Lingüística de corpus e investigación lingüística: El *Corpus del español del siglo* XXI

Guillermo Rojo
http://gramatica.usc.es/persoas/guillermo.rojo

Universidade de Santiago de Compostela / Real Academia Española

Universidad de Salamanca

11 de marzo de 2015

La influencia de la LC en la lingüística actual

1. Con independencia de la orientación teórica en la que se trabaje, parece innegable que la LC es una corriente que ha cambiado considerablemente los modos de trabajo en lingüística.
2. Ese efecto general se hace más evidente en la lingüística española, tradicionalmente poco dada al análisis exhaustivo de datos reales.
3. En términos generales, puede decirse que la LC ha experimentado un fuerte desarrollo en los últimos cincuenta años y ha ampliado nuestros conocimientos tanto en extensión como en profundidad.

Los corpus textuales

1. Un corpus es un conjunto de textos (o fragmentos de textos) naturales, almacenados en formato electrónico, que resultan conjuntamente representativos de una variedad lingüística en su totalidad o en alguno de sus componentes, reunidos con el fin de que puedan ser estudiados científicamente.

Características de los corpus textuales

1. Los textos (o fragmentos de textos) tienen que haber sido producidos en condiciones reales y naturales.
2. Los textos deben estar en formato electrónico.
3. Deben ser representativos de la variedad de la que han sido extraídos. Además, deben estar equilibrados.
4. El corpus tiene que estar construido de modo que sea posible el análisis científico (no solo lingüístico).
5. Además, su organización debe permitir el enriquecimiento del corpus mediante la adición de codificación y anotación morfosintáctica, sintáctica, semántica y pragmática.

Corpus frente a textos

1. Según hemos visto, un corpus consiste en un conjunto de textos o fragmentos de textos con ciertas características adicionales (carácter natural, formato electrónico, carácter representativo de una variedad lingüística determinada, etc.).
2. La insistencia en la evolución de los corpus y los volúmenes que alcanzan en la actualidad, podría inclinar a pensar que la diferencia fundamental entre un conjunto de textos y un corpus está en el tamaño.
3. En esta línea, la diferencia entre lo que tradicionalmente se llamaba un *archivo* y un corpus sería el tamaño y, dada la indeterminación esperable, un nuevo caso de la llamada 'paradoja del montón' (*sorites*).

Corpus frente a textos

1. La diferencia es clara. Un corpus está constituido por textos, pero lo fundamental es que responde a un determinado diseño, consecuencia de la finalidad con la que ha sido concebido.
2. El diseño del corpus establece qué tipos de texto entran en su construcción, en qué proporciones, etc.

Los corpus de referencia

1. En términos generales, los corpus de referencia están situados entre dos extremos bien marcados:
   - A un lado, los corpus pequeños, especializados, con una codificación muy detallada y recursos adicionales (diferentes versiones del mismo texto, traducciones, imágenes de manuscritos, etc.).
   - Al otro, los corpus obtenidos directamente de los textos existentes en la red, formados ahora mismo por miles de millones de formas.

La *web* como corpus

1. En sus resultados más habituales y conocidos, parece claro que un conjunto obtenido a base de descargas automáticas de páginas situadas en la parte pública de la red no es un corpus en el sentido más estricto de la expresión. Sin entrar en muchos otros inconvenientes, carece de diseño.
2. Sin embargo, eso no significa que no puedan ser útiles. Para fenómenos de muy baja frecuencia, por ejemplo, no hay más remedio que recurrir a ellos o directamente a buscadores.
3. Por otro lado, hay que reconocer que las técnicas de construcción de estos conjuntos han mejorado sensiblemente: mejores reconocedores de lengua, filtros para tipos de páginas, detección de repeticiones... Véase, como muestra de lo que se puede hacer ahora mismo el corpus *EsTenTen* que ha construido Adam Kilgarrif.

Los corpus de referencia

1. Los corpus de referencia se sitúan en el centro del difícil balance producido por la tensión entre costes y tamaño por un lado, codificación y diseño equilibrado por la otra.
2. Son el único modo de poder trabajar con ciertos tipos de textos que, por diferentes razones, no se encuentran en la red.
3. La codificación que añaden a los textos es el único modo de lograr la recuperación selectiva de la información y, por tanto, de poder comparar el modo en que un fenómeno o expresión se presentan en textos de diferentes épocas, países, tipos, etc.

Datos y herramientas

1. En la aproximación que practica la mayor parte de las corrientes lingüísticas, los datos están en los textos, que son el resultado de la actividad lingüística de la comunidad correspondiente.
2. Los conocimientos teóricos y las hipótesis de partida nos dicen qué tenemos que buscar, pero necesitamos también las herramientas adecuadas para extraer los datos que nos interesan y poder analizarlos.
3. Con una imagen sencilla, la importancia de las herramientas se valora adecuadamente si pensamos en la contemplación del cielo a simple vista, con unos prismáticos, con telescopios de diferentes alcances, etc.

