantes con las normas definidas en el DTD se han estructurado correctamente, si las etiquetas empleadas se han definido previamente, etc.

Para que un documento sea válido es necesario que esté bien formado, pero no necesariamente al revés, pues un documento puede estar bien formado sin haber definido las etiquetas.

8.7. Ejemplificación de XML para el CORCODE

A continuación se muestran un ejemplo de documentos XML creado en el Grupo de Ingeniería Lingüística para el corpus CORCODE.

En el mismo archivo se tiene la DTD y el conjunto de contextos definitorios que conforman el corpus. En el segundo se muestra un fragmento de la hoja de esquema para los corpus CLI, CHEM y CSMX, que a pesar de tener estructuras diferentes, se trató de homogeneizar en una misma DTD. Ambos ejemplos se encuentran explicados en los comentarios.

Etiquetado en el CORCODE

El Corpus de Contextos Definitorios (CORCODE) está constituido por un archivo que contiene el conjunto de Contextos Definitorios (CD) en el cuerpo del documento y la DTD que describe las etiquetas. Estas etiquetas XML delimitan a cada CD de forma integral, junto con los atributos que brindan las características que especifican sus valores, estructura, etc. Así, algunos elementos se encuentran dentro de otro elemento, como una especie de subelementos. La estructura de un CD se representa en la siguiente figura 8.1, en donde se observa, por ejemplo, que la Predicación Verbal Definitoria (PVD) contiene cuatro constituyentes subordinados: el verbo definitorio (VD), el clítico se (SEmar), el verbo auxiliar (VAUX) y el nexo (NX).

DTD en el mismo documento

El prólogo del CORCODE junto con un ejemplo de tres contextos definitorios se muestra a continuación. Obsérvese que el primer renglón es obligatorio y es el código que iden-

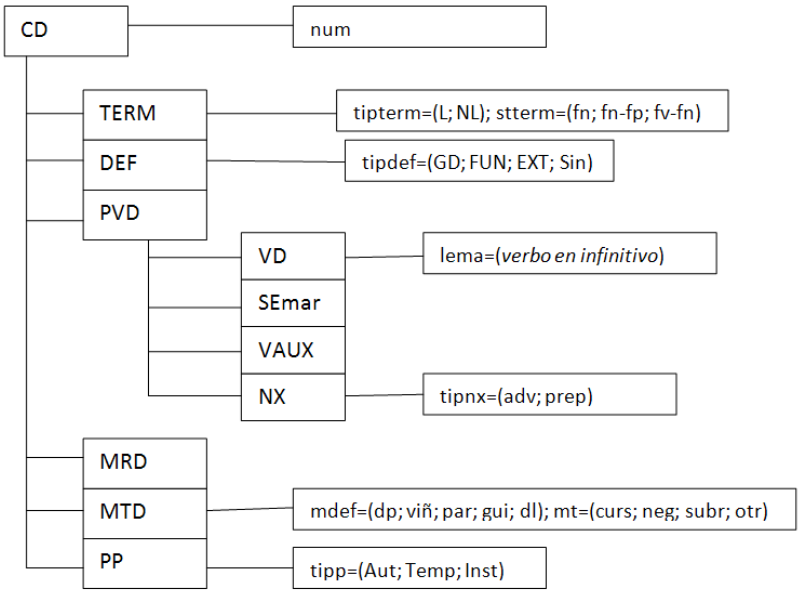


Figura 8.1: Estructura de un CD.

tifica que se trata de XML; además se agrega encoding="iso-8859-1" para poder aceptar caracteres del español.

```
|1 <?xml version="1.0" encoding="ISO-8859-1">
```

Ahora hay que declarar el tipo de documento, esto es, la DTD que puede venir a continuación o en otro archivo. En el caso de que venga a continuación, la DTD se introduce a partir del corchete.

```
|2 <!DOCTYPE CORCODE [
```

Si la DTD viene en otro archivo, en este caso CORCODE.DTD, entonces la línea dos sería:

```
|2 <!DOCTYPE CORCODE SYSTEM "CDCORPUS.DTD">
```

El primer elemento a declarar es la etiqueta raíz que al final debe también ser declarada como etiqueta de cierre. Contiene uno o más contextos definitorios.

```
|3 <!ELEMENT CORCODE (CD)*>
```

Se declara la etiqueta contexto definitorio y lo que puede contener, a saber: texto, predicaciones pragmáticas, marcadores discursivos, término, predicación verbal definitoria, marcador tipográfico definitorio, definición. Además, tiene un atributo, que es el número de CD.

```
|4 <!ELEMENT CD (PP|MRD|TERM|PVD|MTD|DEF)>
|5 <!ATTLIST CD num ID #REQUIRED>
```

La etiqueta término tiene dos atributos. El primero, el tipo de término, que puede ser lingüístico o no. El segundo, la estructura sintáctica, que puede ser frase nominal, frase nominal con frase preposicional o frase verbal seguida de frase nominal.

```
|6 <!ELEMENT TERM (#PCDATA)>
|7 <!ATTLIST TERM tipterm (L|NL) #REQUIRED>
|8 <!ATTLIST TERM stterm (fn|fn-fp|fn) #REQUIRED>
```

Se declara la predicación verbal definitoria con sus diferentes elementos: el marcador SE, un verbo auxiliar, un verbo definitorio, un marcador discursivo y un nexo. Luego se declaran cada uno de estos elementos, algunos de los cuales tienen atributos.

```
|9  <!ELEMENT PVD (#PCDATA|SEmarc|Vaux|VD|MRD|NX)*>
|10 <!ELEMENT SEmarc ANY>
|11 <!ELEMENT Vaux ANY>
|12 <!ELEMENT VD ANY>
|13 <!ATTLIST VD lema CDDATA #REQUIRE>
|14 <!ELEMENT NX ANY>
|15 <!ATTLIST NX tipnex (adv|prep) #REQUIRED>
|16 <!ELEMENT DEF ANY>
|17 <!ATTLIST DEF tipnex (GD|FUN|EXT|Sin) #REQUIRED>
|18 <!ELEMENT MRD ANY>
|19 <!ELEMENT MTD ANY>
|20 <!ATTLIST MTD mdef (dp|viñ|par|gui|dl) #REQUIRED>
|21 <!ATTLIST MTD mt (curs|neg|subr|otr) #REQUIRED>
|22 <!ELEMENT PP (#PCDATA>
|23 <!ATTLIST pp tippp (aut|Temp|Inst) #REQUIRED>
```

Cuando han sido declarados todos los elementos y atributos se cierra la DTD y empieza el cuerpo del documento con la etiqueta raíz, en este caso, CORCODE.

```
|24 ]>
|25 <CORCODE>
```

Fragmento del esquema del CORCODE

A diferencia de la DTD, el esquema es más complejo, pero permite ser más precisos en los datos y en el orden de los mismos. En el fragmento del esquema para el CORCODE

se usaron sangrías para facilidad de lectura, pero en realidad los espacios vacíos no son relevantes en la notación de la DTD o del esquema.

Al igual que la DTD, la primera línea del esquema comienza con la declaración de un documento XML. Luego se introduce la etiqueta raíz, CORCODE, que está conformada por uno o más contextos definitorios. La primera etiqueta corresponde al contexto definitorio, con los mismos seis elementos. En el esquema es necesario indicar el número de veces que puede haber de cada elemento. En el caso de contexto definitorio, debe aparecer al menos una vez, pero los elementos del contexto pueden no necesariamente aparecer en todos los contextos definitorios. Además se señala que la etiqueta CD tiene el atributo número.

```
|1 <?xml version="1.0" encoding="ISO-8859-1">
|2 <xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema
|3 targetNamespace="http://your_namespace"
|4 xmlns="http://your_namespace">
|5 <xsd:element name="CORCODE">
|6     <xsd:complexType>
|7  <xsd:sequence>
|8  <xsd:element ref="CD" minOccurs="1" maxOccurs="unbounded"/>
|9 </xsd:sequence>
|10    </xsd:complexType>
|11 </xsd:element>
|12 <xsd:element name="CD">
|13    <xsd:complexType mixed="true">
|14  <xsd:choice minOccurs="0" maxOccurs="unbounded">
|15  <xsd:element ref="TERM" minOccurs="0" maxOccurs="unbounded"/>
|16  <xsd:element ref="PVD" minOccurs="0" maxOccurs="unbounded"/>
|17  <xsd:element ref="DEF" minOccurs="0" maxOccurs="unbounded"/>
|18  <xsd:element ref="MRD" minOccurs="0" maxOccurs="unbounded"/>
|19  <xsd:element ref="MTD" minOccurs="0" maxOccurs="unbounded"/>
|20  <xsd:element ref="PP" minOccurs="0" maxOccurs="unbounded"/>
|21  </xsd:choice>
|22  <xsd:attribute name="num" use="required">
|23     <xsd:simpleType>
|24  <xsd:restriction base="xsd:token"/>
|25     </xsd:simpleType>
|26  </xsd:attribute>
|27    </xsd:complexType>
|28 </xsd:element>
```

Finalmente, y solo a manera de ejemplo para ver las diferencias con la DTD, se muestra el esquema para el elemento predicación verbal definitoria, así como para el elemento nexo y sus atributos.

```
|50 <xsd:element name="PVD">
|51    <xsd:complexType mixed="true">
```

```
|52  <xsd:choice minOccurs="0" maxOccurs="unbounded">
|53  <xsd:element ref="VD" minOccurs="0" maxOccurs="unbounded"/>
|54  <xsd:element ref="SEmarc" minOccurs="0" maxOccurs="unbounded"/>
|55  <xsd:element ref="VAUX" minOccurs="0" maxOccurs="unbounded"/>
|56  <xsd:element ref="NX" minOccurs="0" maxOccurs="unbounded"/>
|57  </xsd:choice>
|58    </xsd:complexType>
|59 </xsd:element>
|126<xsd:element name="NX">
|127   <xsd:complexType mixed="true">
|128 <!--<xsd:choice minOccurs="0" maxOccurs="unbounded"/> -->
|129  <xsd:attribute name="tipnx" use="required">
|130 <xsd:simpleType>
|131       <xsd:restriction base=xsd:token">
|132   <xsd:enumeration value="adv"/>
|133   <xsd:enumeration value="prep"/>
|134        </xsd:restriction>
|135         </xsd:simpleType>
|136  </xsd:attribute>
|137   </xsd:complexType>
|138 </xsd:element>
```

Muestra del CORCODE en XML

Ya sea para la DTD o para el esquema, un ejemplo del CORCODE en XML quedaría etiquetado de la siguiente manera.

```
|1 <CORCODE>
|2 <CD num="1">Los<TERM tipterm="L" stterm="fn">aparatos</TERM><PVD><SEmarc>se</SEmarc>
|3 <Vaux>pueden</Vaux><VD lema="caracterizar">caracterizar</VD><NX tipnx="adv"> omo</NX>
|4 </PVD><DEF tipdef="GD">poseedores de una impedancia definida de esta secuencia</DEF>;</CD>
|5 <CD num="2">Para obtener la estabilidad de una tubería se deben considerar fenómenos que
|6  pueden ocurrir simultáneamente, ya que el <TERM tipterm="L" stterm="fn-fp"> El soporte
|7  de Bluetooth </TERM> <PVD><VD lema="permitir"> permite </VD></PVD><DEF tipdef= "FUN">la
|8  conectividad inalámbrica para un gran número de eauipos (ordenadores, móviles, PDA's,
|9 etc.), a través de un protocolo industrial estándar </DEF>.</CD>
|10<CD num="3"><PP tipp="Aut">El análisis económico</PP> <PVD><VD lema="entender">entiende</VD>
|11</PVD><TERM tipterm="L" stterm="fn-fp">la acción del gobierno</TERM><PVD><NX tipnx="adv">
|12como </NX></PVD><DEF tipdef="GD">guiada por dos criterios que pueden resultar conflictivos:
|13 la equidad y la eficiencia</DEF>.</CD>
|14<CD num="4">Otro ejemplo de atención selectiva es <TERM tipterm="L" stterm="fn">el efecto
|15 cóctel</TERM><MTD mdef="dp" mt="">:</MTD> <DEF tipdef="GD">la capacidad para atender
|16  selectivamente a una sola voz entre muchas</DEF>.</CD>
|17</CORCODE>
```

La interfaz de consulta

Las etiquetas XML permiten que se puedan realizar, de manera fácil y rápida, búsquedas específicas de los elementos constitutivos de un CD, con base en los valores de sus atributos. En la siguiente figura 8.2 se muestra un ejemplo de búsqueda en la interfaz web.

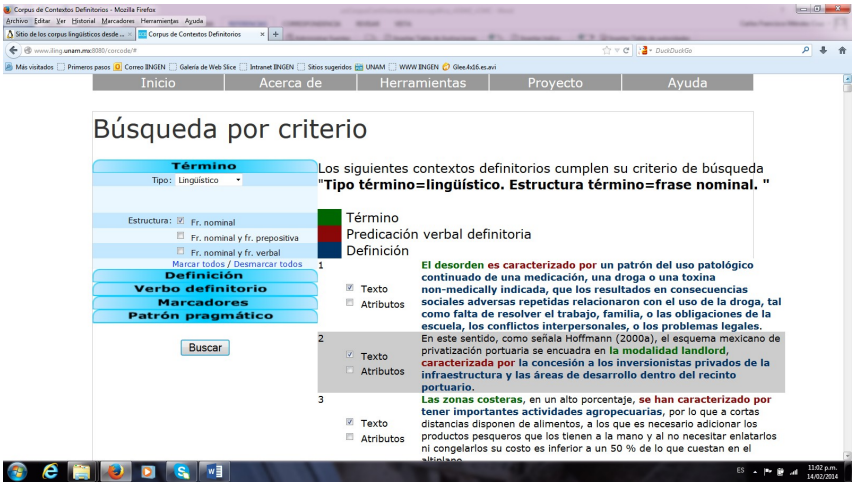


Figura 8.2: Interface de consulta de CORCODE.

8.8. Referencias

Lecturas sugeridas

Guerra, Javier Pérez (1998). Introducción a la lingüística de corpus: un ejercicio con herramientas informáticas aplicadas al análisis textual. Santiago de Compostela: Tórculo Edicións. (Véase sección 4.4).

Goldfarb, Charles F. y Paul Prescod (1998). The XML handbook. Oxford: Prentice-Hall.

Capítulo 9

Tipos de anotación

Los tipos de anotación que se pueden hacer sobre un corpus están determinados por los niveles de análisis de la lengua, y deben hacerse de acuerdo con ciertos principios. Los tipos de anotación de acuerdo con los niveles de análisis de la lengua se exponen en la siguiente tabla 9.1, junto con los subniveles correspondientes que seran explicados en los apartados siguientes.

Tipos de anotación o codificación	Textual	Estructura textual
		Tipología textual
	Fónica	Ortográfica
		Fonética
		Fonológica
		Prosódica
	Morfológica y morfosintáctica	Lematización
		Etiquecado POST
	Sintáctica	Parsing parcial o chunking
		Parsing total
	Semántica	Características semánticas
		Ontológica
	Discursiva	Relaciones semánticas
	Pragmática	Anafórica y referencial

Tabla 9.1: Tipos de anotación que se pueden realizar en un corpus.

9.1. Anotación textual

La anotación textual ayuda a los procesos de búsqueda y recuperación, y facilita el almacenamiento de la información en repositorios o bases de datos. Asimismo, permite una mejor visualización de los elementos del corpus y el orden en el que se pueden dividir o clasificar. Existen tres tipos de anotación textual diferentes.

Anotación por estructura textual

En primer lugar, la anotación por estructura textual. Esta anotación se refiere al marcaje de estructuras determinadas por el usuario, para el procesamiento del corpus o para ubicarse dentro de él. Así, pueden distinguirse las unidades en que se divide: los capítulos formados por secciones, éstas por párrafos y cada uno de ellos por oraciones. Se pueden diferenciar algunos párrafos particulares, como los títulos y subtítulos, las citas y ejemplos. Para el caso de corpus paralelos, por ejemplo, es común llegar al etiquetado del número de oraciones. En poemas se llegan a etiquetar las estrofas y los versos. Si bien en la mayoría de los corpus existentes no se utiliza, también puede etiquetarse el tipo (negritas, itálicas y subrayado) y muy rara vez el tamaño de letra (principalmente se marca cuando existe una diferencia de tamaño sobre el resto del documento) y el tipo de fuente (arial, helvética, times, etc.).

Anotación por tipología textual

Los elementos del corpus pueden etiquetarse por su tipología textual: artículo de revista, tesis, informe académico, poesía, novela, cuento, etc. En caso de requerir este nivel de clasificación, el usuario del corpus debe definir las tipologías textuales que utilizará y lo marcará de acuerdo con sus objetivos.

Anotación ortográfica

La anotación ortográfica o transliteración es la más común en corpus orales y consiste en asociar escritura común (ortográfica) a los elementos de un corpus. En este tipo de anotación se consideran aspectos como la acentuación, puntuación, división de las palabras, uso de mayúsculas y minúsculas, uso de siglas y abreviaturas, entre muchas otras, que dependen del proyecto.

9.2. Anotación fónica

El etiquetado fónico suele ser de varios tipos, aunque se puede delimitar. Una vez más, se etiqueta conforme a la finalidad del corpus. Sin embargo, hay tres tipos de

anotación que debemos mencionar en el marco fónico por ser los más comunes: el etiquetado fonológico, el fonético y el prosódico. En primer lugar, debemos diferenciar el etiquetado fonológico del fonético; en el fonológico se busca marcar los sonidos desde el punto de vista de su función en la lengua y establecer valores distintivos dentro del conjunto de sonidos que la componen; por otro lado, el etiquetado fonético busca describir la producción y percepción de los sonidos, tomando como punto principal sus manifestaciones físicas, es decir, a diferencia de la fonología, no busca agrupar sonidos para delimitar un sistema funcional, sino prepondera la descripción de los mismos para fines variados.

La anotación fonológica resulta de interés cuando se tienen textos sin norma académica, esto es, sin restricciones ortográficas, como sucede hoy en día en las redes sociales que se encuentra quesadilla como kesadiya o kezadilla. Asimismo, en el español antiguo no existía un estándar en las grafías, de manera que vivir podría escribirse como bibjir, vjbir o vivir. Gracias a la anotación fonológica y mediante el uso de un transductor se pueden realizar búsquedas en el corpus con diferentes representaciones fonológicas.

Por último, la anotación prosódica comprende el marcaje de elementos suprasegmentales, es decir, las marcas de tiempo, amplitud y frecuencia que afectan los segmentos y pueden producir cambio de significado, como lo son: acento, melodía, entonación, pausas, velocidad de elocución, ritmo y cualidad de la voz. Estos elementos también son llamados prosodemas.

Acento. Puede definirse como la prominencia de una sílaba en contraste con las que la rodean. Esta prominencia se manifiesta acústicamente y es percibida por los hablantes. Mediante el acento se establece un contraste entre sílabas prominentes y sílabas no prominentes.

Melodía. Es un elemento suprasegmental que se manifiesta en el nivel del enunciado.

Entonación. Se refiere a la variedad de manifestaciones de la frecuencia fundamental en la sílaba.

Pausa. Constituye una interrupción en la producción del habla. Existen pausas silenciosas (empty pauses) relacionadas con la respiración, y pausas sonoras (filled pauses) relacionadas con la planificación del discurso, por ejemplo, los alargamientos vocálicos y los elementos vocales: “eh”, “mmm”.

Velocidad. Es el resultado de la distribución temporal de los acentos y las pausas a lo largo de un enunciado. Repetición, alternancia o recurrencia de acentos, patrones melódicos y pausas.

Ritmo. Se determina por el número de segmentos o sílabas producidos por unidad de tiempo. Puede variar en un mismo locutor pues depende, entre otros factores, de la relevancia informativa de los elementos que configuran el discurso. La velocidad de elocución también puede reflejar estados emotivos del hablante. Es habitual realizar una distinción entre velocidad de habla y velocidad de articulación:

- Velocidad de habla (speaking rate), calculada a partir del tiempo total de emisión.
- Velocidad de articulación (articulation rate), calculada a partir del tiempo total de vocalización, excluyendo las pausas silenciosas.

Cualidad de la voz. Características globales debidas a los ajustes laríngeos y a los ajustes de las cavidades supraglóticas (ajuste articulatorio, articulatory setting) de un hablante.

9.3. Anotación morfológica

La anotación morfológica permite conocer cómo están constituidas las palabras y consiste en la identificación de los morfemas de una palabra y de sus rasgos como género, número, persona y modo. En español, dos fenómenos morfológicos resultan de interés: la flexión y la derivación.

Flexión. En la flexión se modifica una palabra canónica con variaciones funcionales o gramaticales de la palabra original, de manera que la categoría gramatical sigue siendo la misma. Por ejemplo, algunas flexiones de la palabra gato son gato, gatas, gatita. En la flexión nominal, los sustantivos presentan variación de género y de número, regularmente asociados a los sufijos –a, –e y –o, y a los sufijos –s y –es, respectivamente. En la flexión verbal tenemos la identificación de los tiempos, modos y personas, que para los verbos regulares se tienen patrones fijos, en tanto para los verbos irregulares no existen patrones.

Derivación. En la derivación se modifica una palabra canónica para formar otra nueva con un significado diferente y, en muchos casos, otra categoría gramatical. Por ejemplo, el sustantivo abogado puede derivarse en abogacía o abogar, en tanto el verbo medir se puede derivar en medida o medición.

Para la identificación de los morfemas y de la secuencia en que se ordenan éstos han existido diferentes métodos automáticos, pero no existe a la fecha un método estandarizado. Asimismo, existen programas que permiten flexionar y derivar automáticamente las palabras mediante los patrones regulares y uso de diccionarios. Sin embargo, de particular interés para diversos desarrollos de tecnologías del lenguaje se encuentra el proceso de lematización.

Lematización

Lematizar consiste en remitir a su forma canónica una palabra flexionada o una familia de palabras. Se puede definir como un proceso en el que se eliminan partes no esenciales de los términos (sufijos, prefijos) para reducirlos a su parte esencial (lema), con el fin de facilitar la indización y la consiguiente recuperación. A la palabra lematizada también se le conoce como forma de diccionario. Por ejemplo, la forma ser es la forma canónica que se convierte en etiqueta y se asocia a las flexiones soy, eres, fuiste, serás; en tanto, la forma canónica tigre se asocia a los sustantivos tigresa y tigres.

En los corpus es común anotar el lema de cada una de las palabras, seguido de la etiqueta que define la parte de la oración a la que pertenece, lo que se conoce como anotación morfosintáctica.

9.4. Anotación morfosintáctica

La anotación morfosintáctica o etiquetado de las partes de la oración (también conocido como POST por las siglas en inglés de Part Of Speech Tagging), consiste en anotar la categoría léxica o clase de palabra, es decir, si una palabra es un sustantivo, verbo, pronombre, etc. Es el tipo de anotación más común que suele añadirse a los corpus y permite refinar la precisión de las búsquedas de palabras, pues con él se pueden seleccionar usos tanto nominales como verbales de un lema y posibilita el acceso a métodos de codificación más sofisticados.

Existen cuatro puntos secuenciales a considerar para obtener el etiquetado de las partes de la oración:

- Identificación de las palabras o unidades léxicas a etiquetar.
- Definición de las clases de palabras, desde el punto de vista gramatical, que quiere realizarse.
- Definición de las etiquetas con las que se van a anotar las clases de palabras.
- Métodos para etiquetar las partes de la oración.

Identificación de palabras

Al hablar del etiquetado de las partes de la oración, es decir, de las palabras, parece evidente que palabra sea un concepto primitivo. Sin embargo, la realidad es que no resulta del todo trivial. No podemos decir que una palabra es el conjunto de caracteres entre espacios vacíos, pues entonces salta la presencia de los signos de puntuación (puntos,

comas, paréntesis, guiones cortos y largos, etc.) que van unidos a las “palabras”. Una respuesta inmediata sería excluir los signos de puntuación, pero tampoco es siempre posible. Existen tres posibilidades que van más allá de la correspondencia uno a uno entre la palabra ortográfica (aquélla entre espacios vacios) y la palabra morfosintáctica (aquélla cuya categoría gramatical vamos a analizar):

Multipalabras o unidades léxicas
Aquí existe una correspondencia de más de una palabra ortográfica con una palabra morfosintáctica. Esto es, existen unidades léxicas o conjuntos secuenciales de palabras que deben considerarse una sola unidad, ya que su sentido es diferente a la suma de los sentidos de las palabras que la componen. Entre las multipalabras cabe mencionar las locuciones: por favor (prep. + sust.) y sin embargo (prep. + sust.).

La anotación de las multipalabras permite llegar a análisis más confiables y precisos. Por ejemplo, mientras en un conteo de palabras en un corpus conviene considerar cada palabra por separado, en un conteo de palabras significativas hay que considerar las unidades léxicas. Asimismo, conviene señalar:

> **Los nombres propios,** incluyendo el título nobiliario, como Lic. Ezequiel Servando Urbina de la Tejera y Asociación Mexicana de Procesamiento de Lenguaje Natural, normalemente son considerados y anotados como una unidad léxica en muchos corpus. La identificación y categorización automática de nombres propios (nombres de personas, lugar, instituciones y empresas) es un tema de interés que sigue estudiándose.
>
> **Las fechas** (tres de diciembre de 1974, 03.12.74, 3Dic1974, 3-XII-74), las horas (17:30 hs., 5.30 PM, 1730 horas) y los números (317,645.25) tienen patrones más regulares y son anotados explícitamente.
>
> **Los términos,** como unidades significativas en su contexto, de caracter denominativo y valor referencial, llegan a estar combinados con números y otros signos; en ocasiones llegan a estar formados por siete u ocho palabras (*unidad de distribución de energía del subsistema de propulsión* (*PPDU*)). Véase el libro de Ana María Cardero (2003), *Terminología y Procesamiento*, UNAM.

