Lingüística de corpus e investigación lingüística

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

Universidade de Santiago de Compostela

Máster de Lengua española

Universidad Autónoma de Madrid

27 de noviembre de 2014

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.

La división de las ciencias

1. La primera gran división en las disciplinas científicas es la que se da entre ciencias formales (como las matemáticas y la lógica) y las ciencias empíricas.
2. Dentro de las ciencias empíricas, existe una diferencia crucial, que es la que se da entre las ciencias empíricas naturales y las ciencias empíricas culturales.
3. La estructuración anterior deja claro que la diferencia entre ambos tipos procede del distinto tipo de hechos con que han de enfrentarse: en los objetos culturales (los comportamientos individuales, los movimientos sociales, las lenguas, etc.) no puede esperarse, por su propia naturaleza, la regularidad que permite, en cambio, el establecimiento de las llamadas leyes naturales.

El método hipotético-deductivo

1. Las ciencias empíricas actúan mediante el método hipotético-deductivo:
   - formulación de la hipótesis;
   - deducción de sus consecuencias;
   - contraste de las deducciones con las observaciones.
2. El físico Richard Feynman [1918-1988] compara el trabajo científico a lo que sucedería en el caso de que pudiéramos observar, desde una esquina del tablero, la forma en que unos seres sobrenaturales juegan al ajedrez y tuviéramos que descubrir las reglas del juego a partir de la observación de los movimientos de los jugadores.

Tipos de datos

1. Los datos utilizados en la investigación lingüística pueden ser intuitivos o no-intuitivos. Los primeros proceden de la introspección o bien de los juicios (también introspectivos) formulados por otros hablantes.
2. Los no-intuitivos consisten en los actos lingüísticos realizados, lo que los hablantes dicen o escriben.
3. Pueden ser
   - fragmentarios o parciales;
   - extraídos de corpus textuales.

Tipos de datos

1. Los datos parciales son los utilizados típicamente por la lingüística descriptiva tradicional, que trabaja habitualmente mediante la recogida selectiva de casos.
2. La LC adopta, en cambio, el método de la *explicabilidad total (total accountability)*, para usar la expresión popularizada por Labov y utilizada luego por, entre muchos otros, Leech y Quirk.
3. Evidentemente, la diferencia entre estas dos opciones no está en el tipo de datos, sino en la metodología de la recogida.

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. Sin embargo, parece claro que eso no es así. Algunos autores, como Tognini-Bonelli, han insistido en las diferencias:
   - un texto se lee línea a línea; es *parole*;
   - un corpus se lee en conjunto, en bloque; es *langue*.
2. Podemos añadir una caracterización más, basada en la conocida distinción de Halliday: el texto es el tiempo, mientras que el corpus es el clima.

Corpus frente a textos

1. En realidad, los rasgos anteriores son consecuencia de una diferencia más básica: un corpus está basado en un determinado diseño.
2. El diseño de un corpus, consecuencia de los objetivos perseguidos con su construcción, es el fondo contra el que se proyecta su perfil.

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.

Revoluciones conceptuales y revoluciones instrumentales

1. Se trata de una importante distinción, cuyo ámbito es el de la totalidad de la actividad científica, establecida por Freeman Dyson [1923-].
2. Las revoluciones conceptuales son, en una caracterización rápida, las que se destacan en una aproximación a la historia de la ciencia como la que parte de los trabajos de Thomas Kuhn [1922-1996].

Las revoluciones conceptuales

1. Ejemplos típicos de cambios de paradigma:
   - de la concepción geocéntrica a la heliocéntrica;
   - de la física newtoniana a la einsteniana primero y a la cuántica después;
   - para los partidarios de esta orientación: de la lingüística anterior a la de orientación chomskyana.
2. Dyson las ha llamado **revoluciones conceptuales** (*concept-driven revolutions*) y considera que su efecto más visible es 'explicar cosas antiguas de nuevas maneras' (Dyson, 1997: 50).
3. Son muy llamativas y tienen gran repercusión en la historia de las disciplinas científicas, pero son muy poco frecuentes.

Las revoluciones instrumentales [1]

1. Frente a las anteriores, las **revoluciones instrumentales** (*tool-driven revolutions*) surgen con la aparición de un nuevo instrumento o una nueva herramienta de análisis con la que se puede acceder a zonas que hasta entonces estaban ocultas o resultaban inaccesibles.
2. El ejemplo típico de revolución instrumental es la producida por la aparición del telescopio. La primera noche en que Galileo enfocó la Luna y Júpiter con el telescopio rudimentario que él mismo había construido, pudo ver algo que los seres humanos no habían podido contemplar hasta ese preciso momento.