Los textos y la codificación

1. Aunque es inevitable que, al hablar de corpus textuales, se proyecte una imagen según la cual lo que contienen en su interior es un conjunto más o menos amplio de versiones electrónicas de lo que antes fueron textos impresos de novelas, noticias, etc. o transcripciones de conversaciones, entrevistas, etc., la construcción de un corpus supone un enorme trabajo de codificación de los textos.
2. La codificación consiste en la adición al texto nuclear (es decir, la novela, la noticia, la transcripción de la conversación, etc.) de aquellos datos necesarios para que luego se pueda realizar la extracción selectiva de la información.

Los textos y la codificación

1. Esas indicaciones adicionales, que reciben con frecuencia el nombre de *metadatos*, tienen que figurar de modo que estén claramente diferenciadas del texto en sí, para que la aplicación de búsqueda pueda localizarlas e identificarlas.
2. Para lograrlo, hay que utilizar un lenguaje de marcación. El utilizado habitualmente en este momento es XML (de *eXtended Mark-up Language*).

Codificación extratextual e intratextual

1. A la codificación extratextual corresponden fundamentalmente los datos bibliográficos (en textos escritos y sus equivalentes en textos orales) y los valores que presenta cada texto con respecto a los rasgos utilizados en la confección del corpus (país, año, tipo de texto, etc.; sexo, edad, nivel sociocultural, etc.).
2. Todos esos datos van en lo que se llama habitualmente cabecera del texto, que es el lugar al que irá a buscarlos la aplicación de consulta.

La codificación intratextual

1. Comprende, por una parte, la referida a la estructura del texto (en la parte que sea conveniente reflejar): división del texto en capítulos, actos, escenas, cuadros, intervenciones de hablantes, etc.
2. Por otra, todo lo que se centra en aquellos elementos del texto que, de una u otra forma, implican ciertas convenciones: desarrollo de abreviaturas, distintas manos, citas, erratas, palabras cortadas, solapamientos, etc.

La codificación intratextual

1. Es importante tener en cuenta que el peso de la tradición escrita puede dar lugar a algunos desajustes que debemos evitar.
2. Por poner un ejemplo claro: una cosa es indicar que en el texto impreso que estamos digitalizando hay una secuencia en letra cursiva y otra señalar que se trata del desarrollo de una abreviatura.
3. En términos más generales, se trata de la diferencia entre codificar el significado y codificar el significante utilizado para transmitirlo en un medio diferente.

La anotación

1. Consiste en la información lingüística que se añade a los elementos de diversos niveles que se encuentran en el texto.
2. La más elemental, imprescindible, es la conocida habitualmente como anotación morfosintáctica.
3. En su forma más habitual, incorpora la indicación del lema y la clase de palabras a que pertenece una forma y los valores que presentan en ese caso las categorías gramaticales que son de aplicación.

La representatividad

1. Es uno de los conceptos fundamentales en la etapa clásica de la LC y uno de los rasgos que siempre se mencionan en la caracterización de los corpus.
2. Sin embargo, es una noción mal definida y muy probablemente inaplicable en este terreno.
3. Se trata de un concepto estadístico que se refiere a la relación que debe existir entre la muestra que se utiliza y la población de la que esa muestra ha sido extraída.
4. No es difícil ver que, para los corpus generales, el concepto es inaplicable por la sencilla razón de que desconocemos las características cuantitativas de la población.

La representatividad en los corpus iniciales

1. La necesidad de que el corpus fuera representativo se planteaba con mucha importancia en los primeros tiempos de la LC por dos razones diferentes:
   - El tamaño de los corpus, muy reducido, obligaba a construir conjuntos formados por fragmentos de textos, buscando la mayor variedad y diversidad posibles.
   - El corpus se construía como un conjunto único, con lo que las respuestas a las consultas venían dadas en bloque, sin posibilidad de diferenciar entre los diferentes tipos de texto que había en su interior.
2. Es evidente que en conjuntos con esas características y aplicaciones de consulta incapaces de hacer recuperación selectiva, la composición tiene una importancia crucial.

La representatividad en los corpus de referencia

1. El aumento de tamaño de los corpus hace que una buena parte de esas características desaparezca.
2. Pero el cambio fundamental viene del hecho de que la codificación introducida en los textos hace posible la recuperación selectiva de información, esto es, la posibilidad de trabajar con subcorpus (o corpus virtuales) y contrastar los resultados obtenidos en el análisis de dos o más de ellos.
3. Lo que tienen que garantizar los corpus de referencia es la existencia, en las proporciones adecuadas, de textos de los diferentes tipos que entran en su diseño. Es decir, el corpus debe estar equilibrado.