Para anotar una unidad léxica como un solo elemento, existen *ditto tags*: se asigna la misma parte de la oración a cada palabra de la unidad y se señala el número de elemento que la conforman y el número en la secuencia que le corresponde a cada uno; por ejemplo:

```
de\_prep31 esta\_adj32 manera\_sust33
```

Acortamiento
Este término es equivalente a *mergers*. Se trata cuando existe una correspondencia de

una palabra ortográfica a más de una palabra morfosintáctica. Esto es, cuando en una secuencia de letras, algunas veces con signos ortográficos o signos de puntuación, se juntan dos o más palabras.

Clíticos. Los clíticos son un caso de acortamiento. Para el español, son las formas que la suceden (proclíticos) o la preceden (enclíticos) a una palabra determinada, como es el caso de los pronombres átonos. Así, cometelo = come+te+lo está dado por el imperativo segunda persona del verbo comer, más el dativo del pronombre personal de la segunda persona, más el acusativo del pronombre personal de la tercera persona. Se tiene una forma ortográfica para tres partes de la oración. Entre otras alternativas para etiquetar las contracciones y los clíticos se encuentran:

- Emplear picoparéntesis para mostrar la interdependencia de las palabras. Por ejemplo: come_vb> te_pp< lo_pp<.
- Dar la palabra junta y las etiquetas unidas a cada parte: come_vbte_pplo_pp.

Contracciones gramaticales. Las contracciones gramaticales son un tipo de acortamiento en donde se tiene una forma ortográfica para más de una parte de la oración. Por ejemplo, del = de(Prep.) + el(art.). En el caso del inglés se tienen casos en donde una forma ortográfica corresponde a tres palabras, como *dunno* = *do* + *not* + *know*.
Contracciones idiomáticas. Las contracciones idiomáticas separadas con apóstrofes es más común en inglés, pero también se dan casos en español, sobre todo en transcripciones de lengua hablada coloquial. Ejemplos del inglés: *don't* que puede separarse en las dos formas *do not*, pero todavía más complejo está el caso de los posesivos, en donde incluso cambia el orden de las palabras, como en *Zapata's head* = *head of Zapata*. Como ejemplo en español: p'al = para el.
Siglas. Las siglas son un ejemplo de acortamientos, en donde se tiene una forma ortográfica para refereirse a más de una palabra. Es común etiquetar las siglas, aunque normalmente no se le da ningún valor gramatical, el cual debiera anotarse para poder llevar un adecuado análisis sintáctico del contexto.
Abreviaturas. Si bien las abreviaturas no encuadran estrictamente en esta división por la correspondencia entre palabras ortográficas y morfosintácticas, ya que una abreviatura, en general, es una forma ortográfica a una morfosintáctica, nosotros las incluimos aquí por ser acortamientos; no obstante, cabe señalar que existen abreviaturas de multipalabras, como *V.gr.* o *R.S.V.P.*. Si bien es relativamente fácil identificar una abreviatura por el punto que la antecede, no siempre llevan el punto y pueden estar formadas por mayúsculas, minúsculas y/ o números, como dB = decibel, a = amperio, A = amperaje, H2O = agua. Es común anotar las abreviaturas como tales; sin embargo, hay que observar que éstas tienen un valor gramatical, que puede ser sustantivo o adjetivo.
Composición
La composición (compounds) consiste en la combinación de palabras completas para dar

origen a nuevas formas. Leech la considera como la correspondencia de una o más palabras ortográficas con una o más palabras morfosintácticas. La razón de que sea una o más palabras ortográficas es que, por un lado, puede escribirse de diferentes formas (por ejemplo, *eye strain*, como dos palabras; *eyestrain*, como una palabra; o *eye-strain*, como dos palabras separadas por un guión) y, por el otro, la composición llega a fosilizarse y perderse el sentido de unión de dos palabras, como el caso de *peliagudo*. Si bien se ha resuelto que la composición dada por la fusión de dos o más palabras, como *pelirrojo* que es resultado de la unión de pelo y rojo, sea considerada una sola palabra con su correspondiente parte de la oración, en el caso de los compuestos dados por dos palabras separadas con guión *hombre-rana* se llega a etiquetar cada parte por separado. En este último caso, cuando se realiza un etiquetado automático, hay que tener cuidado con los guiones que separan a dos palabras distintas, como en el caso de *San Luis Potosí-Puerto Vallarta*.

Definición de las clases de palabras

La categorización de las partes de la oración es un tema que aún entre los mismos gramáticos no se ponen de acuerdo. La definición y presición o detalle de las partes de la oración depende del objetivo particular para el que se está haciendo un corpus lingüístico. Se pueden definir desde unas cuantas, las más elementales con la información sintáctica elemental, hasta varias centenas, con una estructura más detallada que contemple los distintos aspectos morfológicos, las características de los verbos, etc. Sin embargo, cuando se busca construir un corpus multipropósito, es conveniente pensar en que las clases y subclases definidas lleguen a mayor detalle.

EAGLES proporciona una serie de recomendaciones para reconocer las partes de la oración en tres niveles:

Características obligatorias. Son aquellas partes de la oración básicas que deben ser anotadas en cualquier etiquetado de las partes de la oración. EAGLES reconoce las siguientes principales: sustantivo, verbo, adjetivo, pronombre/determinante, artículo, adverbio, aposición, conjunción, numeral, interjección, único (partícula negativa not y marcador de infinitivo to para el inglés), residual (por ejemplo, palabras extranjeras y símbolos matemáticos) y puntuación.

Características recomendadas. Aquellas categorías gramaticales ampliamente reconocidas y que deben ser anotadas de ser posible. Por ejemplo, para el sustantivo: número, género, caso y tipo (común o propio, por ejemplo).

Características opcionales. Aquellas que pueden ser usadas para propósitos específicos, pero que no son lo suficientemente importantes para ser consideradas obligatorias o recomendadas. Pueden ser de dos tipos:

- Características genéricas. Las que son aplicables a la mayoría de los lenguajes (oficiales de la Comunidad Europea). Por ejemplo, la subcategorización de sustantivos en contables, concretos, abstractos, colectivos, etc.
- Características específicas del lenguaje. Las que aplican a una o pocas lenguas.

Definición de las etiquetas morfosintácticas

Una vez definido el nivel al que se quiere llegar en la categorización de las partes de la oración, el siguiente paso es asignar las etiquetas correspondientes. Por ello, se sugiere que las etiquetas sean escogidas con base en tres criterios:

Concisión. Los nombres de las etiquetas deben ser breves, en preferencia a las etiquetas largas, aunque estas últimas sean más completas en cuanto a la descripción.

Perspicuidad. En el sentido de claras y transparentes, las etiquetas deben ser fácilmente recordadas e interpretadas.

Analizabilidad. Esto es, en los nombres de las etiquetas deberán ser distinguidas y separadas las partes lógicas y gramaticales que las componen.

Ejemplos de etiquetas morfosintácticas

Etiquetas usadas en el Penn Treebank corpus. Estas etiquetas fueron hechas para el inglés, pero existen adaptaciones para el español.

Etiquetas para el español conforme a EAGLES. EAGLES es un consorcio europeo, que intento hacer la anotación para todas las lenguas europeas. Permite grandes niveles de complejidad, pues su etiquetado abarca varios tipos de etiquetado y permite seleccionar de ella sólo algunas.

Etiquetas del British National Corpus. Tiene sus propios estándares de etiquetado, para el inglés.

Etiquetas del Proyecto Corpus, del IULA. Los etiquetarios diseñados para el marcaje del corpus del IULA fueron codificados con el estándar SGML y siguiendo las directrices marcadas por el *Corpus Encoding Standard* (CES) de la iniciativa EAGLES. Hay etiquetarios para las lenguas catalana, castellana e inglesa.

En el siguiente ejemplo se puede observar que se marcaron tanto clase de palabra como el género y el número. La tabla 9.2 presenta la descripción de cada etiqueta usada.

```
(1)La/AFS pérdida/NFS de/Pl/AMS cromosoma/NMS
20/X es/V3S lo/ANS más/D frecuente/
```

Palabra	Etiqueta	Significado de la etiqueta
La	/AFS	(artículo, femenino, singular)
pérdida	/NFS	(nombre o sustantivo, femenino, singular)
de	/P	(preposición)
l	/AMS	(artículo, masculino, singular (el))
cromosoma	/NMS	(nombre o sustantivo, masculino, singular)
20	/X	(número)
es	/V3S	(verbo, tercera persona, singular)
lo	/ANS	(artículo, neutro, singular)
más	/D	(adverbio)
frecuente	/J	(adjetivo)

Tabla 9.2: Ejemplo de etiquetado POST.

Métodos para etiquetar las partes de la oración

Con el fin de resolver las ambigüedades léxicas, el etiquetado de las partes de la oración (POST) puede realizarse con dos tipos básicos de algoritmos y uno híbrido:

Etiquetado basado en reglas. Usan una base de datos grande con reglas de desambiguación que indican, por ejemplo, que una palabra ambigua es sustantivo, en lugar de verbo, cuando va después de un determinante. El método basado en reglas consta de dos etapas. En la primera etapa, se ejecuta un programa para identificar las posibles partes de la oración de cada palabra, a partir de un lexicón en donde a cada palabra le corresponde su o sus partes de la oración. En la segunda etapa, se ejecuta un programa con un conjunto de reglas (1,100 para el inglés, aprox.) aplicadas a las palabras ambiguas.

Etiquetado estocástico. Usa un corpus entrenado para calcular la probabilidad de que una palabra tenga cierta etiqueta dado un contexto determinado. El fundamento de los métodos estadísticos está dado por una generalización de "escoja la etiqueta más probable de esta palabra", basada en el enfoque Bayesiano. Para una oración o secuencia de palabras dadas, los algoritmos basados en las cadenas de Marcov seleccionan la secuencia de etiquetas que maximice el siguiente producto:

$$P(palabra \mid etiqueta) * P(etiqueta \mid n\ etiquetas\ previas)$$

Los modelos basados en las cadenas de Marcov seleccionan una secuencia de etiquetas para una oración completa, más que para una palabra sola.

Etiquetado basado en transformación. El más conocido es el desarrollado por Eric Brill y se conoce como el *Brill tagger*, que comparte características de los dos algoritmos anteriores. Este método asigna las categorías gramaticales a las palabras del corpus con base en las reglas de transformación, es decir, reglas de etiquetado que cambian una etiqueta por otra de acuerdo con una determinada condición. Las condiciones se expresan en forma de plantillas de transformaciones, las cuales pueden ser de dos tipos: las léxicas (que toman en cuenta una secuencia de longitud predefinida de letras finales o iniciales de la palabra, como una representación simplificada de su estructura morfológica) y las contextuales (que consideran las categorías gramaticales de las palabras vecinas, como una representación simplificada de la estructura sintáctica). Las reglas de transformación se obtienen por medio de la comparación del etiquetado inicial realizado por el programa con el etiquetado manual de un experto. Se ponen a prueba todas las posibilidades de etiquetado para cada palabra y se miden estadísticamente las mejorías que se producen. De esta manera, se seleccionan las reglas que se relacionan con el mayor número de mejorías en el etiquetado. El proceso se repite nuevamente hasta obtener una calidad del corpus suficientemente cercana al corpus etiquetado por el experto. Así, el método genera un conjunto de reglas ordenadas que indican las condiciones y contextos en los cuales deberá ser asignada una etiqueta.

9.5. Anotación sintáctica

Comúnmente, el paso que sigue al marcaje de partes de la oración es la anotación sintáctica. Este proceso consiste en encontrar las relaciones sintácticas entre dichas partes, lo que constituye el análisis de la oración o parsing. El proceso de parsing tiene dos etapas; el parsing parcial o chunking y el parsing total.

Parsing parcial o chunking

El chunking es un análisis de constituyentes sintácticos básicos; puede entenderse como “romper el texto en pedazos”. Es altamente preciso, pues normalmente ignora el contenido léxico y sólo requiere identificar las partes de la oración, en la que describe los patrones en los sintagmas nominales, verbales, preposicionales y adverbiales.

Este nivel de etiquetado es suficiente para muchas aplicaciones del Procesamiento del lenguaje natural (PLN). En él, se hace uso de patrones de etiquetas para construir

reglas; por ejemplo, para un sintagma nominal se puede definir la unión de un artículo con un sustantivo:

```
la/AFS(Art) + pérdida/NFS(Sust)= Sintagma Nominal [la pérdida]
(Chunk: Sintagma Nominal)
```

donde AFS=Artículo Femenino Singular y NFS=Nombre Femenino Singular.

Características de los chunks:

- Los chunks son regiones del texto que no se intersectan

  ```
  En [un lugar] de [la Mancha] de [cuyo nombre] no quiero acordarme
  ```

- Los chunks son no recursivos, esto es, un chunk no puede contener otro chunk.
- Por el contrario, los constituyentes son recursivos

  ```
  ([un lugar de [la Mancha]])
  ```

- Los chunks no son exhaustivos, es decir, no todas las palabras están incluidas en éstos.

Parsing total

Por su parte, el parsing total es un proceso que tiene por objetivo el etiquetado sintáctico y consiste en un análisis completo de constituyentes de la oración y sus relaciones sintácticas, de acuerdo con las reglas de una gramática. Como resultado de ese análisis se obtiene un árbol sintáctico o su representación mediante paréntesis categorizados, como el ejemplificado:

```
[O [SN El_Art hombre_Sust SN] [SV vio_Vb [SP a_Prep [SN la_Art
nena_Sust SN] SP] [SP en_Prep [SN el_Art parque_Sust SN] SP][SP
con_Prep [SN el_Art telescopio_Sust SN] SP] SV]O]
```

Una desventaja del parsing total es que es poco preciso (ambiguo), adaptable a un dominio específico y lento en realizarse. Asimismo, llega a ser muy costoso y de poco beneficio para PLN.

Para el parsing total se puede hacer uso de diferentes gramáticas (independiente del contexto, de dependencia o funcional).

9.6. Anotación semántica

Se pueden tener varios objetivos de anotación semántica, entre ellos están desambiguar las palabras del corpus (Word Sense Disambiguation), es decir, asignar a cada palabra el sentido más apropiado del diccionario, y otro objetivo es detectar las relaciones léxico-semánticas de las palabras del corpus.

Es prudente aclarar que no existen estándares establecidos para la anotación semántica. El tipo de marcaje es definido por los usuarios del corpus, de acuerdo con los objetivos de la investigación que se realice. A pesar de lo anterior, si se toman en cuenta características muy generales de los diferentes tipos de anotación semántica existentes es posible ubicarlos en tres grandes grupos: características semánticas, anotación ontológica y anotación de relaciones semánticas.

Características semánticas

Se puede anotar las características semánticas de una palabra, entendiendo esas características como los significados de las palabras. También se puede agregar a las palabras del corpus identificadores de entidades para que puedan ser distinguidas de acuerdo con sus características semánticas; por ejemplo, tenemos la anotación de predicados en la que el verbo agacharse se puede asociar con el marco semántico general de movimiento y al escenario conceptual de cambio de postura.

Anotación ontológica

Es de suma importancia para los estudios en materia de anotación semántica de documentos de la web. En este tipo de marcaje se utilizan tanto las características como las relaciones semánticas de las palabras. La novedad radica en que es posible hacer referencias a esas características y relaciones por medio de metainformación o metadatos que se agrega al contenido de las páginas web. Dicho de otra forma, la anotación ontológica consiste en hacer la descripción formal de los conceptos del corpus y enlazar las relaciones entre esos conceptos. Así pues, con la anotación ontológica se pretende conseguir el acceso inteligente a diversos recursos y que la navegación y búsqueda de información en internet sea más fácil y rápida.

Anotación de relaciones semánticas

Se refiere al marcaje de relaciones léxico-semánticas que se pueden establecer entre elementos del corpus y que van desde la sinonimia, hiperonimia, meronimia, etc., hasta las relaciones de elementos relacionables del texto tales como: agentes, pacientes

y participantes de acciones concretas. Algunas de las relaciones léxicas se exponen a continuación:

Homonimia. Es la relación que existe entre palabras que tienen la misma forma, pero con significados no relacionados. Los elementos que tienen este tipo de relación se llaman homónimos. Por ejemplo, banco (institución vs. conjunto de peces). Los homónimos que tienen diferentes partes de la oración no son problemáticos a la lingüística computacional, pues un tagger distinguiría, con base en sus elementos sintácticos, el significado correspondiente a cada homónimo.

Polisemia. Es el fenómeno de múltiples significados relacionados para un mismo lexema. Se puede llegar a confundir con homonimia, pero aquí se trata de lexemas que comparten mismos semas, como banco (institución/ grupo de peces/ asiento). Cabe señalar que, computacionalmente, polisemia y homonimia no guardan grandes diferencias, y ocasionalmente, ciertas palabras que han perdido transparencia (es decir, posibilidad de análisis) pueden confundirse como homónimas, cuando realmente son polisémicas.

Homofonía. Se refiere a los distintos lexemas que se escriben diferente pero tienen la misma pronunciación. Los homófonos presentan dificultades en diferentes aplicaciones de la ingeniería lingüística, como corrección de escritura, reconocimiento de voz y sistemas de recuperación de información.

Sinonimia. Se define como los diferentes términos que, en cierto contexto, tienen el mismo significado. Desde nuestro punto de vista, dos lexemas son sinónimos si estos pueden ser sustituidos uno por el otro en una oración sin cambiar el significado o la aceptabilidad de la oración. El adecuado uso de una palabra en función de la noción de identidad de significado es importante para la extracción y recuperación de información.

Antonimia. Se dice de las palabras que expresan ideas opuestas o contrarias. Por ejemplo, blanco-negro, día-noche, etc. Pese a que semánticamente guardan un significado opuesto, los pares de antónimos siempre pertenecen a una misma categoría léxica, es decir, que el antónimo de un verbo será un verbo, el de un sustantivo un sustantivo, etc.

Hiponimia. Se refiere a pares de lexemas donde uno denota una subclase del otro. Por ejemplo, la relación entre cánido y perro es del tipo hiponímico. Al no ser una relación simétrica, se usa el término hipónimo a lexema más general, e hiperónimo al lexema más específico.

Hiperonimia. Se dice que una palabra es hiperónima de otra cuando su significado la incluye. Constituye una taxonomía de las lenguas naturales, pues describe una

manera de agrupar y ordenar el mundo. Un ejemplo es la palabra mamífero, en la que se incluyen los términos 'perro', 'gato', 'hombre', etc.

Meronimia. Es una relación semántica entre un lexema que denota una parte correspondiente a un todo, en tanto holonimia es la relación entre un lexema que denota el todo correspondiente a una parte. Al primer tipo se le conoce como relación hasa (tiene un/una). Por ejemplo, 'el automóvil tiene un volante'.

Holonimia. Es la relación semántica antagónica a la meronima. Se dice que una palabra es holónima de otra cuando la abarca, cuando esta palabra es una de las partes que la constituye. Por ejemplo, la palabra bicicleta es holónima de 'llanta', 'pedal' y 'manubrio'.

9.7. Anotación discursiva

La anotación discursiva en corpus lingüísticos es de las áreas en que menos se ha incursionado por parte de quienes trabajan corpus lingüísticos informatizados. El avance se ha limitado a la creación de sistemas de etiquetado que sean útiles para ubicar elementos discursivos dentro de los textos, tales como emisores y receptores, el tema o los temas en construcción, las normas que regulan la situación y los efectos que la comunicación produce. Sin embargo, cabe aclarar que estas marcas son definidas por el usuario y no existe estandarización alguna para ellas, caso contrario a la anotación fónica, por ejemplo, en la que existen convenciones de marca como son los alfabetos fonéticos y las marcas de rasgos suprasegmentales.

Anotación de relaciones discursivas

La Rhetorical Structure Theory (RST) es una teoría que permite describir la estructura discursiva de cualquier texto, de manera jerárquica en forma de árbol, y se basa en la existencia de dos tipos de relaciones discursivas: (a) núcleo-satélite o mononucleares, si una unidad mínima discursiva depende de la otra, y (b) núcleo-núcleo o multinucleares, si las dos unidades tienen la misma importancia de acuerdo con el propósito del autor. Ejemplos de relaciones mononucleares serían causa, propósito, etc. y de relaciones multinucleares unión, contraste, etc. La lista de relaciones a considerar depende del autor y de los propósitos que tenga su investigación.

En la figura 9.1 se muestran las etiquetas y las relaciones retóricas para un texto de matemáticas seleccionado del corpus RST Spanish Treebank.

Algoritmo para resolver exactamente sistemas de ecuaciones lineales con coeficientes enteros En este trabajo se presenta un algoritmo para resolver sistemas de ecuaciones lineales con solución única, cuando sus coeficientes son números enteros. Siendo una variante de la eliminación Gaussiana, posee características didácticas ventajosas sobre la misma. Durante el proceso, que utiliza solo aritmética entera, se obtiene el determinante de la matriz de coeficientes del sistema, sin necesidad de cálculos adicionales.<group id="6" type="span" parent="2" relname="elaboración" /><group id="7" type="span" parent="8" relname="span" /><group id="8" type="span" parent="9" relname="span" /><group id="9" type="span" />

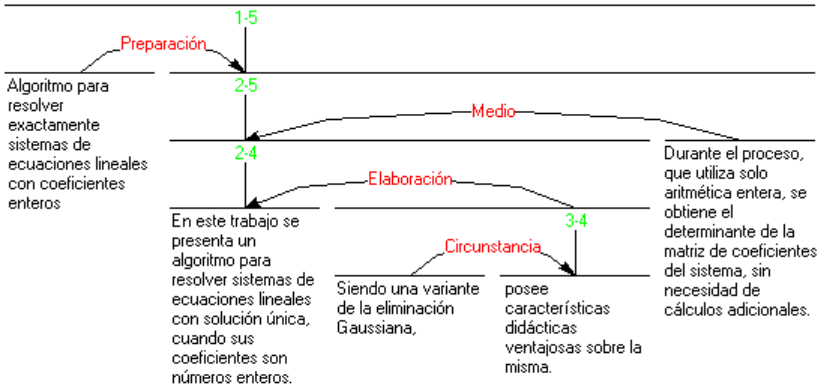


Figura 9.1: Etiquetas y relaciones retoricas para un texto del corpus RST Spanish Treebank.

Anotación de correferencias

Uno de los problemas principales que se busca solucionar por medio de sistemas de marcaje en la anotación discursiva es la resolución automática de anáforas. La anáfora es el hecho de hacer referencia a algo mencionado con anterioridad, de manera que intervienen dos elementos básicos: la expresión anafórica, que es la entidad que refiera a lo mencionado antes, y antecedente, que es justo ese algo de lo que previamente se mencionó. Por ejemplo, en la frase “me gusta el libro, ya lo leí”, el pronombre “lo” está haciendo referencia a “libro”.

Un reto que se presenta en la lingüística computacional consiste en resolver la resolución automática de anáforas, esto es, detectar automáticamente el antecedente de una expresión anafórica. Por ello, en la actualidad, se ha realizado mucho trabajo en marcar

de manera manual las relaciones de correferencia entre elementos del corpus. De esa forma será posible buscar patrones que permitan delimitar reglas y que, eventualmente, resulten en la obtención de relaciones automáticas entre anáforas y sus correferencias.

Un ejemplo de etiquetado de relaciones anafóricas en un texto complejo con varias expresiones anafóricas y diferentes antecedentes se puede apreciar a continuación:

```
<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.
```

9.8. Anotación pragmática

La pragmática estudia los usos de la lengua, esto es, los procesos que se activan durante la codificación o decodificación de los enunciados en un contexto dado. No es lo mismo lo que se dice (que puede analizarse mediante la sintaxis o la semántica), que lo que se quiere decir (que requiere una anotación pragmática). El enunciado “¿me puedes pasar la sal?” va más allá de una pregunta, pues se trata de una petición.

La anotación pragmática se restringe al marcaje de elementos que el usuario del corpus desea analizar. Las marcas de actos de habla, entonación, pausas e, inclusive, la intensidad y los gestos con que se produjo cierta muestra lingüística son ejemplos de anotación pragmática. Asimismo, como parte de la anotación pragmática se encuentra aquella que permite identificar la ironía, los sentimientos y las opiniones, entre otras.

Anotación de actos de habla

Los actos de habla son unidades de análisis sobre las diferentes acciones que se realizan con el lenguaje, tales como el afirmar, negar o corregir algo, el pedir o el ofrecer, el ordenar, el preguntar o el consentir. Todas estas acciones tienen, por tanto, un significado y resultan de interés para distintos proyectos.

En el corpus DIME, por ejemplo, con el fin de desarrollar un sistema conversacional, se anotaron las intenciones comunicativas de los diálogos, tomando como base cada uno de los enunciados. Para segmentar los enunciados se apoyaron en los cambios en el tono

de voz, las pausas, las interjecciones y chasquidos, entre otros elementos. Gracias a la anotación de los diferentes actos de habla de su interés ha sido posible construir un robot de servicio que interprete la intención de la persona con quien se comunica.

Anotación de polaridad

Una línea reciente de la ingeniería lingüística es la detección de emociones, de sentimientos y de opiniones, lo cual tiene varias aplicaciones. La construcción de corpus anotados con etiquetas sobre polaridad, intensidad, emoción o la opinión positiva o negativa, entre otras, resulta vital para el desarrollo de varios sistemas. Distintos grados de complejidad entran según el nivel del segmento a etiquetar, pues en algunos casos será importante llegar a nivel de frase, en tanto en otros será suficiente tener un etiquetado a nivel de todo el texto completo, ya sea una noticia o un boletín. Entre las etiquetas de emoción podemos encontrar: tristeza, miedo, ira, placer, vergüenza, humor, enojo. De la polaridad de opiniones tenemos: positiva, negativa, neutra o sin opinión. Un problema común es la subjetividad de los anotadores, razón por la que es usual que cada segmento deba ser anotado por dos personas al menos. Esta situación se complica cuando se tienen textos cortos, como en el caso de twitter, y cuando existe ironía por parte del autor del texto.

