Informática y Lingüística: Las lenguas en la sociedad del conocimiento

PONENCIAS

Computing and Linguistics: Languages in Knowledge Society

Guillermo Rojo

Resumen

Está muy extendida la idea de que la Informática es imprescindible en las disciplinas consideradas científicas o técnicas, mientras que su papel en las llamadas 'humanidades' va poco más allá del aprovechamiento de las ventajas que proporciona la ofimática, los paquetes estadísticos o la información captable gracias a Internet. Tal consideración es errónea en lo que a la Lingüística se refiere por varias razones que se explicarán a continuación.

La interacción entre Informática y Lingüística se produce a tres niveles. El más elemental se refiere a la comodidad que proporcionan las computadoras para realizar el trabajo, en segundo lugar está la capacidad que proporcionan las máquinas para manejar gran cantidad de datos y el tercer nivel que es el más interesante es la pretensión de que las computadoras interactuen con seres humanos utilizando las lenguas naturales como código de intercambio.

Palabras clave: Informática y Lingüística, corpus, análisis lingüístico, CREA, CORDE, lingüística computacional.

Summary

It is widely known that Computing is essential for scientific and technical subjects whilst the role played in the ones called "Humanities" is little far than to take advantage of office computing, statistical packets or take information available from Internet. Such a thought is wrong concerning Linguistics for the various reasons explained in the paper.

Interaction between Computing and Linguistics takes place in three different levels. The most basic one is related to the convenience provided by computers in order to do de job, in a second level is placed the enormous capacity provided by computers to deal with great quantity of data and the third level -and the most interesting one- is the pretension of make computers interact with human beings using natural languages as interchange code.

Keywords: Computing and Linguistics, corpus, linguistic analysis, CREA, CORDE, Computational Linguistics.

Está muy extendida la erronea idea de que la Informática es imprescindible en las disciplinas consideradas científicas o técnicas pero no en las llamadas 'humanidades'

1.- Introducción

La existencia de recursos informáticos ha cambiado el modo de trabajar de los lingüistas

Permítanme que comience esta exposición haciendo manifestación pública de la doble sensación que me invade en este momento. De una parte, un profundo agradecimiento a los organizadores de estas Jornadas técnicas de RedIRIS, que han brindado a la Real Academia Española la oportunidad de hacerse presente en un foro de altísimo nivel en el que participan destacados especialistas en temas en los que la Academia es muy poco más que beneficiaria de las posibilidades que el trabajo que desarrollan nos brinda a todos. De otra, la satisfacción, que no voy a ocultar, resultante de esa invitación. Estar aquí hoy con todos ustedes supone —al menos, eso quiero creer— el reconocimiento a la labor que la Academia ha venido realizando a lo largo de estos últimos diez años. Para una institución como la que yo represento hoy aquí, obligada a ser tradicional en el mejor sentido del término, el haber sido invitada a pronunciar la sesión inaugural supone un gran honor y trae consigo el compromiso de continuar progresando por este camino en la medida de sus posibilidades.

Mi intención es presentarles en los minutos siguientes una visión forzosamente sumaria de las líneas de trabajo que las Academias de la Lengua Española han venido desarrollando en los últimos años en aquellos aspectos que supongo de su interés. Para situarlas en el marco adecuado, debo dedicar antes cierta atención a algunas cuestiones generales que presentan como factor común la interrelación de Informática y Lingüística o, si lo prefieren, el grado en el que la existencia de recursos informáticos ha cambiado el modo de trabajar de los lingüistas y cuáles son las consecuencias de todo ello.

2. Informática y Lingüística

Como una faceta más de la distancia existente entre las conocidas habitualmente como 'las dos culturas', está muy extendida la idea de que la Informática es imprescindible en las disciplinas consideradas científicas o técnicas, mientras que su papel en las llamadas 'humanidades' va poco más allá del aprovechamiento de las ventajas que proporciona la ofimática, los paquetes estadísticos o la información captable gracias a Internet. Tal consideración es errónea en lo que a la Lingüística se refiere por varias razones.