El futuro: integración de posibilidades

1. Gracias a las mejoras en la velocidad y capacidad de las máquinas, el refinamiento de la codificación, los avances en PLN y la difusión de la red, estamos empezando a poder trabajar con corpus que integran posibilidades distintas y, hasta ahora, separadas.
2. Los corpus orales, por ejemplo, tienen la posibilidad de recuperar el texto de la transcripción y el sonido, lo cual supone una notable ampliación de sus posibilidades sin que sea necesario complicar la transcripción con rasgos fónicos.
3. Por ese camino, empieza a haber proyectos de integración de transcripciones ortográficas con sonido alineado, vídeo, una capa de texto con información morfosintáctica, otra con el texto analizado sintácticamente, etc.

Antecedentes: el CREA y el CORDE

1. Como es bien sabido, la Real Academia Española decidió en 1995 dar un cambio fuerte en sus sistemas de documentación y emprendió la confección de estos dos corpus.
2. El CREA, cerrado en 2008, consta de algo más de 160 millones de formas ortográficas, procedentes de todos los países hispánicos, editados o producidos entre 1975 y 2004, de los más diferentes géneros y tipos.
3. El CORDE, cerrado también en 2008, consta de unos 250 millones de formas ortográficas incluidas en textos procedentes de todos los países hispánicos desde los orígenes del idioma hasta 1974, de los más diferentes géneros y tipos.

Características deL CREA

1. El CREA tiene la gran virtud de poseer una enorme flexibilidad para la obtención selectiva de datos, basada en una codificación muy ajustada a las características generales del español contemporáneo.
2. Responde, como es lógico, a las líneas habituales de actuación en la época en que fue concebido, lo cual produce, a casi veinte años de su inicio algunos inconvenientes muy notables:
   - Tamaño: las últimas fases (las que tienen mayor volumen) poseen únicamente 7,5 millones de formas anuales.
   - Distribución: 50 % América y 50 % España.
   - Carece de lematización y anotación sintáctica.
   - Tipología textual muy escasa.
   - Presencia baja de algunos géneros textuales y ausencia total de otros que han aparecido con posterioridad.
   - Aplicación de consulta eficiente, pero muy envejecida.

Inicio del CORPES

1. Conscientes de estas deficiencias y también de la importancia crucial de los recursos lingüísticos, las Academias de la lengua española decidieron, en el congreso celebrado en Medellín (Colombia), en marzo de 2007, encomendar a la Real Academia Española la construcción del *Corpus del español del siglo* XXI. Desde entonces, la RAE ha estado trabajando para cumplir este encargo:
   - con el asesoramiento y la colaboración de las demás Academias de la lengua española;
   - con el patrocinio del Banco Santander;
   - con la colaboración de grupos editoriales y autores;
   - con la participación de equipos de codificación pertenecientes a diferentes instituciones.

Características diferenciales del CORPES

1. Tamaño: 25 millones de formas ortográficas por año. Por tanto, al final de su segunda fase (en 2018), tendrá 400 millones de formas correspondientes a textos publicados entre 2001 y 2016.
2. Distribución: 70 % América; 30 % España.
3. Adición de textos de Filipinas y Guinea Ecuatorial.
4. Anotación morfosintáctica y lematización
5. Tipología textual muy enriquecida y ampliada.
6. Aplicación de consulta muy flexible y apta para lograr una auténtica recuperación selectiva de la información.
7. Cálculo dinámico de coapariciones de lemas, con posibilidad de selección por diferentes parámetros.

Características generales del CORPES

1. El CORPES es un corpus de tamaño considerable, que continuará creciendo a un ritmo de 25 millones de formas por año, y será, en sentido estricto, un auténtico corpus de referencia del español del siglo XXI.
2. En su construcción y en su desarrollo se ha intentado combinar la representatividad de los grandes corpus textuales con el enriquecimiento que proporciona la anotación y lematización y la flexibilidad de la recuperación selectiva de la información.

Características generales del CORPES

1. La idea básica consiste en compatibilizar un gran volumen de textos con una codificación refinada, que permita trabajar con subcorpus virtuales, y una aplicación de consulta flexible y potente, adecuada a esas características.
2. Incluirá búsquedas por formas, lemas y características gramaticales (con posibilidad de combinación). También por signos de puntuación.
3. Dará frecuencias totales y normalizadas para cualquier combinación de parámetros utilizada.
4. Incluirá textos de nuevos géneros: blogs, entrevistas digitales, etc.
5. En los textos orales, tendrá posibilidad de búsqueda también por los parámetros empleados habitualmente en sociolingüística.
6. En los que tengan sonido y texto alineados, habrá posibilidad de hacer las búsquedas por texto y recuperar el sonido asociado al fragmento correspondiente.

Desarrollo del CORPES hasta marzo de 2015

1. Siguiendo la misma línea adoptada con CREA y CORDE en 1998, en diciembre de 2013 se abrió al público una versión provisional (la 0.6).
2. A comienzos de julio de 2014 se abrió la versión 0.7, que añade un buen número de textos (ahora unos 185 millones de formas) e incorpora algunas mejoras en el sistema de consultas.
3. En marzo de 2015 aparecerá la versión 0.8., con más textos (unos 207 millones de formas), algunos de ellos con sonido alineado, e importantes elementos adicionales en el sistema de consultas.