9.9. Referencias

Lecturas sugeridas

Garside, Roger, Geoffrey Leech y Tony McEnery (1997). Corpus Annotation: Linguistic Information from Computer Text Corpora. New York: Addison Wesley Longman. (Véase capítulo 1).

Grefenstette, Gregory y Pasi Tapanainen (1994). “What is a word, What is a sentence? Problems of Tokenization”. Proc. 3rd Conference on Computational Lexicography and Text Research (COMPLEX’94), Budapest, pp. 79-87.

Navarro-Colorado, Borja (2007). Metodología, construcción y explotación de corpus anotados semántica y anafóricamente. Tesis Doctoral, Universidad de Alicante, España.

Parte IV

Herramientas y técnicas de análisis

Capítulo 10

Técnicas de análisis

10.1. Conteo de palabras

La identificación de las palabras de un texto es un proceso que se llama segmentación. El punto de partida para el conteo de palabras consiste en conocer la diferencia entre token y type. El token o palabra es cada una de las formas que aparecen en el texto, sin importar cuántas veces ocurra cada una. El número total de tokens definirá el tamaño del corpus. Por su parte, el type se refiere a cada una de las formas o palabras diferentes que aparecen en un texto. Se acostumbra indicar la frecuencia absoluta de cada una de los types. La relación entre los type y token de un texto nos dará su riqueza léxica. La suma de las frecuencias de todos los types será la suma de todos los tokens de un corpus.

Listas de palabras

Definiremos lista de palabras al listado de los types o formas que se escriben diferentes en el corpus. A cada palabra le acompaña al menos su frecuencia absoluta, esto es, el número de veces que se repite dicha palabra en el corpus. Adicionalmente se puede acompañar de la frecuencia relativa, que es el número de veces que ocurre la palabra en relación con el total de palabras en el corpus. Su valor es igual a la frecuencia absoluta entre el número total de tokens (o la suma de las frecuencias absolutas).

En el caso de corpus anotados con partes de la oración, se puede acompañar a cada type con la categoría gramatical que le corresponde. Palabras homónimas con distinta categoría gramatical se presentarán en diferentes líneas.

Tipos de listas

Lista simple. Esta es la lista más común de todas. Aquí en cada línea se presenta un type diferente, normalmente acompañado de su frecuencia en el corpus. Sus aplicaciones son muy diversas.

Lista de formas canónicas. Esta lista está ordenada alfabéticamente por cada uno de las formas canónicas o entradas de un diccionario, y a cada línea le suceden las palabras que le corresponden a dicha forma canónica. Como ejemplo, el trabajo de Juan López Chávez y Marina Arjona de Lexicometría y fonometría del Primero Sueño de Sor Juana Inés de la Cruz.

Lista de lemas. Aquí, de manera parecida a la lista de formas canónicas, se encuentra ordenada por lemas o raíces, sin importar la parte de la oración de la que se trate la palabra. Por ejemplo, ante el lema nación se pueden agrupar naciones, nacionales, nacionalizar, nacionalmente y nacionalización.

Listas de dos o más palabras. En estas listas no se presentan palabras simples, sino pares, tríos, etc. de palabras que ocurren contiguamente. De igual manera, cada una se acompaña de su frecuencia. Su importancia radica en la localización de unidades multiléxicas.

Lista de partes de la oración. Si bien en las anteriores listas, excepto en la de dos o más palabras, se puede indicar la categoría gramatical a la que pertenece cada unidad, se puede tener una lista que nada más traiga las distintas partes de la oración, con el fin de tener una referencia de las categorías gramaticales para fines estadísticos.

Orden de la lista de palabras

Orden alfabético. La lista de palabras ordenadas de manera alfabética, normalmente ascendente, resulta útil, en primera instancia, para localizar más fácilmente las palabras de la lista. Además, se puede visualizar las palabras que empiezan con la misma raíz y, de esa manera, agrupar palabras que contengan la misma raíz.

Orden por frecuencias. En este apartado, las palabras de la lista se encuentran en orden decreciente, generalmente, de acuerdo con la frecuencia absoluta de cada una las palabras. Esta presentación permite conocer las palabras más frecuentes de un texto y comparar las palabras con otras listas. En corpus de referencia, equilibrados y balanceados, se puede observar las palabras más frecuentes. Su aplicación es muy diversa; por ejemplo, en el terreno de la lexicografía (Diccionario Básico del Español de México y Longman), la enseñanza de idiomas, el desarrollo de prototipos (traducción automática, sistemas de búsqueda de información, etc.).

A menos que se ocupe una stoplist como filtro, las palabras más frecuentes serán palabras funcionales. Con el empleo de una stoplist, aplicado a un corpus técnico, se puede conocer la terminología mediante la comparación de las palabras de contenido de ese corpus contra las de un corpus de otro tema ajeno.

Orden alfabético inverso. Este orden alfabético a partir de la primera letra del lado derecho de la palabra puede servir para el análisis de las flexiones de una lengua o como un diccionario de rimas.

Orden de aparición. Las palabras se ordenan según van apareciendo en el texto. Esto es diferente al texto mismo, pues solo se muestra una sola vez cada palabra, acompañada con el valor de su frecuencia. Sirve para conocer la distribución y organización de las palabras en el texto. Por ejemplo, las palabras que ocurren con frecuencia similar, pero que aparecen por primera vez en diferentes lugares del texto, indican un cambio del tema o del tipo de vocabulario empleado.

Orden por longitud de las palabras. Las palabras van ordenadas de manera ascendente según el número de caracteres de cada palabra, de manera que en primer lugar aparecen las palabras de una letra, después de las de dos letras y así consecutivamente hasta llegar a la más larga. Cada grupo se ordenará además en orden alfabético.

Orden por categoría gramatical. Aquí se presenta la lista de palabras ordenadas según las diferentes partes de la oración. Para cada categoría se pueden presentar las palabras ordenadas alfabéticamente o por frecuencias. Las listas de palabras, cualquiera que sea su tipo o su orden, no nos permitirá identificar los diferentes sentidos de las palabras. Para hacer este trabajo, normalmente recurrimos a las concordancias, que es nuestro siguiente objetivo.

Problemas con las listas de palabras

Segmentación. Un punto clave en el conteo de palabras es la segmentación o identificación de los tokens. Sin embargo, no todo lo que aparece entre espacios vacíos puede ser considerado como tal. Los guiones (cortos o largos) son normalmente eliminados del conteo de palabras, por lo que las palabras separadas con guiones a fin de línea son consideradas dos palabras erróneamente.

Signos de puntuación. En general, los signos de puntuación son eliminados. Por ello, en las listas llegan a aparecer letras sueltas o sin ningún sentido que pertenecen a otras palabras o que son partes de contracciones (por ejemplo, don y t en don't). Al perderse los signos de puntuación, se llega a confundirse el significados de las palabras (por ejemplo, la letra s de Gerardo's, s.s.s., y m/s). Asimismo,

los números grandes separados por comas y puntos se convertirán en dos o más palabras a contarse.

Eliminación de cifras. Es común que las cifras sean eliminadas del conteo de palabras, pues al menos no se consideran palabras como tales; no aparecen en la lista, pues no proporcionan información relevante. De esta forma, se eliminan no solo los números que indican los capítulos y secciones, sino otras cifras que pueden ser significativas, como las fechas. Por otro lado, hay que considerar que no acostumbran eliminarse los números cuando aparecen escritos, lo cual es contradictorio. En tal caso, los números escritos de varias palabras acostumbran contarse cada una por separado, por ejemplo, cuarenta y tres mil setecientos sesenta y tres.

Eliminación de mayúsculas. En la mayoría de la lista de palabras se acostumbra convertir todas las letras a minúsculas, a fin de no contar como diferentes types a dos palabras iguales, pero de las cuales una es inicial de párrafo o después de punto, y la otra no. Sin embargo, hay que tener ciertas consideraciones:

- Muchas siglas que normalmente van con todas sus letras en mayúsculas, al ser cambiadas llegan a ser confundidas o desconocidas.
- Hay palabras que tienen diferente significado cuando van con inicial mayúscula, como el caso de Papa y papa.

Soluciones a los problemas con las listas de palabras

Entre las etapas de digitalización del texto y de etiquetado se tiene una etapa de preprocesamiento. En esta etapa se pueden diferenciar las partes del texto que quieren hacerse distinguir de las demás, como son los nombres propios, números, fechas, siglas, abreviaturas, etc., mediante etiquetas especiales, las cuales podrán ser tratadas de manera especial por los programas de conteo y lista de palabras. De igual forma se pueden transformar las palabras, en particular los acortamientos, en las unidades significativas que las componen (como el caso de don't). Para ello, hay que marcar en el texto tanto la palabra original como las transformadas, e indicar al programa si se consideran las primeras o las segundas. Se pueden utilizar expresiones regulares para que se haga una especie de preproceso automático al corpus antes de correr el programa. Se pueden seleccionar y correr los programas que, como subrutinas, realicen los cambios requeridos.

Actualmente existen varios programas y herramientas que permiten hacer conteo de palabras. Sin embargo, para los casos en que se quiere diseñar una herramienta para hacer conteos y listas de palabras de diferentes corpus, entonces se puede pensar en diseñar un programa versátil, adaptable a diferentes circunstancias. Así, el programa

podrá contar o no los números, y en el primer caso considerará las cifras como una sola unidad (275,376.25); podrá conservar las mayúsculas en nombres propios y siglas, etc. El programa puede ofrecer la serie de opciones que puede realizar, para que el investigador decida las que mejor se ajustan a sus análisis.

TF-IDF

En el conteo de palabras cabe destacar una medida muy empleada, el método de TF-IDF (Term Frequency - Inverse Document Frequency). Dicho método genera listas de palabras clave asignando a cada palabra del corpus un peso que indica su relevancia con respecto al documento seleccionado y al corpus general. TF-IDF es la unión de dos métricas: frecuencia del término (Term Frequency, TF) y frecuencia inversa de los documentos (Inverse Document Frequency, IDF). TF parte de la idea de que las construcciones que ocurren con frecuencia en un documento están fuertemente relacionadas con el contenido de los textos. Esta métrica asigna los pesos a las palabras con base en su frecuencia relativa en cada uno de los documentos a analizar. IDF realiza la asignación de pesos calculando el número de documentos del corpus en los que aparecen las palabras. Así, la medida TF-IDF se basa en la determinación de la frecuencia relativa de las palabras en un documento específico comparada con la proporción inversa de la frecuencia de dicha palabra en todo el corpus. La medida califica con pesos altos las palabras con alto valor de importancia para cada uno de los documentos analizados.

10.2. Concordancias

Las concordancias son también listas de palabras que aparecen en un corpus, pero en lugar de estar las palabras aisladas, lo cual ya tiene su importancia, estas se encuentran en el contexto. Por ello, una concordancia también se le conoce como Key Word In Context, que se abrevia como KWIC.

Para análisis literarios de obras cortas, se pueden traer las concordancias en forma de lista ordenada alfabéticamente, y por cada palabra se presenta el contexto en el que aparece. Entre algunos ejemplo de concordancias de obras literarias puede encontrarse la que hicieron Hans Flasche y Gerd Hofmann en 1980 para los Autos Sacramentales de Calderón, o el trabajo paralelo de Susana Arroyo en 1993 y de Juan López Chávez y Marina Arjona en 1994 sobre el Primero Sueño de Sor Juana Inés de la Cruz.

Hoy en día existen distintas herramientas informáticas que permiten obtener las concordancias que se requieran de un conjunto de textos propios o incluso de textos en Internet. En tales casos, el usuario introduce la palabra o el patrón de palabras a buscar,

y el programa arroja sus concordancias.

En algunos programas, las concordancias son un paso posterior a la lista de palabras, pues de ésta se selecciona la palabra a traer en concordancia.

Las concordancias pueden traer el contexto sin modificaciones, esto es, con mayúsculas y minúsculas y con los signos de puntuación (a menos que se quiera lo contrario).

Por facilidad de análisis, la palabra que se analiza viene en una columna centrada, y del lado izquierdo la parte del texto que le antecede, y del lado derecho el texto que le precede (ver figura 10.1).

En las concordancias sólo se trae el texto y no las etiquetas de marcaje. Por ello, no se distinguen en una concordancia la tipografía del texto (fines de párrafo o tipos de letra, por ejemplo), pero podría diseñarse un programa para que trajera las etiquetas y así identificar la tipografía.

El tamaño de la ventana

La ventana se define como la cantidad de texto que puede traerse acompañando a la palabra que se analiza en una concordancia. Una ventana puede ser de tamaño variable, ajustándose a una oración o frase, como el caso de las concordancias de Los Autos Sacramentales de Calderón, o incluso a un párrafo.

Una ventana puede ser de tamaño más fijo y sujetarse a un número de palabras o caracteres tanto a la izquierda como a la derecha de la palabra a analizarse. Cuando se trata de número de caracteres, la ventana es de tamaño fijo, pero las palabras pueden cortarse. En caso de número de palabras, éstas no se cortan, pero el tamaño de la ventana es más variable.

Elementos de una concordancia

Una concordancia puede además traer datos sobre la fuente, tales como:

- El código identificador del documento que se trate.
- Datos bibliográficos de la fuente.
- Además de los datos de la fuente puede incluirse el tipo de fuente según su clasificación (como el CREA).

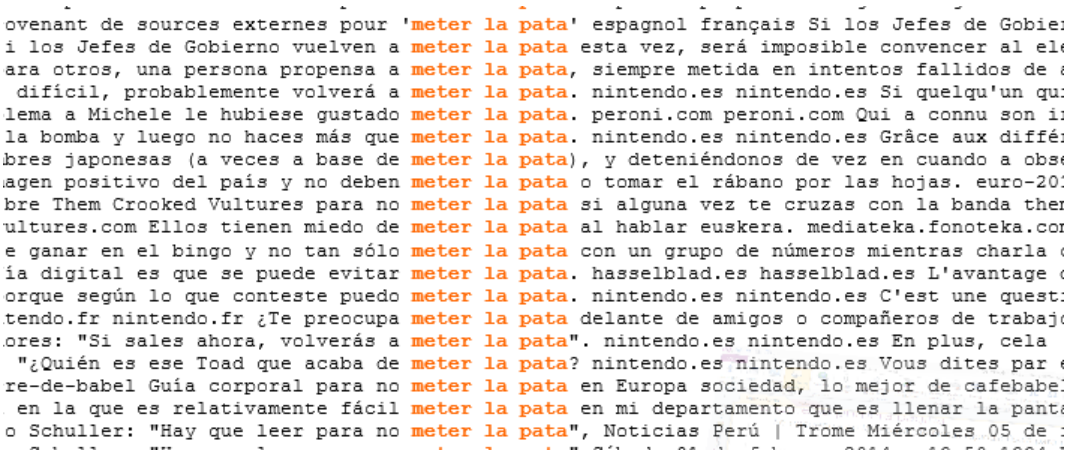


Figura 10.1: Resultado de concordancias para la búsqueda del patrón de palabras 'meter la pata'.

- La ubicación de la concordancia en el texto (página, número de línea, etc.).
- En el caso de un corpus de definiciones, la fuente puede ser la palabra a definir.

La base para traer una concordancia es una palabra, pero también:

- Un conjunto de palabras.
- Lemas.
- Todas las palabras que tengan una categoría gramatical.
- Combinaciones de las anteriores (BwanaNet del IULA).

Métodos para búsquedas complejas

- Operadores Booleanos AND, OR, NOT.
- Expresiones regulares.
- Comodines asterisco e interrogación.

Orden en que pueden presentarse las concordancias

- Por orden de aparición.
- Por orden alfabético de la palabra.
- Por orden alfabético de una palabra del contexto.

10.3. Colocaciones

En un texto, las palabras vienen acompañadas, y normalmente ocurren en conjuntos. Una colocación se define como la ocurrencia de dos o más palabras que se encuentran cercanas en un texto y que tienden a ocurrir cercanas en ciertos contextos. En este sentido, se considera una colocación como la combinación frecuente de palabras, así como la combinación en la que una palabra requiere la presencia de otra para expresar un sentido dado. Esto último ha correspondido a los diccionarios explicativos y combinatorios. Entre ellos, cabe destacar el Diccionario Combinatorio del Español Contemporáneo, publicado en 2004, y el Diccionario Combinatorio Práctico del Español Contemporáneo, publicado en 2006, ambos dirigidos por Ignacio Bosque.

Las colocaciones tienen varios usos dentro de la traducción, la lexicografía y la enseñanza de idiomas, así como para el desarrollo de sistemas de generación de documentos. Entre diferentes tipos de colocaciones se pueden encontrar:

Expresiones terminológicas: procesamiento de lenguaje natural, movimiento uniformemente acelerado, labio leporino, cinturón de seguridad.

Expresiones discursivas: por favor, cómo te va, sin embargo, vamos a ver.

Locuciones (nominales, adjetivales, verbales, etc.): blanco y negro, caer de bruces, punto de partida, de la misma manera, vino tinto, atención médica, tomar una decisión, hacer un favor.

Expresiones idiomáticas: tomar el pelo, meter la pata, jalarle las orejas, empinar la jarra.

Nombres de organismos: Sistema Nacional de Protección Civil, Consejo Nacional de Ciencia y Tecnología, Grupo de Ingeniería Lingüística.

Gracias a las colocaciones podemos observar que cuando ocurre la palabra hospital normalmente también se encuentran las palabras doctor y enfermera; que para el sustantivo dinero se asocian los verbos comprar, gastar y vender; que a la palabra miedo se le correlacionan los verbos tener y sentir, con mayor frecuencia, y experimentar y padecer, con menor frecuencia.

El análisis de colocaciones evalúa las coocurrencias (y su frecuencia) de dos o más palabras en un contexto dado.

La cercanía de las colocaciones puede ser desde una distancia cero (palabras contiguas) hasta una distancia relativamente pequeña (por ejemplo, 6 palabras de distancia una de otra).

El análisis de colocaciones consiste normalmente en la determinación de los patrones de su ocurrencia. Se puede determinar el patrón de ocurrencia actual y compararlo con el patrón de ocurrencia esperado.

El análisis de colocaciones parte de una evaluación de las concordancias.

Elementos de una colocación

Nodo. También llamada base o palabra clave/llave, es el elemento léxico del cual se está buscando la colocación. En el caso de la tabla mostrada a continuación, el nodo en este libro sería la palabra *corpus*.

Colocativo o correlato léxico. Es cualquier elemento léxico que coocurre con el nodo en un contexto. Para la tabla correspondiente a este libro, los correlatos léxicos considerados son únicamente sustantivos, pero puede considerarse cualquier palabra o bien podrías ser todos los verbos.

Ventana (span en inglés). Es el contexto en el cual ocurren los colocativos, que para el caso de la tabla 10.1 se tiene una ventana de cinco elementos a la izquierda y cinco elementos a la derecha.

Conteo de colocaciones

Lo más sencillo es contar las palabras que se encuentran cercanas a otra palabra dada. Se parte de la concordancia de la palabra llave a la que se desea determinar sus colocaciones. Se cuentan todas las palabras que ocurren (anotando su frecuencia) en una

ESPAÑOL	5	1	0	0	0	CORPUS	0	17	20	6	4
TEXTOS	4	6	3	12	0	CORPUS	0	11	4	4	8
ORALES	0	3	1	0	0	CORPUS	21	0	6	2	3
LINGÜÍISTICO	4	0	0	0	0	CORPUS	26	0	0	1	2
MÉXICO	3	4	5	0	0	CORPUS	0	2	0	1	17
TEXTUALES	0	1	0	3	0	CORPUS	23	0	0	2	2
PALABRAS	3	2	7	9	0	CORPUS	0	1	0	1	3
ANOTACIÓN	1	1	4	10	0	CORPUS	1	0	0	2	1
ETIQUETADO	2	3	0	4	0	CORPUS	5	3	1	1	1
LINGÜÍISTICA	2	0	1	14	0	CORPUS	0	0	0	0	2
CASO	2	1	5	9	1	CORPUS	0	0	0	0	1
ESTUDIO	1	2	0	2	0	CORPUS	0	13	0	0	0
CONSTRUCCIÓN	2	3	6	4	0	CORPUS	0	0	0	0	1
EJEMPLO	0	0	0	9	0	CORPUS	0	1	3	0	0
TEXTUAL	0	2	0	0	0	CORPUS	6	3	0	1	2
INGENIERÍA	0	2	4	0	0	CORPUS	0	0	5	2	1
CEMC	2	0	1	2	0	CORPUS	2	0	1	0	6
LINGÜÍISTICOS	1	1	2	0	0	CORPUS	9	0	0	0	0
REFERENCIA	0	0	0	1	0	CORPUS	0	10	1	1	0
MEXICANO	1	0	0	0	0	CORPUS	0	0	9	0	3
ELEMENTOS	1	0	2	6	0	CORPUS	0	0	0	2	1
TAMAÑO	1	1	2	5	0	CORPUS	0	0	1	2	0
CONTEXTOS	3	2	1	0	0	CORPUS	0	5	0	0	1
CONTEMPORÁNEO	1	1	0	0	0	CORPUS	0	1	0	9	0
DEFINITORIOS	1	3	2	1	0	CORPUS	0	0	5	0	0
DOCUMENTOS	0	2	1	2	0	CORPUS	0	0	1	4	2

Tabla 10.1: Ejemplo de los elementos de colocación.

posición definida hacia la derecha o hacia la izquierda. Por ejemplo, en la Tabla XXX, a la palabra CORPUS le precede 21 veces la palabra ORALES inmediatamente, pero ocurre 11 veces la palabra TEXTOS a dos posiciones a la derecha, conformando en esta última CORPUS DE TEXTOS.

Información Mutua

La información mutua es una medida de gran importancia en la teoría de la información, que consiste en la información aportada por una variable aleatoria sobre la otra. En el estudio de colocaciones, la información mutua mide la fuerza de asociación entre dos palabras. Es una medida estadística que determina la cantidad de información que la aparición de una palabra nos da sobre la aparición de otra. Para ello, calcula la probabilidad de que dos palabras aparezcan juntas, y la compara con la probabilidad de que dichas palabras aparezcan por separado.

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

De la fórmula anterior, la información mutua entre dos palabras x y y, se expresa como $MI(x,y)$ o como $I(x,y)$, y corresponde al logaritmo base dos del cociente de la probabilidad de que las dos palabras ocurran juntas entre la probabilidad de ocurrencia de cada una de las palabras en el corpus.

A mayor valor de MI se tendrá que existe una asociación fuerte, de forma que la probabilidad de que aparezcan juntas dos palabras deberá ser mucho mayor que la de que aparezcan de forma independiente. En caso de que los dos valores de frecuencia sean muy similares, la coocurrencia de las dos palabras no suele considerarse muy significativa.

El criterio de costos en colocaciones

Un método para extraer colocaciones de más de dos palabras a partir de corpus es el criterio de costos (cost criterion). Este considera que no se puede determinar una colocación únicamente por su frecuencia absoluta, sino que también debe considerar el número de palabras en la colocación. La fórmula para calcular el criterio de costos $K(a)$ es:

$$K(a) = (/a/ - 1)(f(a) - f(b))$$

donde a es la colocación candidata, $/a/$ es la longitud de la colocación a o su número de palabras, $f(a)$ es la frecuencia de ocurrencia de la colocación a, y $f(b)$ es la frecuencia de la colocación siguiente en cuanto al número de palabras. La de mayor costo será la colocación. Por ejemplo, si "a number" ocurre 102 veces en un corpus, "a number of" ocurre 51 veces, y "a number of times" ocurre 20 veces, entonces:

$$\begin{aligned}
K(\text{a number}) &= (/\text{a number}/ - 1)(f(\text{a number}) - f(\text{a number of}) \\
&= (2-1)(102-51) \\
&= 1 \times 51 \\
&= 51 \\
K(\text{a number of}) &= (/\text{a number of}/ - 1)(f(\text{a number of}) \\
&\quad -f(\text{a number of times})
\end{aligned}$$

$$
\begin{aligned}
&= (3-1)(51-20) \\
&= 2 \times 31 \\
&= 61
\end{aligned}
$$

Por tanto, “a number of” es la colocación.

10.4. Referencias

Lecturas sugeridas

Barrios, María A. (2008). “Diccionarios combinatorios del Español: diferencias y semejanzas entre ‘Redes y Práctico”’. Actas del II Congreso Internacional de Lexicografía Hispánica, pp. 197-203.

Charniak, Eugene (1996). Statistical Language Learning. Cambridge: The MIT Press.

McEnery, Tony y Andrew Wilson (1996). Corpus linguistics. Edinburgh: Edinburgh University Press.

Oakes, Michael P. (1998). Statistics for Corpus Linguistics. Edinburgh: Edinburgh University Press. (Véase sección 3.2.9 Collocations involving more than two words: the cost criterion).

Ramírez, Gaspar, James L. Fidelholtz, Héctor Jiménez y Grigori Sidorov (2006). “Elaboración de un diccionario de verbos del español a partir de una lexicografía sistemática”. Avances en la Ciencia de la computación, Actas de ENC–2006, San Luís Potosí, México, pp. 270–275.

Sinclair, John (1991). Corpus, concordance, collocation. Oxford: Oxford University Press.

Obras citadas

Arroyo, Susana (1993). El Primero Sueño de Sor Juana: estudio semántico y retórico. México: UNAM-ITESM, CEM.

López Chávez, Juan, Marina Arjona (1994). Lexicometría y fonometría del Primero Sueño de Sor Juana Inés de la Cruz. México: Universidad Nacional Autónoma de México.

Capítulo 11

Herramientas de análisis textual

11.1. WordSmith Tools

WordSmith Tools es un software comercial desarrollado por Mike Scott de la Universidad de Liverpool y distribuido por Oxford University Press. Se trata de un conjunto de herramientas destinadas para estudiar el comportamiento de palabras en el corpus.