Las revoluciones instrumentales [2]

3 Nuevamente en palabras de Dyson (1997: 50):

> *El efecto de una revolución impulsada por herramientas es descubrir cosas nuevas que tienen que ser explicadas.*

La difusión de las computadoras en lingüística

1. Supone una auténtica revolución instrumental, que ha cambiado nuestra disciplina en mucho más de lo que implica la lingüística de corpus.
2. Por un lado, la lingüística computacional y todas sus derivaciones han supuesto una reorganización total de buena parte de las disciplinas lingüísticas.
3. En el caso de la lingüística de corpus, las computadoras permiten recuperar con rapidez y comodidad los casos que nos interesan de un subcorpus construido de forma dinámica en un conjunto de formas constituido por cientos o miles de millones de formas.

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 anotación

1. Anotar morfosintácticamente un texto supone introducir en un programa informático el conocimiento de hablantes y lingüistas acerca del texto y las formas que lo integran. Se trata, evidentemente, de un trabajo muy complicado.
2. Las dos fuentes más importantes de problemas para el análisis automático en una lengua con las características del español son las siguientes:
   - En primer lugar, las derivadas de la discrepancia entre el sistema gráfico y la estructura léxico gramatical (contracciones, formas con elementos enclíticos, locuciones, etc.).
   - En segundo lugar, las homografías.
3. En una capa superior, los problemas propios de la tarea que hay que realizar, que supone la aplicación automática de un conocimiento gramatical con importantes diferencias entre los expertos.

Los textos

1. Como es lógico, todos estos problemas se incrementan en la medida en que el corpus esté abierto a la variabilidad diacrónica o diatópica, que complican por un lado la identificación de los elementos y por otro la consideración unitaria de sus variantes
2. Por otro lado, hay que tener en cuenta todos los factores que pesan sobre la edición de textos, que no se limitan a los procedentes de épocas relativamente lejanas.

¿Nueva teoría, nueva disciplina, nueva metodología?

1. No es ninguna de las tres cosas. Leech (1992: 106) la ha definido como “a new research enterprise, and in fact a new philosophical approach to the subject”.
2. Tognini-Bonelli la caracteriza a partir de tres rasgos:
   - es una aproximación empírica a la descripción del uso lingüístico;
   - opera en el marco de una visión contextual y funcional del significado;
   - utiliza las nuevas tecnologías.

Lingüística de corpus y lingüística descriptiva tradicional

1. En la conjunción de esos tres factores radica su diferencia con la lingüística descriptiva tradicional.
2. Es cierto que en algunas de sus áreas se ha operado siempre mediante la recogida sistemática de datos externos (lingüística diacrónica, dialectología, sociolingüística, etc.), pero, por las limitaciones de los recursos técnicos existentes, no podía hacerse casi nunca una recogida sistemática y exhaustiva.
3. La lingüística de corpus aspira a la exhaustividad, a la explicabilidad total (*total accountability*), que, tal como propuso Labov para los análisis sociolingüísticos, supone el estudio de todos los casos presentes en el corpus y también de los casos en los que el elemento estudiado está ausente (pero podría estar presente).

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 2006, 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 2006, 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, en su primera fase (de 2001 a 2012), tendrá 300 millones.
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 noviembre de 2014

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 febrero de 2015 aparecerá la versión 0.8., con más textos e importantes elementos adicionales en el sistema de consultas.

Referencias bibliográficas

1. Feymann, Richard P.(1999): *The pleasure of finding things*. Jackson (TN): Perseius, 1999. Trad. esp. de Javier García Sanz: *El placer de descubrir*. Barcelona: Crítica, 2000, págs. 23-25.
2. Dyson, Freeman (1997): *Imagined Worlds*. Harvard University Press, 1997. Cito por la trad. esp. de Joandomènec Ros: *Mundos del futuro*. Barcelona: Crítica, 1998.
3. Leech, Geoffrey (1992): "Corpora and theories of linguistic performance", en Svartvik, Jan (ed.): *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82* (= Trends in Linguistics. Studies and Monographs, 65), Berlin, Mouton - de Gruyter, 1992., 105-147.
4. Tognini-Bonelli, Elena (2001): *Corpus Linguistics at Work*. Amsterdam, John Benjamins, 2001.

Lectura complementaria

1. Rojo, Guillermo: “Hispanic Corpus Linguistics”, en Lacorte, Manel (ed.): *The Routledge Handbook of Hispanic Applied Linguistics*. Nueva York: Routledge, 2014, 371-387.