El programa ofrece al usuario tres herramientas básicas: Wordlist, KeyWords y Concord. Cada herramienta a su vez dispone de diversos mecanismos para realizar el análisis lingüístico cuantitativo con base en técnicas estadísticas.

La herramienta WordList genera una lista que incluye todas las palabras del corpus ordenadas alfabéticamente o por su frecuencia de aparición. Además, ofrece estadísticas detalladas sobre la distribución de las unidades léxicas y permite comparar textos con base en estas características. La herramienta proporciona información sobre el tamaño de los textos (el número de types y de tokens), el type-token ratio, el type-token ratio estandarizado, el largo de las palabras y de las oraciones, el número de oraciones y de párrafos, etc. El type-token ratio se obtiene dividiendo el total de types por el total de tokens. Un valor elevado apunta a que los textos objeto de análisis presentan un alto grado de variación léxica. En cambio, un valor bajo indicaría que las palabras se repiten con frecuencia y que, por tanto, el vocabulario empleado en los textos es menos variado. El type-token ratio no es una medida fiable para comparar textos que difieren mucho en su extensión. Por ello, WordSmith Tools ofrece como medida alternativa el llamado type-token ratio estandarizado. Para calcularlo la herramienta divide el texto en fragmentos (el tamaño del fragmento establecido por defecto es de 1,000 palabras), aplica la fórmula a cada fragmento por separado neutralizando de esta manera el impacto del tamaño del corpus sobre los resultados.

La herramienta Concord ofrece información sobre el contexto de aparición de palabras o frases. Es una herramienta de concordancias muy completa que, además de las características usuales, permite identificar unidades fraseológicas y colocaciones, realizando un análisis de patrones léxicos. Así, podríamos decir que Concord sirve para dos propósitos. Por una parte, genera las concordancias, o listas que muestran el contexto lingüístico de cada aparición de la palabra seleccionada por el usuario. Por otro lado, la herramienta ofrece información sobre las agrupaciones frecuentes de palabras con base en varias medidas de asociación (información mutua, Z score, Log-likelihood). Este tipo de información permite detectar los patrones léxicos recurrentes, caracterizar el uso que se hace de una palabra o de una expresión particular en los textos analizados o caracterizar el significado de las palabras a través de los contextos de su aparición.

Finalmente, la herramienta KeyWords identifica las palabras clave y describe la forma en la que éstas se distribuyen en el corpus y las relaciones que existen entre ellas. La herramienta compara dos listas de palabras. Una de ellas se considera como el corpus de referencia que sirve de base para la comparación durante el análisis. La otra se crea a partir del texto que se quiere estudiar y se denomina el corpus de estudio. La comparación resulta en la generación de la lista de palabras clave, es decir, aquellas palabras cuyas frecuencias de aparición en el corpus de estudio son significativamente diferentes de sus frecuencias de aparición en el corpus de referencia. Para establecer la significación estadística de las diferencias se utilizan dos técnicas: el test de la χ^2 o de Pearson y la prueba de verosimilitud log-likelihood.

Asimismo, el programa cuenta con varios instrumentos que complementan las herramientas básicas. Entre ellos se encuentran:

Text Converter y Data Converter. Estos instrumentos permiten modificar el formato de los textos a analizar y de los datos generados por el programa.

Splitter. Este instrumento segmenta los textos en fragmentos. Para hacer uso del mismo se debe indicar al programa el símbolo que representará el fin de los segmentos.

Viewer y Aligner. Esta aplicación tiene varias funciones. En primer lugar, visualiza el texto marcando las palabras de interés indicadas por el usuario. En segundo lugar, puede alinear diferentes versiones del mismo texto a nivel del párrafo o de la oración. Este tipo de alineación sirve, por ejemplo, para comparar el original y la traducción de un texto.

Webgetter. Esta opción permite construir un corpus a partir de los textos extraídos del Internet. El programa descarga los textos del Internet con base en el término y el motor de búsqueda indicados por el usuario.

Minimal Pairs. Este instrumento identifica los pares de palabras que presentan una diferencia mínima entre ellas. Dicha opción sirve, por ejemplo, para detectar posibles erratas en los textos objeto de análisis.

Language Chooser. Por medio de esta opción, el usuario puede especificar el idioma de los textos a procesar. Asimismo, es posible indicar los símbolos que el programa considerará como parte de la palabra. Esta opción puede ser útil, por ejemplo, para el tratamiento de los apóstrofes en inglés o para especificar si el programa debe considerar dos unidades separadas por un guión como una o como dos palabras.

Con todas las herramientas que ofrece, WordSmith Tools es un software sumamente útil y flexible, que puede ser de gran ayuda para estudiar el uso del lenguaje en contexto. Para obtener información más detallada sobre el funcionamiento del programa puede consultarse la página http://www.lexically.net/wordsmith/.

11.2. T-LAB

T-LAB es un software comercial desarrollado por el psicólogo Franco Lancia, quien lleva muchos años trabajando en el análisis textual mediante técnicas computacionales. T-LAB incluye un conjunto de herramientas lingüísticas y estadísticas para el análisis de contenido, el análisis del discurso y la minería de textos. La función fundamental de este programa es extraer, comparar y representar gráficamente las características lingüísticas de los textos. Las herramientas de T-LAB proporcionan información sobre las relaciones significativas entre diversas unidades lingüísticas en forma de tablas, gráficos de barras, gráficos radiales, etc.

El programa es flexible y puede ser adaptado a las necesidades del análisis. El software tiene versiones para inglés, español, francés, italiano y portugués. Cada una de éstas cuenta con un diccionario, con base en el cual el programa realiza las tareas requeridas. Los textos a analizar pueden ser de varios tipos: artículos de periódicos, transcripciones de entrevistas y discursos, respuestas a preguntas abiertas, documentos empresariales, textos legislativos, libros, etc.

La interfaz del programa es fácil de utilizar. La primera etapa del trabajo con el programa es el preprocesamiento del corpus. En esta fase se realizan las tareas de normalización, segmentación, lematización, identificación de las palabras gramaticales y de las palabras de contenido y selección de las palabras clave. La interfaz proporciona al usuario la posibilidad de revisar y corregir los resultados del preprocesamiento. Al terminar esta etapa, el corpus es importado a T-LAB. A continuación se pueden utilizar las herramientas que permiten tres tipos de análisis: el análisis de coocurrencias, el

análisis temático y el análisis comparativo.

En términos generales, el primer tipo de análisis da luz sobre el contexto de aparición de las expresiones lingüísticas. Específicamente, el análisis de coocurrencias permite detectar los contextos elementales de cada expresión, identificar todas las palabras con las que coocurre una expresión y estudiar la forma en la que se distribuye en el contexto la asociación entre dos o tres palabras.

El segundo tipo de análisis caracteriza los contenidos del corpus en términos de los temas principales. El programa ofrece tres opciones básicas para realizar el análisis temático: extracción de listas de contextos significativos que resumen el contenido de los textos, exploración de las relaciones entre los temas principales tratados en el corpus y agrupamiento de documentos con base en el criterio temático.

Finalmente, el análisis comparativo identifica similitudes y diferencias entre los textos del corpus. Para ello, el programa cuenta con tres herramientas. El análisis de especificidades permite identificar las expresiones típicas o exclusivas de un subconjunto de textos en comparación con la totalidad del corpus. El análisis de correspondencias tiene por objetivo determinar qué tan cercanos o lejanos están temáticamente los textos a comparar y cuáles son las unidades lingüísticas que afectan el grado de similitud. Así, el programa arroja datos cuantitativos de distintos aspectos de un corpus, al tiempo que permite interpretaciones cualitativas de ellos a partir de la observación rigurosa de los datos arrojados. Finalmente, el Cluster analysis permite identificar los grupos de objetos que tengan características comunes.

Para lograr los resultados, el programa utiliza diversas técnicas estadísticas, tales como índices de asociación, test de la χ^2, agrupamiento semántico, análisis multidimensional, cadenas de Markov, etc.

T-LAB es un software costoso, pero muy útil para el análisis de grandes cantidades de datos lingüísticos. En la página web http://www.tlab.it/es/presentation.php se puede encontrar información más detallada sobre el funcionamiento del programa, así como las condiciones para la obtención del mismo.

11.3. Goldvarb

Goldvarb es un software desarrollado en el Departamento de Lenguas y Ciencias Lingüísticas de la Universidad de York. El programa es gratuito y está disponible en Internet.

Goldvarb fue diseñado para el estudio de la variación lingüística mediante las técnicas de análisis estadístico multivariable, ya que se ha mostrado que la variación lingüística no es producto del azar y está condicionada por múltiples factores. El análisis multivariable permite identificarlos y cuantificar su impacto en la aparición de las variantes lingüísticas objeto de estudio.

Existen diversos programas que aplican el análisis de reglas variables. El nombre genérico con que se conocen tales programas es VARBRUL (Variable rule). Ha habido numerosas versiones optimizadas del primer programa de este tipo desarrollado por David Sankoff y Pascale Rousseau, entre las cuales se encuentra GoldVarb.

En términos generales, el objetivo del análisis que realiza Goldvarb es averiguar hasta qué punto la variación viene determinada por factores lingüísticos (contextuales y funcionales) o factores extralingüísticos (sociales y situacionales). Más precisamente, el análisis estadístico multivariable da luz sobre la relación que existe entre más de dos variables y calcula las probabilidades de que se manifiesten diversas variantes en determinadas condiciones lingüísticas o situacionales.

Normalmente, para realizar este tipo de análisis se necesita que el fenómeno lingüístico analizado presente variación; que el uso de una u otra variante no implique un cambio semántico o pragmático; y que la variación esté relacionada con las condiciones lingüísticas (contexto fonético, contexto sintáctico, función, etc.) y extra-lingüísticas (estatus social del hablante, tipo de contexto situacional, tipo de interlocutor, etc.) en que se produce.

El objeto de estudio se considera como variable dependiente (grupo de factores dependientes), mientras que las condiciones lingüísticas y socio-situacionales que se toman en consideración para estudiar el objeto se denominan variables independientes o explicativas (grupos de factores independientes o explicativos).

El modelo estadístico en el que se basa Goldvarb se denomina modelo logístico de regresión. Con base en las frecuencias de aparición de unidades de análisis en un conjunto de datos lingüísticos, este modelo estima la probabilidad de que un fenómeno sujeto a variación se manifieste en una de sus formas dada la coocurrencia de determinadas circunstancias. Las técnicas estadísticas sirven para determinar el grado de verosimilitud de las probabilidades calculadas e identificar las circunstancias que, al darse simultáneamente, explican el hecho lingüístico analizado.

Goldvarb realiza esta tarea de la siguiente manera. El programa recibe un conjunto de fichas en las que se registran los factores que se toman en consideración en el análisis. En la primera columna de cada ficha se registran las variables objeto de estudio

(variables dependientes) y, en las siguientes, las variables independientes cuyo impacto en el comportamiento de las variables dependientes se desea estimar. Al organizar los datos de esta manera, se crea un archivo de condiciones en el que se indican las relaciones de interdependencia que se quieren analizar.

A continuación, la herramienta permite realizar dos tipos de análisis: el análisis binomial de un solo nivel y el análisis de ascenso y descenso. El primero arroja información sobre el peso probabilístico de cada factor analizado por separado, mientras que el segundo correlaciona todos los grupos de factores independientes y determina cuáles de ellos inciden en el comportamiento de las variables dependientes.

Existen numerosos programas estadísticos capaces de llevar a cabo análisis similares a los que realiza GoldVarb, pero este último está especialmente preparado para trabajar sobre los datos de la variación lingüística y, además, presenta los resultados de una forma adecuada a los intereses de los lingüistas.

Goldvarb ha sido ampliamente utilizado en los estudios de corte sociolingüístico, pero también pueden aplicarse en otros contextos, cuando el uso de los recursos lingüísticos por parte del hablante está condicionado por diversos factores cuya influencia en la selección lingüística de los hablantes se quiere estudiar.

Dada la complejidad de los procedimientos estadísticos empleados por el programa, deben tenerse buenos conocimientos de estadística para poder leer adecuadamente los resultados cuantitativos que genera.

11.4. Referencias

Lecturas sugeridas

McTait, Kevin (1998). “A Survey of Corpus Analysis Tools”. (Véase en www.corpus.unam.mx/cursocorpus/Survey.pdf)

Programas de análisis textual

- WordSmith Tools: www.lexically.net/wordsmith
- T-LAB: www.tlab.it/default.php
- Goldvarb: www.individual.utoronto.ca/tagliamonte/goldvarb.html

Parte V

Aplicaciones

Capítulo 12

Aplicaciones en lingüística

El empleo de los corpus para investigaciones lingüísticas abarca los diferentes niveles, desde la fonética y fonología hasta la pragmática. A continuación se describen las aplicaciones que se han realizado en el seno del Grupo de Ingeniería Lingüística de la UNAM.

12.1. Fonología

En el ámbito de la fonética y la fonología se ha hecho poco dentro del Grupo de Ingeniería Lingüística de la UNAM. Sin embargo, en épocas recientes se ha incrementado el interés en esta área y existen varios trabajos en desarrollo. Uno de los trabajos terminados en esta área se hizo dentro de las investigaciones del proyecto del Corpus Histórico del Español en México (CHEM).

Reglas de correspondencia entre sonido y grafía en el español hablado en México en el siglo XVI

Con miras a crear un transcriptor automático de los textos que comprenden el Corpus Histórico del Español en México (CHEM), la tesis de licenciatura de Teresita Reyes Careaga para la obtención del título en Lengua y Literaturas Hispánicas, de la UNAM, busca las reglas de equivalencia entre el sonido y la grafía del español de México en el siglo XVI, a partir del análisis de los documentos en el CHEM. El estudio se sustenta en las bases de la fonética histórica y propone que la escritura refleja, con ciertas consideraciones, la pronunciación de los hablantes. Por ello, una parte del estudio se enfoca en conocer y determinar el nivel cultural de los autores o amanuenses y su procedencia geográfica para estos tiempos del español de América. Se da por entendido que a mayor nivel cultural se tendrán escritos más cultos y conservadores, en tanto a menos nivel se tendrá más cercanía al habla real.

El objetivo central de la tesis fue establecer el sistema fonológico del español hablado en México en el siglo XVI para poder crear las reglas de correspondencia entre sonido y grafía que serían implementadas en un transcriptor automático.

Se plantearon también algunos objetivos específicos para el área de la lingüística:

- Observar el proceso de transición del sistema fonológico del castellano medieval al español moderno.
- Observar cuáles fonemas representaban problemas y vacilaciones gráficas por su evolución natural y establecer en qué etapa de esta evolución se encontraban.
- Observar bajo qué condiciones se dio el trasplante del español al continente americano y, en relación con este punto, observar si la mayor parte de la población española que llegó a América era de origen andaluz, como se ha discutido ampliamente en la tradición hispanista.
- Observar qué fenómenos del dialecto andaluz permearon el español americano y cómo se encontraron reflejados en los documentos del corpus de trabajo.

Los textos reunidos en el CHEM correspondientes al siglo XVI son parte de los *Documentos Lingüísticos de la Nueva España*, editados por Concepción Company, también las *Cartas de Diego de Ordaz*, editados por Juan M. Lope Blanch y los *Procesos Inquisitoriales contra indígenas que realizó Fray Juan de Zumárraga en Nueva España*, editados por María Buelna.

Se hizo una minuciosa observación de cada grafía, el contexto en el que se encontraba y la correspondencia oral de dicha grafía. Con esta observación se determinaron las reglas de correspondencia entre sonido y grafía. Los residuos de la investigación, bajo la premisa de que toda regla tiene excepciones, se ubicaron, para hacer más sencillo el trabajo del transcriptor automático, en una lista de excepciones llamada en computación lista de filtrado.

Como resultado de la investigación se obtuvo una lista de reglas correspondientes a cada grafía usada en español del siglo XVI (se incluye, por ejemplo, la ç; la j con valor vocálico, entre otras). Además se incluyó la lista de filtrado mencionada anteriormente, como una lista de palabras con la transcripción fonológica asociada a cada una de ellas (sobre todo con palabras originadas de lenguas indígenas). Finalmente también se incluyó una tabla de correspondencias entre los alfabetos fonéticos de la Asociación Fonética Internacional (AFI, o IPA por sus siglas en inglés), el de la Revista de Filología Española (RFE) y el alfabeto fonético computacional Mexbet (Cuétara, 2004).

Una de las aportaciones notables de este trabajo al CHEM, es que en la interfaz de búsqueda del corpus se puede realizar una búsqueda con ortografía actual y con ello se pueden encontrar todas las formas asociadas a esa palabra. Por ejemplo, si se busca la palabra <licenciado>, en los resultados arrojados se obtendrán formas como <liçenciado, liçençiado, ljcenciado> entre otras (dependiendo de cada palabra buscada, por supuesto). Esto contrasta con otros corpus como los de la RAE: CREA y CORDE, y el Corpus del Español de Mark Davies, en los que cada una de estas formas son palabras diferentes y si se quiere encontrar con una forma en específico (liçençiado, por ejemplo) se debe hacer una búsqueda exacta con la ortografía que se pretende encontrar.

12.2. Morfología

Como ya se ha mencionado en varias ocasiones, para poder utilizar un corpus lingüístico como fuente de datos empíricos sobre el uso de la lengua, es necesario desarrollar herramientas computacionales que faciliten su procesamiento. El tipo más común de anotación de corpus es el etiquetado de partes de la oración (etiquetado POST), que consiste en la asignación automática de categorías gramaticales y rasgos morfosintácticos a las palabras del texto. El etiquetado POST es imprescindible para la realización de las tareas complejas del procesamiento del lenguaje natural y permite acceder a métodos de codificación más sofisticados como el parsing sintáctico o la anotación semántica.

Segmentación morfológica en español

En su tesis doctoral de El Colegio de México, titulada “Investigación cuantitativa de afijos y clíticos del español de México: Glutinometría en el Corpus del Español Mexicano Contemporáneo”, Alfonso Medina explora diversos métodos de segmentación morfológica automática.

Como ya hemos visto antes, la anotación morfológica consiste en la identificación de rasgos gramaticales a partir de los morfemas de una palabra. La primera etapa de este tipo de etiquetado es la segmentación de palabras en morfemas. Las herramientas existentes en el ámbito ofrecen buenos resultados, por tanto la anotación morfológica suele considerarse una tarea resuelta.

No obstante, la mayoría de los estudios presuponen la existencia y la naturaleza de las unidades morfológicas y descuidan el aspecto empírico de la investigación. Así, los métodos más utilizados de etiquetado morfológico se basan en reglas lingüísticas provenientes de gramáticas tradicionales. Al usar las gramáticas tradicionales como única fuente de conocimiento se desatiende una parte importante de hechos reales.

La tesis realizada por Alfonso Medina es un estudio empírico que descubre la morfología del español a través del procesamiento cuantitativo de corpus lingüístico con una mínima intervención del analista, sin agregarle manualmente ningún tipo de información adicional.

El trabajo tenía los siguientes objetivos:

- Desarrollar un procedimiento para la obtención automática de morfemas con base en un corpus lingüístico.
- A partir del catálogo de morfemas obtenido, desarrollar una herramienta que segmente palabras separando raíces y afijos.
- Establecer los criterios necesarios para distinguir entre palabras de contenido y palabras funcionales.

La investigación se centró en las unidades afijales (fragmentos de palabra morfológicamente pertinentes que ocurren adheridos a la base, ya sea al principio (prefijos) o al final (sufijos) de las palabras) y en los clíticos (partículas que no constituyen palabras plenas y tienden a acompañar subordinadamente a otras palabras).

Para fines de la investigación se utilizó el corpus CEMC. Para medir la cualidad de afijo o clítico de los segmentos del corpus se partió de los siguientes supuestos. En primer lugar, las palabras funcionales en general contienen menos información que las palabras léxicas. De la misma manera, los afijos aportan menor cantidad de información que las bases a las que se adhieren. En segundo lugar, en comparación con las bases, los afijos son considerablemente más frecuentes. En tercer lugar, el número de combinaciones en las que participan los afijos es mucho mayor, de manera que poseen una capacidad combinatoria más alta. Finalmente, el hecho de que el número de afijos sea limitado implica una relación de economía entre los elementos del nivel morfológico y los elementos del nivel léxico. Así, los afijos de una lengua se caracterizan cuantitativamente por su número limitado, su alta frecuencia, sus numerosos contextos de aparición, su baja entropía (menor cantidad de información) y su alta participación en varias palabras con numerosos segmentos de baja frecuencia.

Para cuantificar estas características y así poder identificar distintas unidades morfológicas se exploraron los siguientes métodos: estadística de diagramas (test de la χ^2, información mutua, razón de semejanza, coeficiente de Yule), entropía, índices de cuadros y de economía de Kock. A partir de dichos métodos se propuso un índice formal para medir la cualidad de afijo de cualquier segmento de palabra. Los mismos métodos se extendieron para medir el carácter de clítico de los segmentos que aun siendo gráficamente independientes no lo son en el plano lingüístico. Los procedimientos

desarrollados fueron formalizados y aplicados al CEMC, con lo cual se obtuvieron de manera automática los catálogos de afijos, clíticos y palabras funcionales. Dichos catálogos constituyen un tipo de descripción formal del español. La investigación de los métodos para cuantificar la afijalidad y la cliticidad de los segmentos del corpus permitió establecer la noción de glutinosidad cuantitativa entre cadenas de morfemas al interior de la palabra y del sintagma.

Como resultado de la investigación se construyeron diversos programas para los experimentos de descubrimiento y procesamiento de elementos gramaticales tanto al interior como al exterior de la palabra, entre ellos un tokenizador, un separador de raíces y afijos, un programa que ordena palabras funcionales según sus índices de entropía y economía y dos programas que miden el grado de glutinosidad de los segmentos del corpus (uno para el interior y otro para el exterior de la palabra gráfica). Dado su carácter empírico y cuantitativo, los métodos desarrollados son independientes de la lengua.

Identificación automática de categoría gramaticales en el español del siglo XVI

Los métodos tradicionales de etiquetado morfosintáctico no toman en cuenta la morfología de las palabras u ofrecen mecanismos simples para su tratamiento. Una posible explicación para esta situación es la predominancia de estudios enfocados al inglés, una lengua con morfología flexiva pobre. Aunado a ello, existe una falta importante de corpus electrónicos para las lenguas de uso regional, lo cual dificulta el desarrollo de aplicaciones computacionales para las mismas.

En su tesis de Maestría en Lingüística Hispánica de la UNAM, Carlos Méndez propone un método novedoso para la identificación automática de categorías gramaticales a partir del Corpus Histórico del Español de México (CHEM). El método complementa un modelo tradicional de etiquetado con la información sobre la estructura morfológica de las palabras descubierta mediante criterios lingüísticos cuantificables.

La investigación parte de la idea siguiente. Si en lugar de segmentos de longitud predefinida las reglas léxicas usaran unidades morfológicas reales, fundamentadas desde el punto de vista lingüístico, sería posible emplear el método de Brill en otras lenguas sin necesidad de variar el parámetro de longitud. Además, el uso de afijos disminuiría el número de reglas generadas innecesariamente, con lo cual el algoritmo se volvería más eficiente y, probablemente, aumentaría la precisión del etiquetado.

La segmentación morfológica es la descomposición de la palabra en unidades morfológicas. En la tesis se utiliza un método de segmentación y descubrimiento de morfología basado en técnicas cuantitativas, propuesto por Alfonso Medina. Dicho

método segmenta las palabras con base en el índice de afijalidad, que es una medida obtenida a partir de la cuantificación de las características de un afijo, tales como su capacidad combinatoria, la cantidad de información en relación a otras unidades y la economía que aporta al sistema lingüístico.

El objetivo general de la investigación era desarrollar un algoritmo para la identificación de categorías gramaticales basado en el método de Brill que tome en cuenta las unidades morfológicas descubiertas a partir de criterios lingüísticos cuantificables. Asimismo, se establecieron los siguientes objetivos específicos:

- Con base en el corpus obtener una lista de afijos mediante el método de Medina basado en el índice de afijalidad.
- Complementar el método de Brill con las reglas morfológicas basadas en afijos descubiertos de manera automática.
- Comparar el método que usa reglas morfológicas con el método que no lo hace.
- Construir un etiquetador de categorías gramaticales aplicable a un corpus del español de México del siglo XVI, que utilice reglas morfológicas basadas en afijos descubiertos de manera automática.

La investigación se realizó en las etapas siguientes. Primero, se seleccionaron los textos del Corpus Histórico del Español de México del siglo XVI. El tamaño total del corpus alcanzó 16,500 palabras aproximadamente. El corpus fue dividido en dos partes: un corpus de entrenamiento para generar reglas y un corpus de evaluación. Segundo, se determinó el conjunto de etiquetas de categorías gramaticales para la anotación. El estudio tomó como base el conjunto de etiquetas del estándar EAGLES. Tercero, el corpus de entrenamiento fue etiquetado de manera manual con la participación de estudiantes de la carrera de Lengua y Literatura Hispánicas de la facultad de Filosofía y Letras de la UNAM. A continuación, se utilizó el método de Medina para el descubrimiento de afijos en los textos del corpus. De esta manera se obtuvo un inventario de unidades morfológicas que se usaron en la generación de las reglas del etiquetado. Después, se modificaron las plantillas de reglas léxicas de Brill para integrar el inventario de unidades morfológicas descubiertas con el método de Medina. Las plantillas del método original toman en cuenta sólo cuatro caracteres al inicio y al final de la palabra, por tanto, el objetivo de esta etapa fue que el programa de Brill generara reglas usando los afijos y cadenas de afijos previamente descubiertos. Más adelante, se generaron las reglas léxicas con el método de Brill original a partir del corpus de estudio. Con ello, se obtuvo el primer corpus etiquetado y un conjunto de reglas. Después, a partir de los mismos textos, se generaron las reglas morfológicas basadas en los afijos descubiertos. Esto resultó en un segundo corpus etiquetado y un nuevo conjunto de reglas comparables con

las primeras. Finalmente, se comparó el desempeño de ambos métodos con la medida de precisión.

El estudio demuestra que la inclusión de afijos descubiertos a partir de criterios lingüísticos mejoró el método de Brill haciéndolo más económico. Es decir, se obtuvo la misma precisión, pero se redujo el número de reglas generadas por el método. Así, la aportación de la tesis es un conjunto de reglas de etiquetado morfosintáctico aplicables a un corpus del español del siglo XVI y un método más económico para generar reglas de etiquetado, que puede utilizarse para distintas lenguas.

El estudio ilustra la utilidad del corpus no solamente para la construcción de herramientas de lingüística computacional (etiquetador morfosintáctico), sino también para los estudios lingüísticos (descubrimiento de la morfología). Asimismo demuestra la utilidad de la información lingüística para las tareas del Procesamiento del Lenguaje Natural.

12.3. Sintaxis

Para ilustrar el uso de corpus en el ámbito de sintaxis se describen diversas propuestas desarrolladas en el GIL que analizan patrones sintácticos para la extracción automática de contextos definitorios. Ya se ha mencionado que los contextos definitorios son unidades del discurso que introducen el término y su definición. En el GIL se desarrolló el Extractor de COntextos DEfinitorios (ECODE), que obtiene los contextos definitorios a partir de un corpus textual y los clasifica de acuerdo con el tipo de definición. Asimismo, la herramienta es capaz de identificar las partes constitutivas de un contexto definitorio: el término y su definición. Para realizar la extracción el sistema utiliza una gramática de patrones verbales definitorios.

La descripción del ECODE se hará más adelante. Aquí se mencionan los trabajos de investigación en el ámbito de la sintaxis que se han realizado en el Grupo de Ingeniería Lingüística con el fin de optimizar el ECODE.

Análisis lingüístico de contextos definitorios en textos de especialidad

Con miras a establecer las bases necesarias para la extracción de contextos definitorios, se realizó un proyecto de investigación patrocinado por el Consejo Nacional de Ciencia y Tecnología y por la Dirección General de Asuntos del Personal Académico de la UNAM, en donde se exploró el problema de extracción conceptual, se delimitó el concepto de contexto definitorio y de sus elementos constitutivos y se sentaron las bases para el CORCODE. Un primer resultado en este sentido fue la tesis de licenciatura en Lengua y

Literaturas Hispánicas de Rodrigo Alarcón. Se utilizó el Corpus Lingüístico en Ingeniería, constituido entonces por 25 textos, entre tesis, informes a patrocinadores y artículos en las áreas de transporte, logística, sistemas expertos y estructuras bioclimáticas.

Se identificaron dos patrones básicos que se emplean para conectar al término con su definición o para resaltar visualmente su presencia dentro del mismo, a saber:

Patrones tipográficos. La tipografía es un recurso que sirve como ayuda visual para identificar fácilmente los elementos importantes y diferenciarlos del resto del texto común. En este sentido, los patrones tipográficos se utilizan ya sea para resaltar a los elementos constitutivos mínimos de los contextos definitorios o bien para conectar el término con la definición. En la mayoría de los casos el término tiende a ser resaltado, aunque ocurre que la definición también se encuentra señalizada con algún elemento tipográfico o con alguna tipografía específica. En su trabajo de tesis encontró que las tipografías textuales más recurrentes para resaltar los elementos constitutivos son: cursivas, negritas, subrayados, mayúsculas, encabezados, viñetas y paréntesis. En cuanto al uso de signos de puntuación para conectar el término con la definición encontró que los más usados son dos puntos, punto y guión, o punto y seguido.

> **Desastre.** *Perturbación de la actividad normal que ocasiona pérdidas o daños extensos o graves.*

Patrones sintácticos. Suelen utilizarse construcciones sintácticas para unir un término con su definición, así como para referir atributos y características conceptuales del término en cuestión. Se encontraron dos patrones sintácticos:

- Patrones verbales definitorios. Las construcciones verbales unen un término con su definición. Se identificó los verbos metalingüísticos, como definir, entender o denominar; y los verbos comunes de lengua general, como ser y considerar. Asimismo, los patrones pueden contener, además del verbo, otras partículas gramaticales, siendo de las más comunes el pronombre impersonal “se” en posición proclítica o enclítica en relación con el verbo definitorio, las preposiciones “a” o “por”, y el adverbio “como”.

  TÉRMINO se define como DEFINICIÓN
  DEFINICIÓN se conoce también como TÉRMINO

- Marcadores reformulativos. Este tipo de conectores o patrones sintácticos conforman un proceso de reformulación en el que se explica el significado de un término a partir de estructuras sintácticas no verbales que, en este caso, sirven para referirse a los términos como elementos del propio lenguaje. Entre

los marcadores encontrados se encuentran: por ejemplo, es decir, esto es, en otras palabras, dicho de otra manera.

Se pudo observar que, en textos especializados, además de la definición, ocurre otro tipo de información relevante para entender al término dentro del contexto en el cual aparece.

Patrones pragmáticos. Estos patrones son muy útiles, junto con los patrones verbales, para identificar un posible contexto definitorio dentro del texto cuando no existen patrones tipográficos, pues describen el uso de los términos y manifiestan explícitamente las condiciones de uso o de alcance de dicho término, como son el ámbito temático, la ubicación geográfica, las instituciones que utilizan el término, el nivel de especialidad, o la frecuencia de uso, entre otras características pragmáticas. Las estructuras más recurrentes están conformadas por adverbios y frases adverbiales (usualmente, de manera general), frases prepositivas (desde el punto de vista), palabras simples (definición, concepto, término), y estructuras formadas por nombres propios.

La funcionalidad al interior de contextos definitorios con definiciones analíticas: el patrón sintáctico para + infinitivo

Los elementos básicos de una definición son el género próximo y la diferencia específica. El género próximo indica la clase a la que corresponde una entidad o un evento. La diferencia específica describe los rasgos o elementos que distinguen a la entidad o al evento de otros haciendo referencia al género próximo. El género próximo se introduce mediante unidades nominales, mientras que la diferencia específica es introducida por oraciones subordinadas.

De acuerdo con la presencia o ausencia de estos elementos, en el ECODE se identifican cuatro tipos de definiciones: sinonímicas, funcionales, extensionales y analíticas. La definición sinonímica hace explícito el género próximo y establece una equivalencia conceptual con el término definido. La definición funcional hace explícita la diferencia específica en la que se describe la función del término. La definición extensional hace explícita la diferencia específica en la que se enumeran las partes constitutivas del término. La definición analítica introduce de manera explícita tanto el género próximo como la diferencia específica.

Sin embargo, en algunas ocasiones las definiciones que el sistema ECODE clasifica como analíticas también aportan información sobre la función o utilidad de los términos definidos. Octavio Sánchez realizó su tesis de licenciatura de la Carrera de Lengua y Literaturas Hispánicas de la UNAM con el fin de estudiar de qué manera se expresa

dicha información y proponer reglas para su localización automática. Específicamente, se enfocó en el comportamiento del patrón sintáctico para + infinitivo. Los objetivos del trabajo eran:

- Describir el papel que desempeña este patrón en la introducción de información funcional sobre el término en las definiciones analíticas.
- Formular reglas que permitan determinar los casos en los que el patrón introduce dicha información.

Para la construcción del corpus de estudio se utilizó el Corpus Técnico del IULA y la herramienta BwanaNet. Primero, se realizó la extracción de todos los párrafos que contuvieran el verbo ser en tercera persona del singular o del plural, seguido de cualquier determinante, seguido de una ventana de diez palabras como máximo, hasta encontrar la preposición “para”. Como resultado de este procedimiento se obtuvieron 1616 párrafos. A continuación, se realizó la búsqueda automática de los casos en los que la preposición “para” iba seguida de un verbo en infinitivo. Después de este filtrado se obtuvieron 695 candidatos a contextos definitorios que contenían el patrón objeto de estudio. Finalmente, se llevó a cabo una limpieza manual, la cual dio como resultado 290 contextos definitorios que constituyeron el corpus de estudio.

En la primera fase del análisis los contextos definitorios fueron clasificados manualmente en tres grupos: a) el patrón para + infinitivo introduce información sobre la funcionalidad del término; b) el patrón para + infinitivo no introduce tal información; c) es difícil reconocer si el patrón estudiado introduce información de funcionalidad del término.

Para analizar con más detalle los casos problemáticos se llevaron a cabo los siguientes tipos de análisis: determinación del género próximo, determinación de la transitividad del verbo en infinitivo, determinación de la categoría gramatical de la palabra a la izquierda del patrón, determinación del tipo de sintagma a la izquierda del patrón, determinación de la existencia del verbo definitorio dentro de la definición y, finalmente, construcción de árboles sintácticos de dependencia.

El análisis sintáctico de dependencias arrojó la información más relevante con respecto al papel que desempeñaba el patrón para + infinitivo, ya que ponía en evidencia la relación sintáctica entre el sintagma introducido por el patrón y el género próximo. Al realizar este análisis, se pudo determinar que más de tres cuartas partes de los contextos definitorios analizados presentan información de funcionalidad del término definido.

Con base en el análisis sintáctico de dependencias se propusieron las siguientes reglas para la localización de la información sobre la funcionalidad del término introducida mediante el patrón sintáctico estudiado. El patrón sintáctico para + infinitivo que forma

parte de un contexto definitorio analítico introduce la información de funcionalidad del término definido excepto cuando:

- El patrón se encuentra fuera del contexto definitorio.
- El patrón se encuentra alejado del género próximo por más de dos subordinaciones.
- Existe un adverbio de negación que modifica al patrón o un adjetivo cuya semántica niega la funcionalidad del patrón.

De esta manera, el estudio demostró que en un alto porcentaje de casos el patrón sintáctico para + infinitivo aporta información sobre la funcionalidad del término y que es posible determinar y formalizar mediante reglas lingüísticas las condiciones en las que el patrón estudiado no introduce dicha información.

Análisis lingüístico de definiciones analíticas para la búsqueda de reglas que permitan su delimitación automática

Una tarea importante para la extracción automática de los contextos definitorios consiste en determinar los límites de la unidad discursiva que contiene la definición. En la práctica ello conlleva ciertas dificultades debido a que los contextos definitorios varían mucho en cuanto a su extensión y su estructura.

ECODE emplea una forma sencilla pero efectiva para delimitar la extensión de las definiciones, que consiste en cortar la definición en el primer punto. Sin embargo, esta medida no es suficiente debido a que existen casos en los la definición termina antes del primer punto o en los que hay información relevante para la definición que se encuentra fuera de los límites de la oración. Ariadna Hernández realizó su tesis de licenciatura en Lengua y Literaturas Hispánicas con el fin de optimizar el procedimiento de delimitación de contextos definitorios en el sistema ECODE.

Su investigación tenía como objetivo general encontrar patrones sintácticos que permitieran delimitar automáticamente la extensión de una definición.

Los objetivos específicos del trabajo eran:

- Conformar un corpus de contextos definitorios que presenten problemas para la delimitación de las definiciones.
- Plantear una serie de patrones que permitan al ECODE determinar el final de una definición cuando ésta termina antes del primer punto.
- Evaluar la precisión de cada patrón mediante un índice.

Para diseñar el corpus de estudio se usaron los patrones verbales definitorios que se emplean con frecuencia en las definiciones analíticas y tres fuentes textuales: Google, Corpus Técnico del IULA y corpus de la herramienta para la construcción y consulta de corpus lingüísticos Sketch Engine. Para extraer los contextos definitorios de Google se usó la opción de búsqueda avanzada con la introducción de los patrones verbales definitorios. Para extraer los contextos definitorios del Corpus Técnico del IULA se empleó la herramienta BwanaNet. En Sketch Engine los contextos definitorios fueron extraídos mediante la introducción de los patrones verbales definitorios en la opción de búsqueda Concordance. Se seleccionaron solamente aquellos contextos definitorios que tuvieran problemas de delimitación. Este procedimiento dio como resultado la obtención de un total de 70 contextos definitorios que conformaron el corpus de estudio.

La identificación de los patrones para la delimitación de las definiciones se realizó en varias etapas. Primero, los contextos definitorios fueron clasificados según la fuente textual y al patrón verbal definitorio. Segundo, se identificaron y se delimitaron los elementos principales de la definición analítica: el género próximo y la diferencia específica. Tercero, se elaboró de manera manual un primer inventario de patrones léxicos y sintácticos que indicaban la finalización de las definiciones. Finalmente, los candidatos a patrones fueron evaluados de la siguiente manera: se dividió la cantidad de contextos donde el patrón cumplía con su función entre la cantidad total de contextos definitorios encontrados por dicho patrón. Así para cada patrón se obtuvo un índice que indicaba el grado de efectividad que el patrón aportaría al sistema ECODE.

Al realizar este análisis se determinaron dos tipos principales de candidatos a patrones para la delimitación de las definiciones: a) patrones que rompen completamente con la definición para introducir otro término o tema; b) patrones que solamente delimitan los elementos de la definición, pero continúan con la información relevante para el contexto definitorio (es decir, amplían la información definitoria del mismo término). De acuerdo con los resultados de la investigación, los patrones más eficientes del primer tipo son los constituidos por un marcador discursivo contra-argumentativo más un sintagma nominal, por ejemplo, sin embargo + sintagma nominal, en cambio + sintagma nominal y mientras que + sintagma nominal. Los patrones del segundo tipo son: o sea, es decir, tal como, por ejemplo, etc. Dichos marcadores no pueden ser considerados como patrones de delimitación propiamente dichos, ya que indican el final de la diferencia específica pero no el término de la definición. Sin embargo, son de gran utilidad porque aportan información enriquecedora sobre el término definido. Además, la información que introducen puede considerarse parte o no del contexto definitorio según los propósitos del sistema de extracción.

El sintagma nominal en la extracción de relaciones léxico-semánticas de contextos definitorios: el caso de la preposición "de"

En su tesis de licenciatura en Lengua y Literaturas Hispánicas de la UNAM, Irasema Cruz estudió diversos tipos de relaciones semánticas que presentan los sintagmas nominales modificados por sintagmas preposicionales introducidos con la preposición "de" en los contextos definitorios analíticos en textos especializados, con el fin de ampliar las bases lingüísticas para la detección y extracción automática de dichas relaciones semánticas.

Los contextos definitorios analíticos con frecuencia contienen en su definición un sintagma nominal con uno o más sintagmas preposicionales como modificadores, específicamente introducidos por la preposición de. Ésta es una de las preposiciones más ambiguas del español. No obstante, son pocas las investigaciones que se centran en las relaciones semánticas marcadas por dicha preposición.

Teniendo en cuenta esta problemática, se establecieron los siguientes objetivos:

- Realizar un análisis descriptivo del comportamiento de las relaciones semánticas que presentan los sintagmas nominales modificados por sintagmas preposicionales introducidos con la preposición "de" en los contextos definitorios analíticos.
- Identificar y clasificar las relaciones semánticas de los sintagmas preposicionales.
- Proponer una nueva clasificación para las relaciones semánticas que no han encontrado cabida en las clasificaciones tradicionales.

El corpus de estudio contenía 656 contextos definitorios analíticos extraídos de diversas fuentes (el corpus del sistema ECODE, el CORCODE y el portal de la Biblioteca Nacional de Medicina de EE.UU. MedLine). Para la conformación del corpus se tomaron en cuenta diversos patrones verbales definitorios: ser + artículo, se define, se considera, se refiere, etc. Cada contexto definitorio contenía un sintagma nominal modificado por uno o varios sintagmas preposicionales introducidos con la preposición "de".

Para organizar los datos lingüísticos se usaron los siguientes criterios:

- Recursividad de la preposición "de" en el sintagma nominal (es decir, el número de sintagmas preposicionales contenidos en un sintagma nominal).
- Relación semántica entre el sintagma nominal y el sintagma preposicional.

Para la clasificación de las relaciones semánticas se usó una taxonomía tradicional desarrollada por Chaffin cuyas categorías principales son inclusión (por ejemplo, hiponimia o meronimia), posesión y atribución (por ejemplo, componente-objeto,

miembro-colección, porción-masa, material-objeto, etc.).

En primer lugar, el estudio ilustra la alta productividad que posee la preposición “de” en cuanto al establecimiento de relaciones semánticas. Con frecuencia los sintagmas nominales analizados tenían una estructura compleja al contener varios sintagmas preposicionales recursivos, por lo que dentro de un mismo sintagma nominal se encontraban diversos tipos de relaciones semánticas.

En segundo lugar, los resultados del trabajo indican que la atribución es la relación semántica más frecuente en el corpus de estudio. Ello se debe a que en las definiciones analíticas la diferencia específica indica las características distintivas del objeto, que a nivel semántico en numerosas ocasiones corresponden a la relación de atributo.

Finalmente, una de las aportaciones importantes del estudio es la identificación de algunas relaciones semánticas que no se mencionan en las taxonomías tradicionales. A este respecto el trabajo propone una clasificación alternativa que incluye categorías como contenedor + de + contenido, enfermedad + de + tipo, del tipo, etc. Dicha propuesta está basada en el análisis de las estructuras recurrentes encontradas en un corpus específico, pero podría ayudar a identificar los patrones lingüísticos necesarios para la detección y extracción automática de relaciones semánticas en contextos definitorios analíticos de áreas especializadas.

12.4. Semántica

Los estudios semánticos basados en corpus lingüísticos son muy variados y con múltiples aplicaciones. Tanto en el ámbito lingüístico como en procesamiento de le lenguaje natural y en inteligencia artificial, el conocer el significado de las palabras y las asociaciones que ocurren entre ellas resulta de vital importancia. A continuación se presenta un estudio orientado al diseño de sistemas de recuperación de información realizado en el Grupo de Ingeniería Lingüística.

Estructuración semántico-pragmática de léxico

Con el fin de permitir la búsqueda en lenguaje natural en los sistemas de recuperación de información, es necesario contar con bases de datos léxicas que contengan un repertorio de información para cubrir las diferentes características que introducen los usuarios. El diccionario onomasiológico se considera un sistema de recuperación de información que permite introducir la descripción del concepto en lenguaje natural y, por tanto, también requiere de esta base de datos léxica en donde vengan asociadas las palabras por su

similitud semántica.

Con este fin, la tesis de maestría en Lingüística Hispánica de Antonio Reyes, intitulada Estructuración semántico-pragmática del léxico en dominios restringidos para sistemas de recuperación de información, persigue analizar, desde un plano semántico y bajo los criterios de la gramática cognoscitiva, la base de conocimiento léxico del diccionario onomasiológico, en particular los verbos que, sin ser sinónimos, se desempeñan como tales en contextos específicos.

El corpus de análisis lo constituyen 28 pares de verbos que al parecer de diccionarios especializados no guardan relación léxica de sinonimia, pero que con base en el proceso de determinación de grupos semánticos usado en el diccionario onomasiológico sí funcionan como sinónimos en cierto contexto (ver tabla 12.1).

registrar-representar	afectar-caracterizar	atravesar-experimentar	denotar-implicar
utilizar-hablar	considerar-importar	representar-designar	emplear-invertir
determinar-modificar	impresionar-afectar	entender-referir	continuar - preservar
recibir-realizar	describir-barrer	significar-expresar	existir-ocurrir
generalizar-usar	explicar-relacionar	señalar-expresar	propagar-viajar
desempeñar-expresar	obtener-liberar	especificar-expresar	suponer-considerar
marcar-distinguir	emplear-seguir	decir-predicar	reproducir-producir

Tabla 12.1: Corpus de pares de verbos sin relación léxica de sinonimia.

En un primer análisis se observó que por cada par, al intercambiar un verbo por el otro en las definiciones de las que provenían, resultaba que la codificación de la definición conservaba el mismo significado. Esto era razonable, pues los verbos y las definiciones estaban dentro de un contexto bastante delimitado y definido. Por esta razón, se analizó si el comportamiento similar de estos verbos podría extenderse a otros dominios comunicativos, con elementos léxicos distintos.

Para ello, se construyó un corpus de prueba, formado por fragmentos de textos de especialidad, no constituido por definiciones. Se utilizó el Corpus Técnico del IULA y las herramientas del BwanaNet para extraer 50 contextos científicos en los que son utilizados cada uno de los 56 verbos del estudio. En cada uno de estos contextos se revisó si el intercambio de verbos en contextos del corpus de prueba producía un cambio de significado, pudiendo darse tres casos: aceptable, si no causaba modificación; duda, en caso de duda o que el significado resultara ambiguo; inaceptable, si había cambios. Se encontró que las dos primeras tienen un valor alto de incidencia.

Con base en la gramática cognoscitiva y en los conceptos de Marcos Semánticos

de Fillmore, se argumentó semánticamente los factores que producen que el significado, independiente del verbo, no cambie. Con esto se concluyó que, en un discurso comunicativo, existen marcos que engloban estructuras más generales que adoptan significados particulares, dando sentido y cohesión a las construcciones en donde los verbos funcionan como pares semánticos.

12.5. Análisis del discurso

Para la producción del discurso se utilizan diferentes recursos, uno de ellos concierne al empleo de las anáforas. En Procesamiento de Lenguaje Natural, la resolución de la anáfora es uno de los principales problemas más difíciles de resolver.

Análisis de referencias

En el marco del proyecto Extracción de conceptos en textos de especialidad a través del reconocimiento de patrones lingüísticos y metalingüísticos, patrocinado por CONACYT al Grupo de Ingeniería Lingüística, se realizó un estudio para identificar las expresiones referenciales que se manifiestan en los contextos definitorios (CDs), esto es, los fragmentos textuales que introducen un término con su correspondiente definición. Por ello, el corpus de análisis fue el CORCODE.

Un tema interesante, pero complejo tratándose de CDs, es la forma en que las relaciones anafóricas intervienen para su extracción. La anáfora es comúnmente el término que se emplea para hacer referencia a algo que anteriormente ya fue mencionado, y considera cualquier expresión, palabra o frase que recupera algo previamente enunciado. El análisis de relaciones anafóricas juega un papel determinante para la obtención completa de un CD.

En el ejemplo siguiente:

> Este consta de un banco de capacitores sumergidos en aceite en un recipiente de porcelana y conectados en serie (...).

en efecto, vemos un pronombre demostrativo en representación del término del contexto. Si solo extrajéramos este CD incompleto sería imposible determinar a qué término corresponde la definición: “un banco de capacitores sumergidos en aceite en un recipiente de porcelana (...)”. Con base en lo anterior, es evidente la necesidad de una extensión de este tipo de casos.

Por extensión, entendemos el tamaño del fragmento textual que contiene el CD completo, con término y definición, mientras que por expansión se comprenderá la pertinencia de acudir al documento de origen del contexto con el objetivo de verificar la

extensión del CD.

Con la finalidad de resolver este problema, primero es necesaria la identificación de los tipos de relaciones anafóricas que operan con CDs. Con este fin, Valeria Benítez realizó un estudio profundo donde se describen de manera completa relaciones anafóricas presentes. En dicho estudio se encontró que, principalmente, son cuatro las expresiones más frecuentes en CDs. En el primer grupo se encuentran algunos pronombres demostrativos (esto, aquellos, esta), personales (lo, le), relativos (la cual, lo cual, que) e impersonales (el primero). La frecuencia de esta clase de expresiones no es muy alta, pero son las más comunes en la ocurrencia de candidatos incompletos, como puede verse en el siguiente ejemplo, ya que la expresión apunta a un antecedente omitido en la extracción automática.

> Esto es lo que se entiende por enfoque genético de la medicina o "medicina genética".

El segundo grupo abarca los sintagmas nominales con determinante demostrativo, los cuales son expresiones con valor anafórico porque refieren a una parte anterior en el texto. Por ejemplo:

> Estos elementos son parte constitutiva de los compuestos que forman la base material para la vida (. . .)

El tercer grupo lo conforman las expresiones mixtas (pronombres y sintagmas nominales con demostrativo) en las que se muestran cadenas de anáforas o anáforas muy cerca de otras, es decir, que las cadenas de referencia se manifiestan con pronombres y sintagmas nominales que se encuentran en una relación anafórica.

> Esta concepción es lo que se conoce con el nombre de materialismo histórico.

En el anterior ejemplo se observa cómo la expresión anafórica, representada por el pronombre lo hace referencia al sintagma nominal con demostrativo "esta concepción", que a su vez tiene como referente al verdadero término de la definición.

El último grupo está constituido por las expresiones ligadas a una entidad previamente enunciada, las cuales pueden ser sintagmas nominales, elipsis o marcadores discursivos.

> El primer grupo es típico de los buques rápidos y consiste en olas de gran periodo, que sufren poca dispersión al alejarse del barco (. . .)

Una vez llevada a cabo la observación del corpus y después de realizar la clasificación de los elementos más frecuentes en las relaciones anafóricas, Benítez realizó el diseño de etiquetas XML para la identificación de relaciones anafóricas siguiendo los patrones de formación de las etiquetas ya establecidas para el CORCODE.

12.6. Pragmática

La pragmática se interesa por el funcionamiento de la lengua en el proceso de comunicación, haciendo una distinción entre lo que se dice, la intención con que se dice, y el efecto de lo que se dice.

Corpus para el análisis del discurso del concepto de Ad hoc-cracia

El estudio realizado por Margarita Palacios tenía por objetivo analizar el concepto de democracia a través de su uso en el discurso de los legisladores del Congreso de la Unión de México.

Para los fines de la investigación se construyó el corpus de ad-hocracia (para la descripción de este corpus, véase el apartado 2.7). El corpus se compone de dos partes. La primera está conformada por los documentos curriculares de los senadores y diputados y la segunda, por las transcripciones de los debates sostenidos en el Congreso. La inclusión de los documentos curriculares permite relacionar el uso de la lengua con la información sobre los hablantes y el contexto de situación.

El etiquetado del corpus consistió en la identificación de los segmentos en los que aparece el lexema democr. Para ello se utilizó el software WordSmith Tools. La unidad de análisis fue definida como las intervenciones de los legisladores en las que se registró dicho lexema. Las unidades de análisis fueron etiquetadas con categorías gramaticales. Posteriormente, se llevó a cabo el análisis estadístico de los contextos de aparición del lexema. Además, se registraron los tipos de oración (activa o pasiva) en los que aparecía el término objeto de estudio, ya que la tematización y la topicalización de elementos son datos valiosos que manifiestan las intenciones discursivas de hablante y los efectos que pretende producir en el interlocutor.

Asimismo, las unidades de análisis fueron clasificadas según su relación con los cuatro valores descriptivos para un ideal democrático: regla de mayoría, la deliberación, el estado de derecho y la separación de poderes. Así, fue posible observar de qué manera se relaciona el concepto de democracia con dichos valores en el discurso de los legisladores.

La aportación principal de este estudio fue la construcción de un recurso que permite identificar las formas lingüísticas correspondientes al concepto de democracia y los campos semánticos con los que se complementa, describiendo de esta manera la precepción y el uso que hacen de dicha noción los legisladores en México.

12.7. Referencias

Tesis reseñadas

Alarcón, Rodrigo (2003). Análisis lingüístico de contextos definitorios en textos de especialidad. Tesis de licenciatura, UNAM.

Benítez, Valeria (2008). Anáforas en la expansión de contextos definitorios: una propuesta de etiquetado. Tesis de licenciatura, UNAM.

Cruz, Irasema (2011). El sintagma nominal en la extracción de relaciones léxico-semánticas de contextos definitorios: el caso de la preposición DE. Tesis de licenciatura, UNAM.

Hernández, Ariadna (2009). Análisis lingüístico de definiciones analíticas para la búsqueda de reglas que permitan su delimitación automática. Tesis de licenciatura, UNAM.

Medina, Alfonso (2003). Investigación cuantitativa de afijos y clíticos del español de México: Glutinometría en el Corpus del Español Mexicano Contemporáneo. Tesis de doctorado, El Colegio de México.

Méndez, Carlos F. (2009). Identificación automática de categorías gramaticales en español del siglo XVI. Tesis de maestría, UNAM.

Palacios, Margarita (2013). La palabra ”democracia.en el Congreso de la Unión: usos y sentidos (México 2008). Tesis de doctorado, UNAM.

Reyes, Antonio (2006). Estructuración semántico-pragmática del léxico en dominios restringidos para sistemas de recuperación de información. Tesis de maestría, UNAM.

Reyes-Careaga, Teresita (2008). Reglas de correspondencia entre sonido y grafía en el español hablado en México en el siglo XVI para la creación de un transcriptor automático: una aportación al Corpus Histórico del Español de México (CHEM). Tesis de licenciatura, UNAM.

Sánchez, Octavio (2009). La funcionalidad al interior de contextos definitorios con definiciones analíticas: El patrón sintáctico para + infinitivo. Tesis de licenciatura, UNAM.

Lecturas sugeridas

Botley, Simon y Tony McEnery (Eds.) (2000). Corpus-based and Computational Approaches to Discourse Anaphora. Amsterdam: John Benjamins.

Capítulo 13

Aplicaciones en lingüística aplicada

13.1. Lexicografía

De los pioneros en lexicografía basada en corpus ha sido reconocido el trabajo realizado por la Universidad de Birmingham y la editorial Collins, en 1985, para compilar el Collins Cobuild Dictionary of English Language a partir del Collins Cobuild Corpus, un corpus informatizado de lengua oral y escrita, representativo del inglés británico. Sin embargo, es menos conocido que en México, a fines de 1973, varios años antes del proyecto Cobuild, Luis Fernando Lara, en El Colegio de México, inició un proyecto para constituir el Diccionario del español de México, con base en el primer corpus informatizado en lengua española, el Corpus del Español Mexicano Contemporáneo (CEMC).

En otras áreas de la lexicografía computacional basada en corpus, en México se cuenta también con un grupo en el Instituto Politécnico Nacional que ha llevado a cabo varios proyectos, tales como el diccionario explicativo del español o el diccionario de contextos de palabras españolas.

El Grupo de Ingeniería Lingüística tiene la lexicografía computacional como una de sus líneas de investigación y desarrollo. De hecho, con el fin de tener un producto específico aplicado a dominios de especialidad, así como técnicas de punta y de avanzar en el desarrollo de diferentes líneas de investigación de ingeniería lingüística, el eje rector del GIL ha sido la construcción de diccionarios onomasiológicos. Con la metodología desarrollada en las distintas fases de un proyecto de esta naturaleza ha sido posible elaborar, sistemáticamente y en un tiempo razonable, diccionarios integrales que permitan tanto la búsqueda semasiológica como la onomasiológica, aplicados a diversas áreas de conocimiento. Además, se ha contribuido al desarrollo de distintas áreas, tales como lingüística aplicada (en particular, terminología y lexicografía), lingüística computacional, ciencias de la computación e informática, bibliotecología y ciencias de la información.

Diccionario onomasiológico

En tanto un diccionario semasiológico se pregunta por las significaciones, es decir, se inicia del nombre para buscar el sentido o los sentidos ligados a él, en el onomasiológico se pregunta por las designaciones, esto es, se inicia del sentido y se busca el nombre o nombres conectados a éste. En la tesis de doctorado presentada en 1999 en la entonces University of Manchester Institute of Science and Technology (UMIST), se trabajó en el desarrollo de una metodología para crear diccionarios onomasiológicos con tecnología de punta en lingüística computacional y recuperación de información, a partir de corpus lingüísticos. La utilidad de los diccionarios onomasiológicos radica en que los emisores puedan expresarse efectivamente en la escritura y en el discurso, de manera que cuenten con una herramienta para expresar una idea cuando no se recuerda el término adecuado. En particular, este diccionario tiene como objetivo de permitir al usuario introducir en una interfaz el concepto a ser buscado a través de las ideas que éste tenga, usando cualquier palabra en cualquier orden. Lo que el sistema hace, una vez introducido la descripción del concepto, es expandir la formulación inicial del usuario, de forma que se busca, en la base de datos indexada, a los términos que contienen no sólo las palabras clave introducidas por el usuario, sino también todas aquellas palabras clave que están relacionadas semánticamente. Por ejemplo, supóngase que un usuario quiere obtener el término barómetro, el cual tendría la siguiente definición:

Barómetro: instrumento para medir la presión atmosférica.

Si el usuario introduce: aparato con el que determinamos la presión de la atmósfera; el sistema de búsqueda del diccionario busca cada palabra clave en los conjuntos de palabras asociadas a los términos. Dos de estos grupos semánticos serían:

$$\{\text{instrumento, aparato, dispositivo,}\ldots\}$$
$$\{\text{medir, determinar, estimar, conocer, }\ldots\}$$

El punto de partida para la elaboración de este diccionario consiste en la recopilación de textos de especialidad y en su procesamiento para ser manejables por la computadora, esto es, en la obtención del corpus lingüístico. Para formar la base de conocimiento necesaria en el desarrollo del diccionario onomasiológico, una segunda fase se enfoca hacia la búsqueda y extracción tanto de la terminología como de las definiciones o descripciones dadas en dichos textos. La última fase consiste en la determinación de los grupos semánticos, que utiliza un método creado *ex profeso* que aprovecha las definiciones capturadas en la base de conocimiento.

13.2. Terminología

La extracción terminológica consiste en la obtención automática de posibles unidades terminológicas (candidatos a término) a partir de un corpus especializado y responde

a las necesidades de investigación en los ámbitos de lexicografía y terminología. Asimismo, contribuye al desarrollo de numerosas aplicaciones del Procesamiento del Lenguaje Natural, tales como la construcción de glosarios, vocabularios y diccionarios de especialidad, la indexación de textos, la traducción automática, etc.

Una línea de investigación que se ha desarrollado en el Grupo de Ingeniería Lingüística del Instituto de Ingeniería, UNAM, se orienta a la extracción automática de términos a partir de corpus lingüísticos. A lo largo de la historia del Grupo se han concluido cuatro tesis en este tema.

Hacia una obtención computarizada de términos (aplicación concreta al léxico de la física en el nivel bachillerato)

Este primer trabajo, realizado por Antonio Reyes Pérez para obtener el grado de Licenciado en Lengua y literatura hispánicas de la Facultad de Filosofía y Letras, UNAM, se enmarcó en un proyecto patrocinado por CONACYT al entonces Centro de Instrumentos, actualmente Centro de Ciencias Aplicadas y Desarrollo Tecnológico. El objetivo práctico del trabajo era aplicar y explotar la herramienta computacional, WordSmith Tools, para extraer los términos relativos a mecánica pertenecientes al área de la física.

El corpus de estudio, que fue denominado Hipertexto, lo proporcionaron los expertos en física del Centro de Instrumentos de la UNAM. La caracteristica de los textos es que fueran exclusivos para la enseñanza y el estudio de la física en el nivel bachillerato, sin que esto excluyera que deban ser textos especializados en el área, en este caso de mecánica. Se conformó de quince archivos en formato electrónico y con la extensión txt (texto), compuesto de 76,991 palabras (tokens).

Para procesar y analizar el Hipertexto se utilizó WordSmith Tools. Primero se aprovechó la herramienta de lista de palabras (wordlist) y una lista de paro (stoplist) compuesta por 250 palabras, entre las que se encuentran pronombres, artículos, preposiciones, conjunciones y los verbos ser y estar. Como resultado se obtuvo un total de 1674 palabras diferentes (types), las cuales no solo eran términos en mecánica sino las que acompañan al discurso, como verbos, adjetivos y otros sustantivos. Por esa razón, se corrió el programa de búsqueda de palabras clave, contrastando el Hipertexto con dos archivos de referencia, uno dedicado a Octavio Paz y otro sobre salud alimenticia.

Gracias a la herramienta de agrupamientos y de información mutua se analizaron las 218 palabras clave obtenidas, con el fin de comprobar que las palabras clave realmente formaban parte de la terminología y que muchas de ellas eran constituyentes de términos poliléxicos, esto es, constituidos por dos o más términos simples. Con ello, se obtuvo finalmente un total de 307 términos en el área de mecánica, de los cuales 148 son simples,

y 159 poliléxicos. Finalmente, se presentaron los términos obtenidos a la gente del Centro de Instrumentos para que evaluaran y avalaran la terminología derivada del trabajo. Para la evaluación de los expertos se siguió el método Delphi, consistente en este caso de un solo ciclo, en donde cada uno de los expertos evaluó la terminología en privacidad, con el fin de no haber influencias entre los expertos. Los resultados aprobados fueron de un 98 % del total de términos propuestos.

Extracción de la terminología básica de las sexualidades en México a partir de un corpus lingüístico

Con el fin de sentar las bases para el desarrollo de un proyecto interdisciplinario e interinstitucional entre el Grupo de Ingeniería Lingüística con El Colegio de México y AVE de México, orientado a crear un Diccionario Básico de las Sexualidades en México, se han conjuntado varios esfuerzos, principalmente apoyados por CONACYT y por la Dirección General de Asuntos del Personal Académico de la UNAM. Entre los primeros resultados, se creó el Corpus de las Sexualidades en México (CSMX), que como corpus representativo y piramidal consta de 160 archivos distribuidos en ocho subáreas de la sexualidad y éstas a su vez en cinco niveles que señalan una variación diastrática o por grupo social.

Por su parte, Jorge Lázaro Hernández realizó su tesis de licenciatura de la Carrera de Lengua y Literaturas Hispánicas en la Facultad de Filosofía y Letras, UNAM, dedicada a extraer la terminología básica de las sexualidades en México, conteniendo los mil términos más representativos obtenidos a partir del CSMX. De manera análoga al trabajo realizado en la extracción de los términos correspondientes a mecánica en el área física, se utilizó la misma metodología aprovechando el programa WordSmith Tools. En este caso, no sólo se identificaron los mil términos básicos, sino que además se ubicaron éstos por las distintas áreas.

Extracción automática de términos en contextos definitorios

En el marco del proyecto Extracción de conceptos en textos de especialidad a través del reconocimiento de patrones lingüísticos y metalingüísticos, Alberto Barrón Cedeño realizó su tesis de maestría en Ingeniería en Computación, orientada a establecer las bases y la implementación de un programa para la extracción automática de los términos que aparecen al interior de un contexto definitorio. Dicho programa sería una adaptación del algoritmo C-value/NC-value, originalmente diseñado para obtener candidatos a término multipalabra del área de biomedicina en inglés.

Como método híbrido, C-value/NC-value tiene una etapa lingüística y una estadística. La etapa lingüística permite detectar un conjunto de candidatos a término ordenados con base en una medida que toma en cuenta la frecuencia de aparición y la longitud de cada candidato. Para ello, aplica al texto etiquetado con partes de la oración un filtro lingüístico de los patrones característicos de los términos en español, así como una lista de paro de un poco más de 200 palabras que muestran una alta frecuencia de aparición en los corpus utilizados y que no se espera que constituyan términos de las áreas de especialidad tratadas. La etapa estadística consiste en el conteo de frecuencia y longitud de los sintagmas que resultan ser candidatos a término.

Se utilizaron dos corpus de prueba: el Corpus Lingüístico de Ingeniería y un Corpus de Informática en Español obtenido por el Observatoire Linguistique Sense-Texte de la Universidad de Montreal para la elaboración de un diccionario fundamental sobre informática e Internet.

Para la evaluación del programa se utilizaron las medidas precisión (precision) y cobertura (recall) utilizadas ampliamente para evaluar sistemas de recuperación de información. La precisión calcula qué proporción de candidatos a término obtenidos de manera automática son realmente términos, en tanto la cobertura determina el conjunto de términos relevantes que son recuperados con respecto al número total de términos que existen en un documento analizado. Se seleccionaron algunos textos del corpus al azar para evaluar dos parámetros: el tipo de filtro (abierto o cerrado) y el uso u omisión de la lista de paro. Con ello se observó que al utilizar un filtro abierto se beneficia a la cobertura con respecto a uno cerrado, pero se perjudica a la precisión. Además, que al filtrar a los candidatos por medio de la lista de paro se mejora la precisión. De esta manera, si bien se determinó que la precisión (0.265) es mucho menor que la cobertura (0.794), esto se debía a que se incluyeron candidatos a término de una sola palabra, lo que resultó una de las modificaciones que se hizo al algoritmo original, además que se mostró tener un buen desempeño sin necesitar recursos adicionales como los usados en el otros prototipos.

TF-IDF para la obtención automática de términos y su validación mediante Wikipedia

Este último trabajo fue desarrollado por Luis Adrián Cabrera Diego como su tesis de licenciatura, Facultad de Ingeniería, UNAM, y en el marco del proyecto El vocabulario básico científico en México: Una investigación de sus características, componentes y difusión, financiado por CONACYT a El Colegio de México y en el cual el Grupo de Ingeniería Lingüística fue partícipe activo.

Como se ha visto a lo largo de este apartado, existen numerosas herramientas de extracción de términos que se basan en diversos métodos de análisis automático.

Sin embargo, la mayoría de los extractores terminológicos actuales no hacen uso de recursos léxicos para la validación de los candidatos a término obtenidos.

Teniendo en cuenta esta problemática, el estudio se propuso como objetivos generales:

- Desarrollar un sistema de extracción terminológica basado en conocimiento estadístico;
- Proponer un mecanismo para su validación empleando un recurso de información léxica.

La investigación se basó en el Corpus de textos científicos en español de México (COCIEM). Se decidió utilizar este corpus puesto que presenta una gran variedad de materiales y, además, siendo de carácter educativo, debe contener una gran cantidad de unidades terminológicas.

Antes de realizar la extracción terminológica, se llevó a cabo el pre-procesamiento del corpus. En primer lugar, los documentos fueron revisados de manera manual para corregir algunos errores existentes en el corpus debido a que los libros que lo conforman fueron escaneados y procesados con un reconocedor óptico de caracteres. En segundo lugar, con ayuda de la herramienta de análisis lingüístico automático Freeling se llevó a cabo la lematización de los documentos. Finalmente, para poder identificar tanto los términos monoléxicos como los poliléxicos, se extrajeron del corpus todos los unigramas, bigramas y trigramas de palabras.

Con el fin de detectar los candidatos a término a partir de los n-gramas se empleó el método de TF-IDF (Term Frequency - Inverse Document Frequency). Inicialmente, el método TF-IDF fue desarrollado para la creación de listas de palabras clave en el contexto de búsqueda de información. No obstante, su uso se ha ido ampliando a áreas como la extracción automática de términos, ya que en los textos especializados las unidades terminológicas resultan ser también las palabras clave, es decir, las palabras más importantes del texto.

Una vez generadas las listas de candidatos a término, se desarrolló un algoritmo para su validación. El algoritmo está basado en Wikipedia como recurso de información léxico-terminológica. La ventaja de utilizar dicho recurso es su disponibilidad y su amplio alcance. La validación se realizó de la siguiente manera. Primero, se determinaron las categorías de Wikipedia que correspondían a las áreas de conocimiento tratadas en el corpus. Después, a partir de la estructura de páginas y categorías de Wikipedia, para cada candidato a término se calculó su coeficiente de dominio, que indica su grado de pertenencia al ámbito de especialidad. El coeficiente de dominio se calcula con

base en el número o longitud de caminos que relacionan la página de la enciclopedia correspondiente al candidato a término y la categoría que representa el dominio de especialidad en la estructura de Wikipedia.

Los resultados de la extracción y de la validación de candidatos a término fueron evaluados con las métricas de precisión y cobertura. Para ello, las listas de términos obtenidas con el método propuesto fueron comparadas con las listas de términos creadas de manera manual por expertos en los ámbitos de especialidad correspondientes. La evaluación demostró que el método ofrece resultados prometedores.

13.3. Lingüística forense

La lingüística forense es un campo de investigación interdisciplinario que constituye una interfaz entre la lingüística y el derecho. En este ámbito complejo se hace cada vez más presente la participación de Tecnologías del Lenguaje, ya que la interacción entre el conocimiento lingüístico y las técnicas estadísticas y computacionales pueden ser una aportación relevante a la resolución eficiente de cuestiones legales.

Entre las numerosas tareas de lingüística forense se encuentra la detección del plagio. En términos generales, se entiende por plagio la acción de copiar en lo sustancial una obra ajena dándola como propia.

Existen diversos tipos de plagio. El más frecuente se denomina copy-paste e implica una copia literal de fragmentos específicos de un documento sin mencionar la fuente. Este tipo de plagio es relativamente fácil de detectar con métodos computacionales, ya que se trata de comparar el original y la copia e identificar fragmentos idénticos. Una forma no tan evidente de realizar el plagio es por medio de la paráfrasis, la cual consiste en cambiar la sintaxis, utilizar palabras con significado igual o parecido, reordenar oraciones y, en general, expresar el contenido de un trabajo original con recursos lingüísticos diferentes. Cabe mencionar que la paráfrasis es válida, siempre y cuando se haga referencia a la fuente de la información.

Es importante entender que la tarea de la lingüística computacional en este sentido no es determinar si existe el plagio, sino indicar en qué medida hay similitud entre dos documentos y en qué medida dicha similitud es producto del azar. De esta manera, la medición de similitud solamente contribuye a reforzar o disminuir la sospecha sobre la existencia del plagio. En el Grupo de Ingeniería Lingüística se han realizado varios trabajos que buscaban resolver este problema con técnicas diferentes.

Análisis estilométrico para la detección de plagio

Alejandro Rosas realizó su tesis de licenciatura en Lengua y Literatura Hispánicas, en la que la tarea de detección de similitud textual se aborda desde la estilometría. La tesis forma parte del esfuerzo que se realizó en el GIL en relación con un caso específico de detección de plagio. La tarea consistía en procesar y comparar de manera automática dos textos y medir el grado de similitud entre los mismos. Los textos pertenecían a dos autores, uno de los cuales acusaba al otro de haber plagiado su trabajo. El acusado negaba la imputación arguyendo que las coincidencias entre los textos se debían a la cercanía de género y tipo textual.

En efecto, los documentos pertenecían al mismo género y abordaban la misma temática, por lo que hacían uso de un vocabulario muy similar. De esta manera, la detección del plagio en este caso no podía realizarse únicamente mediante técnicas basadas en n-gramas, dado que se tenía que clasificar y medir las diferencias y similitudes entre textos que compartían el léxico, el tema y el género.

El estudio constaba de las siguientes fases. En la primera etapa se procedió a digitalizar los documentos, con lo cual se obtuvo un corpus electrónico de dos textos: uno, presuntamente plagiario, de 10,873 palabras, y otro, supuestamente plagiado, de 27,176.

Después, se añadieron al corpus dos textos de referencia que pertenecían a cada uno de los autores. Dichos documentos servían para dos fines. En primer lugar, era necesario establecer un nivel de variación inter-autor (es decir, el nivel de variación entre los textos escritos por autores diferentes). En segundo lugar, el hecho de tener dos textos más de cada autor permitió determinar si eran consistentes en su propio estilo.

A continuación, se usaron varios métodos para la comparación de los textos en controversia, a saber: la comparación por cadenas textuales, el análisis de frecuencias de aparición de palabras funcionales y el análisis estilométrico y estilográfico. La contribución de la tesis de Alejandro Rosas se centra en este último tipo de análisis.

Se denomina estilometría o estilística estadística al ámbito de la Lingüística Aplicada que hace uso de técnicas estadísticas para la medición de las características estilísticas de los textos con el fin de su sistematización y clasificación. La ventaja de este método consiste, precisamente, en que no se limita a detectar los fragmentos copiados de manera literal, sino que permite medir la similitud en los casos en los que el plagio se ha realizado por medio de la paráfrasis.

Los objetivos de la tesis eran:

- Aplicar el método de estilometría para el análisis y clasificación del estilo, entendido éste como las selecciones lingüísticas independientes del contenido.
- Representar determinados aspectos del estilo mediante datos estadísticos que puedan ser comparados para establecer el grado de similitud entre textos en el contexto de detección de plagio.

Para el análisis estilístico se establecieron los parámetros siguientes:

Longitud de palabra: la frecuencia con la que aparecen palabras de determinada longitud puede ser una marca característica del estilo de un autor.

Longitud de oración: presenta menos variación entre escritos del mismo autor que entre escritos de diferentes autores, sobre todo si los textos pertenecen al mismo género.

Longitud de párrafo: esta medida está estrechamente relacionada con la anterior.

Frecuencia de aparición de signos de puntuación: se toman en cuenta los diferentes signos de puntuación.

Para el procesamiento de datos se usó el programa Signature, elaborado por Peter Millican de la Universidad de Oxford. Dicho programa genera diversas gráficas y tablas que representan la frecuencia de los marcadores de estilo contenidos en los textos a comparar. Asimismo, ofrece al usuario la posibilidad de medir la significación estadística de los resultados de comparación por medio del test de la χ^2.

Detección de similitud textual mediante criterios de discurso y semántica

En su tesis de licenciatura en Lengua y Literatura Hispánicas de la UNAM, Brenda Castro propuso otro método interesante para la medición automática de similitud textual en el contexto de detección de plagio mediante paráfrasis. Como ya se mencionó en el apartado anterior, los métodos de detección de similitud textual pueden agruparse en tres categorías: comparación de textos mediante n-gramas de palabras, búsqueda de copia literal con ayuda de motores de búsqueda en Internet y análisis estilístico. Con todo, estos métodos no siempre representan una ayuda suficiente, con lo cual surge la necesidad de desarrollar métodos de análisis más fino. La tesis de Brenda Castro propone un método novedoso que contempla los niveles léxico-semántico y discursivo.

Con respecto a dichos niveles de comparación se establecieron las hipótesis siguientes:

- Ha de ser posible establecer la similitud entre el contenido semántico de un texto original y un texto que lo parafrasee mediante un cálculo.
- Las estructuras discursivas de un texto original y un texto que lo parafrasee han de ser similares, aunque las palabras y las estructuras sintácticas empleadas sean diferentes.

Los objetivos de la investigación eran:

- Comparar las estructuras discursivas de un corpus de textos originales y de textos parafraseados, para observar si éstas son similares.
- Calcular la similitud semántica entre los textos originales y los textos parafraseados pertenecientes a un corpus, para observar si el resultado obtenido es útil para la detección de similitud textual.
- Comprobar si la comparación de las estructuras discursivas de textos originales y textos que los parafraseen, junto con el cálculo de su similitud semántica pueden resultar de utilidad para la detección de similitud textual.

El estudio se llevó a cabo en varias etapas. Primero se construyó el corpus de textos originales y sus paráfrasis. Para la construcción del corpus se reunieron textos cortos de diversas fuentes y temáticas. Posteriormente se solicitó a varios voluntarios que intencionalmente reformularan o parafrasearan dichos textos. Se tomó como paráfrasis la reformulación de un texto conservando su contenido semántico pero variando el léxico, la sintaxis o la organización textual. La reformulación se realizó a dos niveles: nivel bajo (únicamente variación léxica) y nivel alto (variación léxica, sintáctica y discursiva).

A continuación, se realizó la anotación discursiva del corpus al estilo de la Rhetorical Structure Theory. Después, las estructuras discursivas de los textos originales y sus paráfrasis se compararon de manera manual. Asimismo, se realizó la comparación de las estructuras discursivas de los textos originales dedicados a la misma temática y se contabilizaron las diferencias y las coincidencias en cada caso.

En la siguiente etapa de la investigación, la similitud entre los fragmentos que se consideraron coincidentes en el análisis manual se midió de manera automática con un algoritmo de medición de similitud semántica. El primer paso de este algoritmo es el pre-procesamiento que consiste en la eliminación de los signos de puntuación, normalización a minúsculas, eliminación de palabras funcionales y, finalmente, reagrupación de palabras en familias léxicas. A continuación, la similitud entre los fragmentos es medida con base en la similitud entre todas las parejas de palabras contenidos en los mismos. La similitud a nivel de palabra se calcula a partir de su posición en la base de datos léxica

EuroWordNet. Los cálculos realizados permiten obtener un valor único que representa la similitud semántica entre los fragmentos.

Los resultados de la investigación indican que el grado de similitud entre los textos originales y sus paráfrasis es mayor que entre los textos originales dedicados a la misma temática. De esta manera, el estudio demuestra que con la metodología propuesta es posible identificar las copias realizadas por medio de la paráfrasis. Además, atendiendo a la cantidad de coincidencias encontradas en el análisis manual y el cálculo de similitud semántica, se puede identificar el nivel de complejidad de una paráfrasis.

13.4. Referencias

Tesis reseñadas

Barrón, Alberto (2007). Extracción automática de términos en contextos definitorios. Tesis de maestría, UNAM.

Cabrera, Luis Adrián (2011). TF-IDF para la obtención automática de términos y su validación mediante Wikipedia. Tesis de licenciatura, UNAM.

Castro, Brenda (2011). Detección de similitud textual mediante criterios de discurso y semántica. Tesis de licenciatura, UNAM.

Lázaro, Jorge (2010). Extracción de la terminología básica de las sexualidades en México a partir de un corpus lingüístico. Tesis de licenciatura, UNAM.

Reyes, Antonio (2002). Hacia una obtención computarizada de términos: Aplicación concreta al léxico de la física en el nivel bachillerato. Tesis de licenciatura, UNAM.

Rosas, Alejandro (2011). Análisis estilométrico para la detección de plagio. Tesis de licenciatura, UNAM.

Sierra, Gerardo (1999). Design of a concept-oriented tool for terminology. Tesis de doctorado, UMIST, Mánchester.

Lecturas sugeridas

Biber, Douglas, Susan Conrad, y Randi Reppen (1998). Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press.

Vargas Sierra, Chelo (2010). “Combinatoria terminológica y diccionarios especializados para traductores”. En Ibáñez Rodriguez, Miguel (Ed.): Lenguas de especialidad y terminología. Granada: Comares, pp. 17-46.

Otras citas

DEM. dem.colmex.mx

Diccionario explicativo del español. Gelbukh, Alexander y Grigori Sidorov (2003). Hacia la verificación de diccionarios explicativos asistidos por computadora. Estudios de Lingüística Aplicada 38, pp. 89-108.

Corpus de Informática en Español. Observatoire Linguistique Sense-Texte, Universidad de Montreal.

Freeling. http://nlp.lsi.upc.edu/freeling/

Signature. www.philocomp.net/humanities/signature.htm

C-value. Frantzi, Katerina, Sophia Ananiadou y Hideki Mima (2000). Automatic recognition of multi-word terms: The C-value/NC-value method. International Journal on Digital Libraries 3 (2), pp. 115-130.

EuroWordNet. Vossen, Piek (2004). EuroWordNet: A multilingual database of autonomous and language-specific Wordnets connected via an Inter-Lingual Index. International Journal of Lexicography 17 (2), pp. 161-173.

Capítulo 14

Aplicaciones en tecnologías del lenguaje

En la sociedad actual se cuenta con el acceso a grandes cantidades de información textual en formato electrónico y una de las exigencias básicas a la que todo ciudadano se enfrenta es el aprovechamiento de esta información para construir conocimiento. La sociedad de la información basa su valor en su capacidad de obtener, clasificar, sistematizar, sintetizar y aprovechar esta información. Para ello, las tecnologías del lenguaje, que constituyen la base de la ingeniería lingüística, aplican los conocimientos de la lengua en el desarrollo de sistemas informáticos que puedan reconocer, comprender, interpretar y generar lenguaje humano en todas sus formas.

Las tecnologías del lenguaje requieren de estudios teóricos del lenguaje natural, de una serie de metodologías y técnicas de análisis, así como de una serie de recursos lingüísticos, tales como lexicones, diccionarios y, por supuesto, corpus lingüísticos. A continuación se describen algunas aplicaciones de las tecnologías del lenguaje que se han realizado en el seno del Grupo de Ingeniería Lingüística, y que han utilizado los corpus lingüísticos.

14.1. Extracción de información

Una de las áreas dentro de las tecnologías del lenguaje que han tenido un gran desarrollo en los últimos años es la que se refiere al diseño de sistemas automáticos para la extracción de información, en donde se busca de manera selectiva una serie de estructuras o combinaciones de datos, los cuales se encuentran, de manera explícita o implícita, dentro de un conjunto de textos.

Entre los aspectos que han tomado gran relevancia dentro de la extracción de información, cabe señalar la extracción de información terminológica y conceptual, proyectada para la elaboración de ontologías y diccionarios electrónicos. Dicha tarea consiste en la identificación, almacenamiento y administración de términos y conceptos a partir de textos especializados. Los términos son unidades de la lengua y, por tanto, pueden ser identificados con base en sus características lingüísticas superficiales y su distribución en los textos de especialidad frente a otros tipos textuales. Mientras tanto, los conceptos son unidades de conocimiento abstracto que pueden ser expresados de muy diversas maneras. Por ello, la extracción automática de información conceptual requiere de técnicas más complejas. Una de las fuentes importantes de esta información es el conocimiento definitorio, el cual permite inferir el significado de los términos a partir de la descripción de sus atributos, características o relaciones semánticas en las que participan. Existen dos maneras de obtener el conocimiento definitorio a partir de corpus textuales:

- Extracción de contextos definitorios que proporcionan descripciones generales del significado de los términos.
- Extracción de relaciones semánticas (hiperonimia, hiponimia, holonimia, meronimia, sinonimia, etc.) que se establecen entre los conceptos propios de un ámbito de especialidad.

Extracción de contextos definitorios

La metodología para extraer contextos definitorios está basada en reglas lingüísticas y consiste en la búsqueda automática de ocurrencias de patrones verbales definitorios. El ECODE (Extractor de COntextos DEfinitorios), que fue desarrollado como investigación doctoral por Rodrigo Alarcón en el Instituto Universitario de Lingüística Aplicada (IULA), es una herramienta que obtiene los contextos definitorios a partir de un corpus textual y los clasifica de acuerdo con el tipo de definición. Asimismo, la herramienta es capaz de identificar las partes constitutivas de un contexto definitorio: el término y su definición. Para realizar la extracción el sistema utiliza una gramática de patrones verbales definitorios y abarca un procesamiento automático de los candidatos a contextos definitorios: primeramente, un filtro de contextos no relevantes, esto es, aquellos contextos donde, a pesar de tener un predicación verbal definitoria, no se define un término; luego, la identificación de los elementos constitutivos del contexto definitorio, es decir el término y la definición; finalmente, una ponderación de resultados para determinar cuáles son los mejores contextos definitorios propuestos por el sistema.

Para el desarrollo del sistema se utilizó como corpus de pruebas el Corpus Técnico del IULA en el área de genoma en español. Este corpus de prueba estuvo integrado por las oraciones donde apareciera cualquiera de los 163 términos del Vocabulario Básico del

Genoma Humano y quedó conformado por 1,091,946 palabras, para un total de 38,427 oraciones. Está etiquetado morfosintácticamente con las etiquetas de EAGLES, de forma que se pueden hacer búsquedas de palabras o conjuntos de palabras, de formas gramaticales y de la ocurrencia de uno o varios lemas específicos. Por ejemplo:

```
WORD=‘‘se’’, LEMMA=‘‘definir’’, POS=‘‘RG000’’
```

Gracias a esto, se buscaron 29 verbos definitorios, tomando en cuenta ciertas restricciones verbales que se imponen a la raíz del verbo, ya que contribuye a filtrar contextos no relevantes. Por ejemplo, el verbo “definir” permite recuperar información definitoria prácticamente con cualquier forma verbal o persona gramatical, en tanto que para otros verbos, como permitir o contar, se utilizan solo algunas formas verbales o personas. Asimismo, a cada verbo se le asoció un nexo para recuperar buenos candidatos a contextos definitorios. Entre los nexos se encuentran los determinantes, los adverbios “también” y “como”, así como las preposiciones “de”, “en”, “para” y “por”.

Una vez que se diseñó el algoritmo para extraer contextos definitorios, se utilizó un corpus de evaluación, ahora en el área de medicina, para evaluar el funcionamiento del sistema a partir de la búsqueda de los lemas de los verbos contenidos en la gramática de patrones verbales asociados a distintos tipos de definiciones.

Extracción automática de relaciones léxico-semánticas a partir de textos especializados

El enfoque de la semántica léxica es una de las perspectivas ampliamente utilizadas para la extracción de la información conceptual. La tarea de extracción de relaciones semánticas consiste en la identificación de los patrones más característicos de la relación que servirían para extraer las instancias más confiables. Los principales tipos de métodos utilizados para la extracción de las relaciones semánticas son: correspondencia de patrones, enfoque de agrupamiento basado en la distribución del contexto en un corpus y subsunción de conceptos. Los patrones pueden ser identificados de manera manual o aprendidos por medio de técnicas de aprendizaje automático.

En su tesis doctoral en Ciencia de la Computación, Olga Acosta propone una metodología para la extracción automática de instancias de la relación conceptual hiponimia-hiperonimia con base en un corpus de textos especializados.

El objetivo general del trabajo fue desarrollar una metodología para la extracción automática de un subconjunto de relaciones semánticas de hiponimia-hiperonimia implícitas en candidatos a contextos definitorios extraídos de corpus de dominio específico.

La investigación tenía las siguientes etapas principales. En primer lugar, se constituyó el corpus de estudio. Para ello, fue recolectado un conjunto de documentos del dominio médico a partir de MedLine en español. Los documentos pertenecían a dos géneros diferentes: por un lado, textos de tipo enciclopédico dedicados a algún tema de interés en los que se podía encontrar con frecuencia una definición y más detalles sobre el concepto definido; por otro lado, noticias de salud relacionadas con alguna enfermedad, tratamiento o estudio. Asimismo, se recolectaron cuatro libros en formato PDF perteneciente también al ámbito de la medicina. El tamaño total del corpus fue de 1.3 millones de palabras.

En segundo lugar, se realizó el preprocesamiento del corpus que consistía en la eliminación de fragmentos textuales irrelevantes, segmentación en oraciones y etiquetado POST.

En tercer lugar, se llevó a cabo la extracción de las instancias de la relación hiponimia-hiperonimia. Para ello, se construyó y se aplicó al corpus una gramática de expresiones regulares que considera el comportamiento sintáctico prototípico de los elementos de un contexto definitorio. Se tomó en cuenta que en los textos reales los patrones verbales definitorios no se usan únicamente para introducir definiciones (por ejemplo, el verbo ser). Por tanto, para aumentar la precisión de la extracción automática se propusieron dos filtros adicionales. El primero tenía por objetivo garantizar la existencia de un término, un patrón verbal y un hiperónimo (expresado a través del género próximo) en el candidato a contexto definitorio. El segundo permitió eliminar aquellos casos en los que la predicación expresaba otro tipo de relaciones semánticas como meronimia-holonimia o causalidad.

En cuarto lugar, se realizó la extracción de términos y de sus hiperónimos. En esta fase se tomó en consideración que los núcleos de los sintagmas nominales en los que esperaríamos localizar el hiperónimo pueden ser palabras con un significado muy amplio (por ejemplo, clase, tipo, especie, etc.) cuando se encuentran precediendo un sintagma preposicional introducido con la preposición “de”. En estos casos los hiperónimos se encuentran normalmente en el sintagma preposicional. Por tanto, se consideraban como candidatos para la extracción las unidades nominales precedidas por la preposición “de”.

Por último, a partir de los hiperónimos recolectados se extrajeron los hipónimos correspondientes. Una de las propuestas de la tesis es la extracción de hipónimos que tengan un hiperónimo como núcleo sintáctico. La propuesta está enfocada a la extracción de las categorías subordinadas del hiperónimo más relevantes priorizando los adjetivos relacionales como proveedores de un mayor número de propiedades debido a su origen nominal.

La aportación principal de la tesis es el desarrollo de un programa denominado ExtReLex que a partir de un corpus de dominio específico extrae un conjunto de contextos definitorios analíticos candidatos y las relaciones de hiponimia-hiperonimia implícitas en estos fragmentos textuales. Asimismo, la herramienta extrae un conjunto de hipónimos derivados de los hiperónimos más frecuentes.

14.2. Traducción automática

Históricamente, la traducción automática se considera una de las primeras aplicaciones de Tecnologías de Lenguaje y, a la vez, una de sus aplicaciones más ambiciosas. En las últimas dos décadas el campo ha tenido un desarrollo vertiginoso motivado por las necesidades de la comunicación multilingüe en un mundo globalizado. Actualmente, el Grupo de Ingeniería Lingüística está comenzando su labor en este ámbito complejo, el cual involucra la aplicación y el perfeccionamiento de la mayoría de las tareas del Procesamiento del Lenguaje Natural.

Análisis lingüístico de la traducción automática para su evaluación

Con el fin de indagar en la problemática de la evaluación de sistemas de traducción automática, Marina Fomicheva realizó la tesis de maestría en Lingüística Aplicada de la UNAM, en la cual desarrolló una propuesta metodológica para comparar traducciones humana y automáticas en tres niveles de la lengua: léxico-terminológico, morfosintáctico y discursivo. Dicha propuesta fue aplicada a un corpus paralelo inglés-español de textos especializados del ámbito médico, que incluía textos originales, traducciones humanas y traducciones automáticas. Para el procesamiento del corpus se usaron varias herramientas de análisis lingüístico automático (el extractor terminológico basado en Wikipedia para el nivel léxico-terminológico, el etiquetador morfológico de Freeling para el nivel morfosintáctico y la interfaz de anotación de relaciones discursivas RSTTool para el nivel textual).

El trabajo tenía como objetivo general el estudio de las diferencias lingüísticas sistemáticas entre la traducción automática y la traducción humana y sus implicaciones para la evaluación automática de sistemas. Los objetivos específicos del trabajo eran:

- Detectar las diferencias en la distribución de unidades de análisis (unidades terminológicas, n-gramas de etiquetas POS y unidades discursivas) en la traducción humana y en las traducciones automáticas.
- Identificar las condiciones en las que se producen dichas diferencias teniendo en cuenta los textos originales y las estrategias de traducción humana y automática.

> La metodología del estudio incluía, por un lado, el uso de técnicas estilométricas para caracterizar el lenguaje de la traducción automática frente al de la traducción humana y, por otro lado, el análisis de los textos traducidos en términos de las modificaciones realizadas por los traductores humanos y por los sistemas de traducción automática con respecto al texto original, esto último con el fin de explicar las diferencias detectadas.

Entre los resultados relevantes del análisis, se encontró que a nivel léxico-terminológico las diferencias asociadas a la traducción humana son más frecuentes que las diferencias relacionadas con la traducción automática, debido a que a nivel de la unidad terminológica la traducción que ofrecen los sistemas es cercana al original, mientras que en la traducción humana se producen modificaciones opcionales que están condicionadas por factores estilísticos o pragmáticos y no por las divergencias léxicas entre los sistemas de la lengua fuente y la lengua meta.

A nivel morfosintáctico las diferencias relacionadas con la traducción humana se manifiestan, por un lado, en la explicitación obligatoria de los rasgos gramaticales que no se marcan en inglés, pero cuya marcación es necesaria en español y, por otro lado, en la implicitación opcional relacionada con una preferencia por el uso de construcciones más concisas propias de los textos de alto nivel de especialización. Mientras tanto, los sistemas de traducción automática suelen reproducir de manera literal las estructuras sintácticas de la lengua del original, lo cual en algunos casos resulta en la generación de oraciones agramaticales y en otros conlleva una falta de naturalidad en el discurso.

A nivel discursivo las diferencias que se originan en la traducción humana, reflejan la interpretación del texto original por parte del traductor y la adecuación a los patrones de organización textual en la lengua meta. Así, la modificación de la estructura discursiva en la traducción humana indica el grado de adaptación del original a las convenciones propias de la lengua de llegada. Mientras tanto, el cambio de las relaciones discursivas en la traducción automática es indicio de errores, ya que los sistemas no realizan ningún tipo de modificaciones a nivel extra-oracional.

Con respecto a las implicaciones del análisis realizado para la evaluación de los sistemas de traducción automática, las diferencias entre la traducción automática y la traducción humana relacionadas con las modificaciones opcionales realizadas por los traductores y las diferencias que se deben a la falta de modificaciones obligatorias en la traducción automática no tienen la misma relevancia para evaluar la calidad de esta última. Aunado a ello, el estudio demuestra que en la evaluación automática de sistemas deben utilizarse traducciones humanas de referencia de tipo análogo, es decir, aquellas en las que el número de las modificaciones opcionales se reduce al mínimo. Finalmente, el trabajo resalta la importancia del corpus lingüístico para la evaluación y el desarrollo

de la traducción automática.

El objetivo de la traducción automática es modelar, al menos en parte, el comportamiento de los traductores humanos. El análisis de corpus, que representa un acercamiento empírico a la descripción lingüística y aboga por el estudio de instancias concretas del uso de la lengua, es útil de cara a la descripción de las características de la traducción humana, ya que en numerosas ocasiones las decisiones del traductor se ven influenciadas por los factores sutiles relacionados con el uso de la lengua en contexto. El uso de corpus es crucial para el desarrollo de sistemas empíricos, ya que éstos utilizan modelos estadísticos de lenguaje y de traducción cuyos parámetros se estiman a partir de corpus monolingües y bilingües.

Obtención del léxico de un corpus paralelo náhuatl-español

Actualmente, la traducción automática y, en términos más generales, el procesamiento de datos multilingüe es una necesidad fundamental. Existen numerosas aplicaciones que realizan las tareas relacionadas con el análisis y/o generación de textos en varias lenguas con una calidad satisfactoria. Sin embargo, la gran mayoría de estas aplicaciones trabajan con lenguas de uso internacional como inglés, francés o español.

Así, es necesario también desarrollar recursos para otras lenguas. Para ello es crucial la disponibilidad de textos paralelos. Por este motivo, se realizó un estudio conjunto entre dos alumnos de posgrado de la UNAM, Sergio Páez, del Posgrado en Ciencias de Computación, y Gabriela Bayona, del Posgrado en Lingüística. El objetivo de este estudio fue crear un corpus paralelo informatizado náhuatl-español y diseñar herramientas para la alineación de los textos del mismo y la extracción del léxico bilingüe.

Se decidió usar los textos escritos en la variante clásica del náhuatl debido a que las variantes actuales todavía están en proceso de estandarización. El corpus está compuesto por cinco textos y cuenta con un total de 188,611 palabras en náhuatl y 210,587 palabras en español actual de México. El corpus fue etiquetado en formato XML con la siguiente información sobre la estructura de los documentos:

- <doc> para indicar inicio y fin de los documentos;
- <t> para marcar títulos;
- <st> para marcar subtítulos;
- <p> para delimitar párrafos;
- <o> para delimitar oraciones.

Después del etiquetado, se procedió a alinear los textos a nivel de párrafo. La alineación se llevó a cabo por medio de la enumeración automática de párrafos en los textos originales y en los textos traducidos. Después se verificó de manera manual que todos los documentos originales y sus traducciones tuvieran el mismo número de párrafos.

De la misma manera, para la alineación a nivel oracional se llevó a cabo la enumeración automática de todas las oraciones en los documentos originales y en los textos traducidos. La enumeración se reiniciaba en cada párrafo para simplificar la detección de las diferencias en la segmentación oracional. Se diseñó un programa que identifica las diferencias en el número de oraciones en cada párrafo. El programa recibe en la entrada el documento original y su traducción, y se encarga de verificar que cada párrafo contenga el mismo número de oraciones en ambos documentos. Si se encuentra un párrafo que no cumple con esta condición, el programa se detiene indicando al usuario el párrafo correspondiente. Las diferencias en la segmentación oracional se eliminan de manera manual y el ciclo se repite hasta que los documentos tengas el mismo número de oraciones en cada párrafo.

Después de la alineación oracional, se realizó la extracción del vocabulario bilingüe. Para ello se llevaron a cabo experimentos con varias medidas de asociación. Primero, se aplicó la medida de información mutua, que se utiliza tradicionalmente tanto en el procesamiento monolingüe, para identificar colocaciones, como en el escenario multilingüe, para identificar las posibles traducciones de una palabra. La información mutua mide la cantidad de información que la aparición de una palabra nos da sobre la aparición de otra. En el caso del procesamiento monolingüe calcula la probabilidad de que dos palabras aparezcan juntas, y la compara con la probabilidad de que dichas palabras aparezcan por separado. En el caso del procesamiento de corpus paralelos la información mutua calcula la probabilidad de que dos palabras aparezcan en las mismas oraciones en el texto original y en el texto traducido y la compara con la probabilidad de que aparezcan en oraciones diferentes. Las probabilidades se calculan con base en las frecuencias de aparición o coaparición de las palabras en el corpus.

Los resultados del experimento indican que este método tiene una limitación importante. La medida de información mutua se enfoca en la frecuencia de coaparición de las palabras, y no presta atención suficiente al número de veces en las que las palabras aparecen por separado. Para resolver este problema, se aplicó una medida alternativa, el índice de la $\phi 2$. Esta medida hace mejor consideración de ambos parámetros y, de acuerdo con los resultados del experimento, tiene una precisión más alta a la hora de indicar posibles alineaciones de palabras en el corpus paralelo.

14.3. Clasificación y agrupamiento

En un sentido amplio, agrupar significa reunir objetos que comparten propiedades similares y separar los objetos que no lo son, en tanto clasificar significa agrupar objetos en grupos o categorías previamente establecidos. Ambas actividades se llevan a cabo con múltiples aplicaciones, desde las clasificaciones hechas por expertos como en biología o bibliotecología, hasta los procesos automáticos como la secuenciación de ADN o los estudios de demanda en economía. En tecnologías de lenguaje también se tienen varias aplicaciones. En el Grupo de Ingeniería Lingüística se han usado con fines de lexicografía, lingüística forense o minería de textos. A continuación se muestra un ejemplo para un problema complejo, la clasificación de textos cortos, donde la información contenida en el contexto no aporta suficientes elementos para poder hacer un buen discernimiento entre un grupo u otro.

Agrupamiento de contextos definitorios

El ECODE proporciona finalmente una lista de contextos definitorios asignados a alguno de los tipos de definiciones: analítica, extensional, funcional o sinonímica, y organizada según la probabilidad de que sean en mayor o menor medida mejores contextos definitorios. Además de la clasificación de estos por su tipo de definición, también pueden ser agrupados según sus características semánticas. Esto es, se pueden agrupar los contextos definitorios polisémicos por sus diferentes significados o incluso por las características descritas en su definición. En el primer caso, tenemos por ejemplo el término virus, del cual se pueden tener por un lado contextos correspondientes al área de informática y por otro lado los correspondientes al área de medicina o de biología. En el segundo caso, podemos encontrar por ejemplo contextos con definiciones analíticas para el término gen, que por un lado lo describen como la unidad de la herencia y por otro como una secuencia de ADN. Por esta razón, Alejandro Molina realizó su tesis de maestría en Ciencias de la Computación en la UNAM orientada a desarrollar un algoritmo para poder llevar a cabo el agrupamiento automático de contextos definitorios según su significado, de tal forma que los resultados de la búsqueda de un término polisémico sean presentados mediante una clasificación semántica.

El desarrollo del algoritmo requirió la construcción de un corpus de prueba que fuera conveniente para la clasificación de contextos definitorios referentes a diferentes áreas de conocimiento, denominado Corpus de Términos Polisémicos en Español (CTPE). En este sentido, se propusieron inicialmente diez términos que tuvieran varias acepciones en el diccionario y que fuera factible encontrar una cantidad considerable de información sobre ellos en Internet. Los términos seleccionados fueron: aguja, barra, cabeza, casco, célula, golpe, punto, serie, tabla y ventana. Sin embargo, dado que se obtuvo una cantidad considerable de información, se decidió restringir el estudio a sólo

cuatro términos: barra, célula, punto y ventana. Se obtuvo un total de 14,731 candidatos a contextos definitorios con estos cuatro términos, a los cuales se les aplicó el ECODE, que entregó un archivo de salida con un total de 1,422 contextos definitorios clasificados según el tipo de definición: analíticas, extensionales y funcionales.

El algoritmo lleva a cabo tres grandes etapas. Dentro de la primera, el texto es procesado hasta llegar a su representación vectorial usando diversas técnicas del Procesamiento del Lenguaje Natural. En la segunda, se calcula la distancia entre cada vector utilizando la matriz de energía textual y, en la última etapa, se aplica el agrupamiento jerárquico con el método de los vecinos más próximos.

Se probó el comportamiento del algoritmo para cada término-tipo de definición y se analizaron los resultados tanto a nivel cualitativo como cuantitativo. La evaluación cualitativa consistió en la lectura directa e interpretación de los grupos generados, en tanto la cuantitativa se basa en coeficientes tradicionales de clasificación de información.

La ventaja más importante del algoritmo de agrupamiento es ser independiente del idioma, no requiere de ningún tipo de anotación lingüística, tampoco requiere de un conjunto de entrenamiento previo, ni es necesario indicar el número de grupos a generar y, finalmente, que es fácilmente configurable, ya que sólo depende del valor de corte por distancia.

14.4. Resumen automático

Ante la cantidad de información y la necesidad de generar conocimiento, los resúmenes automáticos se han vuelto una necesidad imperiosa y, por tanto, ha sido ampliamente desarrollada como un área de las tecnologías del lenguaje. Entre otras clasificaciones, se diferencian los resúmenes extractivos de los abstractivos, siendo los primeros extractos o fragmentos de los documentos, sin ninguna modificación, en tanto los segundos implican ya modificaciones, lo que se asemeja más al resumen humano. A continuación se describe una investigación que si bien forma parte de un método por abstracción, también es una adición al resumen extractivo.

Compresión automática de frases

La compresión de frases consiste en eliminar de forma automática las partes menos importantes de una oración o frase, lo que permite reducir la extensión de un texto para incluir mayor información en un espacio reducido. La investigación doctoral de Alejandro Molina, desarrollada en el LIA de la Universidad de Aviñón, en Francia, se orienta a desarrollar un método de compresión de frase aplicado en la generación de un resumen,

en el cual quedan eliminados algunos segmentos intra-oracionales de las frases. En este sentido, la compresión de frases establece un puente desde el resumen por extracción hacia la generación de resúmenes abstractivos.

El punto de partida del sistema es la segmentación discursiva. Se parte de la idea de que si una frase ya es suficientemente simple, puede ser considerada como un solo segmento discursivo, por lo que no tiene necesidad de ser comprimida. Por el contrario, si la frase es larga y compleja, estará formada por muchos segmentos discursivos donde habrá información a omitir. Para este fin, se utilizó DiSeg, un segmentador discursivo para el español, basado en la RST.

Para medir el desempeño de DiSeg, se evaluó hasta qué punto los seres humanos eliminan fragmentos textuales que corresponden a segmentos discursivos identificados por DiSeg. Para tal fin, tuvo lugar un experimento: después de reunir un corpus con textos cortos de cuatro géneros, se solicitó a cinco lingüistas eliminar palabras o grupos de palabras de las frases del corpus bajo las condiciones de no reescribir las frases, no modificar el orden de las palabras, no sustituir las palabras, asegurarse que las frases comprimidas fueran gramaticales y asegurarse que las frases y el texto resultante conservara el significado de origen.

En este trabajo también fueron analizados 3 criterios para decidir si un segmento debía eliminarse o conservarse:

La gramaticalidad de la frase resultante. Se refiere a garantizar que todas las frases del resumen, asi como el resumen en su totalidad, sean gramaticales, ya que basta la existencia de un pequeño error gramatical para que la calidad final del resumen sea puesta en duda. Para ello, se utilizó un modelo probabilístico del lenguaje que estima la probabilidad de generar una secuencia de palabras a partir de las frecuencias de n-gramas obtenidas de un corpus.

La informatividad de los segmentos discursivos. Se entiende como la cantidad de información importante retenida. Para medir esta información se utilizó el modelo de energía textual, que puede ser aplicado incluso cuando se trata con frecuencias unitarias, pues toma en cuenta el contexto de las frases y las relaciones léxicas entre ellas.

La tasa de compresión. Expresa el volumen conservado después de que la frase es resumida.

Al no contar con un corpus en compresión de frases en español, se desarrolló una plataforma de anotación multitudinaria además de que se lanzó una campaña de anotación masiva para conformar el corpus. Primero se eligieron 30 documentos que

fueron segmentados usando DiSeg y CoSeg. Conviene decir que 150 voluntarios fueron registrados para anotar el corpus. Al final de ésta, se capturó más de 60 mil veces la decisión de eliminar o conservar un segmento discursivo y al mismo tiempo también se obtuvieron 2,877 resúmenes manuales.

Por la posibilidad de aplicar tanto la metodología como el marco teórico en otras direcciones e idiomas, se ha puesto a disposición de la comunidad los datos generados de la investigación para futuros estudios.

14.5. Minería de textos

En tanto la minería de datos descubre patrones en grandes cantidades de datos, sean económicos o científicos, en la minería de textos este trabajo se vuelve más complejo, pues aquí se tiene información no estructurada, razón por la que los textos tienen que convertirse a valores medidos a través de alguna métrica, como la presencia de palabras o la frecuencia con que aparecen. Además, también es de considerar la dimensionalidad de información, pues para empezar existe una gran cantidad de textos y cada uno contiene miles de palabras. Los beneficios de la minería de textos resultan en innovaciones tecnológicas que coadyuvan al entendimiento y uso mejor de la información disponible en repositorios de documentos. Estos documentos, para los fines de una investigación, conforman un corpus.

Existen varias aplicaciones de la minería de textos, a continuación se describe una investigación orientada a la minería de opiniones.

Clasificación de opiniones

El análisis de sentimientos o la minería de opiniones es un área de las tecnologías del lenguaje que busca determinar la orientación de la opinión que se tiene sobre un objeto, ya sea un producto o servicio, una persona o un partido, una acción o un evento, por ejemplo. Dentro de esta área, Pavel Soriano presentó la tesis titulada “Clasificación de opiniones mediante aprendizaje de máquinas: el caso de reseñas sobre películas”, para obtener el título de Ingeniero en Computación. En la investigación se ofrece un sistema capaz de clasificar en positivas o negativas, y a nivel oración, opiniones sobre películas. Lo anterior se logra con ayuda de técnicas del aprendizaje de máquinas y del procesamiento de lenguaje natural.

Se constituyó un corpus de artículos sobre películas estrenadas de noviembre 2009 a enero 2010, obtenidas de Wikipedia. Primero se exportó los artículos de películas utilizando la opción Special:Export de Wikipedia; luego se seleccionó las películas del

periodo mediante el empleo de expresiones regulares para extraer los títulos y las fechas de estreno, tomando en cuenta los datos que aparecen de cada película en las cajas de información (o infoboxes); y finalmente se extrajo las reseñas y otros datos de interés, tales como la calificación dada por los usuarios y el número de votos. Con ello, se obtuvo 171 películas y 7,089 reseñas, estas últimas conteniendo 123,878 enunciados en total.

Para entrenar al clasificador, se propusieron y se pusieron marcha cuatro métodos que tienen por objetivo mejorar el desempeño del clasificador. El primer sistema separa los enunciados de la opiniones por polaridad (positivos o negativos) dependiendo de la calificación que los usuarios dieron a la película. En el segundo, se realiza una selección de rasgos de los enunciados y, con base en ellos, se agrupan para obtener conjuntos de enunciados subjetivos similares. En el tercero quedan identificados los enunciados subjetivos por la presencia en ellos de adjetivos, adverbios o disparadores de presuposición. Finalmente, en el cuarto se entrena, prueba y valida un clasificador bayesiano usando los enunciados contenidos. Cabe decir que se alcanzó la mayor exactitud usando enunciados subjetivos identificados por la presencia de adjetivos o adverbios. La tesis termina brindando una conclusión en la cual se presentan las ventajas y desventajas del sistema, así como el trabajo a realizar a futuro.

En este proyecto se utilizó como lenguaje de programación Pyton, dado que es un lenguaje simple con excelente funcionalidad para procesar información lingüística. También se utilizó ECLIPSE IDE, un ambiente de desarrollo de software conformado por un entorno de desarrollo integrado y un sistema de plugins o complementos. Finalmente también se utilizó Pydev, que permite programar en Pyton, Jyton e IronPyton.

14.6. Referencias

Tesis reseñadas

Acosta, Olga (2012). Extracción automática de relaciones léxico-semántica a partir de textos especializados. Tesis de doctorado, UNAM.

Alarcón, Rodrigo (2009). Extracción automática de contextos definitorios en corpus especializados. Tesis de doctorado, IULA, España.

Fomicheva, Marina (2012). Análisis lingüístico de la traducción automática para su evaluación. Tesis de maestría, UNAM.

Molina, Alejandro (2009). Agrupamiento semántico de contextos definitorios. Tesis de maestría, UNAM.

Molina, Alejandro (2013). Compression automatique de phrases: un étude vers la génération de résumés. Tesis de doctorado, Université d'Avignon, Francia.

Lecturas sugeridas

Llisterri, Joaquim (2003). "Lingüística y tecnologías del lenguaje". Lynx. Panorámica de Estudios Lingüísticos 2, pp. 9-71.

McNaught, John. (1993). "User needs for textual corpora in natural language processing". Literary and Linguistic Computing 8(4), pp. 227-234.

Torres Moreno, Juan Manuel (2011). Résumé automatique de documents: une approche statistique. Hermes-Lavoisier, Francia.

[1] [2] [3] [11] [12] [18] [20] [23] [27] [35] [45] [51] [52] [53] [55] [54] [56] [58] [64] [65] [62] [67] [72] [73] [68] [4] [5] [19] [30] [38] [66] [69] [71] [75] [76] [80] [82] [6] [7] [9] [15] [16] [21] [22] [31] [33] [59] [40] [44] [47] [48] [57] [70] [74] [8] [13] [14] [17] [25] [26] [29] [32] [39] [41] [79] [42] [43] [49] [60] [63] [72] [81] [10] [24] [28] [34] [36] [37] [61] [77] [78] [46] [50]

Bibliografía

[1] Acosta López, Olga Lidia: *Extracción automática de relaciones léxico-semántica a partir de textos especializados*. Tesis de Doctorado, UNAM, México, 2012.

[2] Alarcón Martínez, Rodrigo: *Análisis lingüístico de contextos definitorios en textos de especialidad*. Tesis de Licenciatura, UNAM, México, 2003.

[3] Alarcón Martínez, Rodrigo: *Extracción automática de contextos definitorios en corpus especializados*. Tesis de Doctorado, IULA, Universitat Pompeu Fabra, España, 2009.

[4] Arias Álvarez, Beatriz: *Confección de un corpus para conocer el origen, la evolución y la consolidación del español en la Nueva España*. En Enrique Arias, Andrés (editor): *Diacronía de las lenguas iberorrománicas: Nuevas aportaciones desde la lingüística de corpus*, páginas 55–78. Iberoamericana, Madrid, 2010.

[5] Arrarte, Gerardo: *Normas y estándares para la codificación de textos y para la ingeniería lingüística*. En Blecua, José Manuel, Gloria Clavería, Carlos Sánchez y Joan Torruella (editores): *Filología e Informática: Nuevas tecnologías en los estudios filológicos*, páginas 17–44. Editorial Milenio-Universitat Autònoma de Barcelona, Barcelona, 1999.

[6] Arroyo, Susana: *El Primero Sueño de Sor Juana: estudio semántico y retórico*, volumen 16 de *Cuadernos del Seminario de Poética*. UNAM-IIF/ITESM-CEM, México, 1993.

[7] Bach, Carme, Roser Saurí, Jordi Vivaldi y M. Teresa Cabré: *El corpus de l'IULA: descripció*. Papers de l'IULA, Sèrie Informes 17. Universitat Pompeu Fabra-Institut Universitari de Lingüística Aplicada, Barcelona, 1997.

[8] Barcala, Francisco Mario, Cristina Blanco y Victor Manuel Darriba: *Metodología para la construcción de córpora textuales estructurados basados en XML*. Procesamiento del Lenguaje Natural, 36:9–16, 2006.

[9] Barnbrook, Geoff: *Language and Computers: A Practical Introduction to the Computer Analysis of Language*. Edinburgh University Press, Edinburgh, 1996.

[10] Barrios, María A.: *Diccionarios combinatorios del Español: diferencias y semejanzas entre 'Redes y Práctico'*. En *Actas del II Congreso Internacional de Lexicografía Hispánica*, páginas 197–203, Alicante, España, 2008. Biblioteca Virtual Miguel de Cervantes.

[11] Barrón Cedeño, Luis Alberto: *Extracción automática de términos en contextos definitorios*. Tesis de Licenciatura, UNAM, México, 2007.

[12] Benítez Rosete, Valeria Amanda: *Anáforas en la expansión de contextos definitorios: una propuesta de etiquetado*. Tesis de Licenciatura, UNAM, México, 2008.

[13] Berber Sardinha, Tony: *Lingüística de Corpus: Histórico e Problemática*. DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada, 16(2):323–367, 2000.

[14] Biber, Douglas: *Representativeness in Corpus Design*. Literary and Linguistic Computing, 8(4):243–257, 1993.

[15] Biber, Douglas, Conrad Susan y Reppen Randi: *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press, Cambridge, 1998.

[16] Botley, Simon y McEnery Tony: *Corpus-based and Computational Approaches to Discourse Anaphora*, volumen 3 de *Studies in corpus linguistics*. John Benjamins, Amsterdam, 2000.

[17] Busa, Roberto: *The Annals of Humanities Computing: The Index Thomisticus*. Computers and the Humanities, 14(2):83–90, 1980.

[18] Cabrera Diego, Luis Adrián: *TF-IDF para la obtención automática de términos y su validación mediante Wikipedia*. Tesis de Licenciatura, UNAM, México, 2011.

[19] Carré, René: *Los bancos de sonidos*. En Vidal Beneyto, José (editor): *Las industrias de la lengua*, páginas 108–118. Fundación Germán Sánchez Ruipérez/Ediciones Pirámide, Madrid, 1991.

[20] Castro Roldán, Brenda Gabriela: *Detección de similitud textual mediante criterios de discurso y semántica*. Tesis de Licenciatura, UNAM, México, 2011.

[21] Charniak, Eugene: *Statistical Language Learning*. The MIT Press, Cambridge, 1996.

[22] Cole, Ronald, Joseph Mariani, Hans Uszkoreit, Annie Zaenen y Victor Zue (editores): *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, Cambridge, 1996.

[23] Cruz Domínguez, Irasema: *El sintagma nominal en la extracción de relaciones léxico-semánticas de contextos definitorios: el caso de la preposición DE*. Tesis de Licenciatura, UNAM, México, 2011.

[24] Cunha, Iria da, Juan Manuel Torres-Moreno y Gerardo Sierra: *On the Development of the RST Spanish Treebank*. En *5th Linguistic Annotation Workshop*, páginas 1–10, Portland, Oregon, 2011. Association for Computer Linguistics.

[25] Davies, Mark: *Un corpus anotado de 100.000.000 palabras del español histórico y moderno*. Procesamiento del Lenguaje Natural, 29:21–27, 2002.

[26] Estruch, Mónica, Juan Maria Garrido, Joaquim Llisterri y Montserrat Riera: *Técnicas y procedimientos para la representación de las curvas melódicas*. Revista de lingüística teórica y aplicada, 45(2):59–87, 2007.

[27] Fomicheva, Marina: *Análisis lingüístico de la traducción automática para su evaluación*. Tesis de Licenciatura, UNAM, México, 2012.

[28] Francis, Winthrop Nelson y Henry Kučera: *Manual of Information to Accompany A Standard Corpus of Present-day Edited American English, for Use with Digital Computers*. Informe técnico, Department of Lingustics, Brown University, Providence,Rhode Island, 1979.

[29] Frantzi, Katerina, Sophia Ananiadou y Hideki Mima: *Automatic recognition of multi-word terms: The C-value/NC-value method*. International Journal on Digital Libraries, 3(2):115–130, 2000.

[30] Garduño, Gabriel, Gerardo Sierra y Alfonso Medina: *Herramientas de análisis para el Corpus Lingüístico en Ingeniería*. En Arias Estrada, Miguel y Alexander Gelbukh (editores): *Avances en la Ciencia de la Computación*, páginas 219–226. Sociedad Mexicana de Ciencia de la Computación, Colima, 2004.

[31] Garside, Roger, Geoffrey Leech y Tony McEnery: *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Addison Wesley Longman, New York, 1997.

[32] Gelbukh, Alexander y Grigori Sidorov: *Hacia la verificación de diccionarios explicativos asistidos por computadora*. Estudios de Lingüística Aplicada, 22(38):89–108, 2003.

[33] Goldfarb, Charles F. y Paul Prescod: *The XML handbook*. Prentice-Hall, Oxford, 1998.

[34] Grefenstette, Gregory y Pasi Tapanainen: *What is a word, What is a sentence? Problems of Tokenization*. En *3rd Conference on Computational Lexicography and Text Research (COMPLEX'94)*, páginas 79–87, Budapest, 1994.

[35] Hernández Angulo, Ariadna Carolina: *Análisis lingüístico de definiciones analíticas para la búsqueda de reglas que permitan su delimitación automática*. Tesis de Licenciatura, UNAM, México, 2009.

[36] Hernández Mena, Carlos Daniel y Abel Herrera Camacho: *CIEMPIESS: A New Open-Sourced Mexican Spanish Radio Corpus*. En *9th International Conference on Language Resources and Evaluation (LREC'14)*, páginas 371–375, Reykjavik, Islandia, 2014.

[37] Jungl, Manuel, Silvia Molina, Pilar Pérez y Juan Morales: *El proyecto AGLE (Archivo Gramatical de la Lengua Española): desarrollo y perspectivas*. En *XXXV Simposio Internacional de la Sociedad Española de Lingüística*, páginas 1048–1059, León, España, 2006. Biblioteca Virtual Miguel de Cervantes.

[38] Kahrel, Peter, Ruthanna Barnett y Geoffrey Leech: *Towards cross-linguistic standards or guidelines for the annotation of corpora*. En Garside, R., Geoffrey Leech y Tony McEnery (editores): *Corpus Annotation. Linguistic Information from Computer Text Corpora*, páginas 231–242. Longman, London, 1997.

[39] Kilgarrif, Adam y Gregory Grefenstete: *Introduction to the special issue on the web as corpus*. Computational linguistics, 29(3):333–347, 2003.

[40] Lara, Luis Fernando, Roberto Ham Chande y Ma. Isabel García Hidalgo: *Investigaciones lingüísticas en lexicografía*. El Colegio de México, México, 1979.

[41] Leech, Geoffrey: *Corpus annotation schemes*. Literary and Linguistic Computing, 8(4):275–281, 1993.

[42] Llisterri, Joaquim: *Transcripción, etiquetado y codificación de corpus orales*. Revista española de lingüística aplicada, 1:53–82, 1999.

[43] Llisterri, Joaquim: *Lingüística y tecnologías del lenguaje*. Lynx. Panorámica de Estudios Lingüísticos, 2:9–71, 2003.

[44] López Chávez, Juan y Marina Arjona Iglesias: *Lexicometría y fonometría del Primero Sueño de Sor Juana Inés de la Cruz*. Universidad Nacional Autónoma de México, México, 1994.

[45] Lázaro Hernández, Jorge Adrián: *Extracción de la terminología básica de las sexualidades en México a partir de un corpus lingüístico*. Tesis de Licenciatura, UNAM, México, 2010.

[46] MacMullen, John: *Requirements Definition and Design Criteria for Test Corpora in Information Science*. SILS Technical Report 2003-03, School of Information and Library Science, University of North Carolina, Chapel Hill, NC, 2003.

[47] Macwhinney, Brian: *The CHILDES Project: The database*. Psychology Press, 2000.

[48] McEnery, Tony y Andrew Wilson: *Corpus linguistics*. Edinburgh University Press, Edinburgh, 1996.

[49] McNaught, John: *User Needs for Textual Corpora in Natural Language Processing*. Literary and Linguistic Computing, 8(4):227–234, 1993.

[50] McTait, Kevin: *A Survey of Corpus Analysis Tools*. (Véase en www.iling.unam.mx/cursocorpus/Survey.pdf), 1998.

[51] Medina Urrea, Alfonso: *Investigación cuantitativa de afijos y clíticos del español de México: Glutinometría en el Corpus del Español Mexicano Contemporáneo*. Tesis de Doctorado, El Colegio de México, México, 2003.

[52] Medina Urrea, Alfonso y Carlos Méndez Cruz: *Arquitectura del Corpus Histórico del Español en México (CHEM)*. En Hernández, A. y J.L. Zechinelli (editores): *Avances en la Ciencia de la Computación*, páginas 248–253. Sociedad Mexicana de Ciencia de la Computación, 2006.

[53] Méndez Cruz, Carlos Francisco: *Identificación automática de categorías gramaticales en español del siglo XVI*. Tesis de Licenciatura, UNAM, México, 2009.

[54] Molina Villegas, Alejandro: *Agrupamiento semántico de contextos definitorios*. Tesis de Licenciatura, UNAM, México, 2009.

[55] Molina Villegas, Alejandro: *Compression automatique de phrases: un étude vers la génération de résumés*. Tesis de Doctorado, LIA, Université d'Avignon, Francia, 2013.

[56] Navarro-Colorado, Borja: *Metodología, construcción y explotación de corpus anotados semántica y anafóricamente*. Tesis de Doctorado, Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, España, 2007.

[57] Oakes, Michael P.: *Statistics for Corpus Linguistics*. Edinburgh University Press, Edinburgh, 1998.

[58] Palacios Sierra, Margarita: *La palabra "democracia" en el Congreso de la Unión: usos y sentidos (México 2008)*. Tesis de Doctorado, UNAM, México, 2013.

[59] Pérez Guerra, Javier: *Introducción a la lingüística de corpus: un ejercicio con herramientas informáticas aplicadas al análisis textual*. Tórculo Edicións, Santiago de Compostela, 1998.

[60] Quirk, Randolph: *Towards a description of English usage*. Transactions of the Philological Society, 59(1):40–61, 1960.

[61] Ramírez, Gaspar, James L. Fidelholtz, Héctor Jiménez y Grigori Sidorov: *Elaboración de un diccionario de verbos del español a partir de una lexicografía sistemática*. En *VII Encuentro Nacional de Ciencias de la Computación, ENC-2006*, páginas 270–275, San Luís Potosí, México, 2006. Avances en Ciencia de la Computación.

[62] Reyes Careaga, Teresita Adriana: *Reglas de correspondencia entre sonido y grafía en el español hablado en México en el siglo XVI para la creación de un transcriptor automático: una aportación al Corpus Histórico del Español de México (CHEM)*. Tesis de Licenciatura, UNAM, México, 2008.

[63] Reyes-Careaga, Teresita Adriana, Alfonso Medina Urrea y Ge
btxfnamespacelong rar
btxfnamespacelong do Eugenio Sierra Martínez: *Un corpus para la investigación en la extracción de términos y contextos definitorios: hacia un diccionario de las sexualidades en México*. Debate Terminológico, 7:24–35, 2011.

[64] Reyes Pérez, Antonio: *Hacia una obtención computarizada de términos. (Aplicación concreta al léxico de la física en el nivel bachillerato)*. Tesis de Licenciatura, UNAM, México, 2002.

[65] Reyes Pérez, Antonio: *Estructuración semántico-pragmática del léxico en dominios restringidos para sistemas de recuperación de información*. Tesis de Licenciatura, UNAM, México, 2006.

[66] Rojo, Guillermo: *La explotación de la Base de Datos Sintácticos del español actual (BDS)*. En De Kock, Josse (editor): *Lingüística con corpus. Catorce aplicaciones sobre el español*, volumen 7, páginas 255–286. Universidad de Salamanca, España, 2001.

[67] Rosas González, Alejandro: *Análisis estilométrico para la detección de plagio*. Tesis de Licenciatura, UNAM, México, 2011.

[68] Sierra, Gerardo: *Design of a concept-oriented tool for terminology*. Tesis de Doctorado, UMIST-University of Manchester, Inglaterra, 1999.

[69] Sierra, Gerardo, Rodrigo Alarcón, César Aguilar, Alberto Barrón, Valeria Benítez y Itzia Baca: *Corpus de contextos definitorios: una herramienta para la lexicografía y la terminología*. En Estrada, Z. y A. Munguia (editores): *IX Encuentro Internacional de Lingüística en el Noroeste*, páginas 15–17. Universidad de Sonora, Hermosillo, México, 2006.

[70] Sinclair, John: *Corpus, concordance, collocation*. Oxford University Press, Oxford, 1991.

[71] Sinclair, John: *Creación de corpus*. En Vidal Beneyto, José (editor): *Las industrias de la lengua*, páginas 95–107. Fundación Germán Sánchez Ruipérez/Ediciones Pirámide, Madrid, 1991.

[72] Sánchez Sánchez, Mercedes y Carlos Domínguez Cintas: *El banco de datos de la RAE: CREA y CORDE*. Per Abbat: Boletín filológico de actualización académica y didáctica, 2:137–148, 2007.

[73] Sánchez Velázquez, Octavio: *La funcionalidad al interior de contextos definitorios con definiciones analíticas: El patrón sintáctico para + infinitivo*. Tesis de Licenciatura, UNAM, México, 2009.

[74] Torres Moreno, Juan Manuel: *Résumé automatique de documents: une approche statistique*. Hermès-Lavoisier, Francia, 2011.

[75] Torruela, Joan y Joaquim Llisterri: *Diseño de corpus textuales y orales*. En Blecua, J.M., G. Clavería, C. Sanchez y J. Torruela (editores): *Filología e informática: Nuevas tecnologías en los estudios filológicos*, páginas 45–77. Editorial Milenio-Universitat Autònoma de Barcelona, Barcelona, 1999.

[76] Vargas Sierra, Chelo: *Combinatoria terminológica y diccionarios especializados para traductores*. En Ibáñez Rodriguez, Miguel (editor): *Lenguas de especialidad y terminología*, páginas 17–46. Comares, Granada, 2010.

[77] Villaseñor, Luis, Antonio Massé y Luis Pineda: *The DIME Corpus*. En Zozaya, C., M. Mejía, P. Noriega y A. Sánchez (editores): *3º Encuentro Internacional de Ciencias de la Computación ENC'01*, Aguascalientes, México, 2001. SMCC, Tomo II.

[78] Villaseñor-Pineda, Luis, Manuel Montes-y Gómez, Dominique Vaufreydaz y Jean Francois Serignat: *Elaboración de un Corpus Balanceado para el Cálculo de Modelos Acústicos usando la Web*. En León, J. Díaz de, G. González y J. Figueroa (editores): *XII Congreso Internacional de Computación (CIC-2003)*, páginas 198–200, México, 2003. Avances en Ciencias de la Computación, IPN.

[79] Villayandre Llamazares, Milka: *Lingüística con corpus (I)*. Estudios humanísticos. Filología, 30:329–349, 2008.

[80] Volk, Martin: *Using the Web as Corpus for Linguistic Research*. En Pajusalu, R. y T. Hennoste (editores): *Tähendusepüüdja. Catcher of the Meaning. A Festschrift for Professor Haldur Õim*, páginas 1–10. Department of General Linguistics 3, University of Tartu, EstoniTartu, 2002.

[81] Vossen, Piek: *EuroWordNet: A Multilingual Database of Autonomous and Language-Specific Wordnets Connected via an Inter-LingualIndex*. International Journal of Lexicography, 17(2):161–173, 2004.

[82] Zampolli, Antonio: *Corpora de referencia*. En Vidal Beneyto, José (editor): *Las industrias de la lengua*, páginas 119–124. Fundación Germán Sánchez Ruipérez/Ediciones Pirámide, Madrid, 1991.