En primer lugar, el concepto mismo de 'humanidades' resulta bastante impreciso y probablemente no sirve para mucho más que para hacer grandes agrupaciones de especialidades en la organización de, por ejemplo, los estudios universitarios. Piénsese que en esta especie de cajón de sastre entran especialidades tan distantes entre sí como la Ética o la Metafísica y la Arqueología o incluso la Geografía.

La Lingüística es una disciplina empírica (estudia datos observables de forma intersubjetiva), que se diferencia de las que se presentan como ciencias típicas en tanto que su objeto de estudio no es natural, sino cultural

La Lingüística se ocupa del estudio del modo en que funcionan y evolucionan las lenguas desarrolladas históricamente en las sociedades humanas, incluyendo también bajo la idea de 'funcionamiento' todo lo relacionado con las interacciones sociales en las que la lengua juega un papel importante y, además, con una cierta vocación de prolongación hacia todo lo que, en general, se puede entender como estudio de los sistemas de comunicación. Es, por tanto, una disciplina empírica (estudia datos observables de forma intersubjetiva), que se diferencia de las que habitualmente se presentan como ciencias típicas en tanto que su objeto de estudio no es natural, sino cultural. Eso es lo que los especialistas consideran actualmente el campo de trabajo de la Lingüística, pero hay que reconocer que la idea general acerca de esta disciplina es algo distinta y se basa en que la rama más visible de los estudios lingüísticos tradicionales está constituida por poco más que un conjunto de normas acerca de cuál es la forma correcta o más elegante de utilizar una lengua, cómo debe pronunciarse o escribirse una palabra, cuál es su significado correcto, cuáles son los buenos autores, etc. La más visible probablemente, al menos en una consideración superficial, pero no la única y, desde luego, no la realmente técnica.

La Lexicografía ha manejado tradicionalmente grandes conjuntos de datos con los que montar la documentación sobre la que construir el diccionario o la gramática de una lengua viva o ya desaparecida

La otra cara, la que desde hace mucho tiempo ha venido tratando de mostrar el modo en que evolucionan las lenguas, cómo están constituidas y cómo funcionan los elementos que las componen, ha manejado tradicionalmente grandes conjuntos de datos con los que montar la documentación sobre la que construir el diccionario o la gramática de una lengua viva o ya desaparecida. Los ficheros de los lexicógrafos son tan antiguos como la Lexicografía misma y las concordancias, que muchos creen surgidas con la introducción de los ordenadores existen —hechas a mano, naturalmente— desde hace varios siglos. Permítaseme aludir, simplemente para fijar datos, que el gran nombre inicial de la Lingüística informática, el jesuita Roberto Busa, viajó a Estados Unidos y entró en contacto con IBM ya en 1949. Busa estaba haciendo su tesis sobre el concepto de presencia e interioridad en la obra de Tomás de Aquino y, en un cierto momento, se dio cuenta de que la confección de las fichas y el estudio posterior de los sustantivos correspondientes no proporcionaban todo lo que había que tener en cuenta en el autor estudiado. Era necesario procesar también todo lo introducido por la preposición in, lo cual convertía la recogida de datos en algo considerablemente más trabajoso. Lo hizo y reunió más de 10.000 ejemplos, pero llegó a la conclusión de que tenía que hacer algo para evitar que él mismo o cualquier otro estudioso que pretendiera trabajar con la obra de Tomás de Aquino tuviera que pasar de nuevo por la lectura exhaustiva de varios miles de páginas. Alguien del MIT lo encaminó hacia IBM y el resultado fue el *Index Thomisticus*, que empezó siendo un conjunto enorme de fichas perforadas.

De otra parte, aunque habrán oído ustedes miles de veces que los ordenadores fueron creados para el cálculo numérico, quizá incluso con el corolario de que de ahí vienen algunos de los problemas que

ciertamente sufrimos quienes buscamos en ellos otras utilidades, debo recordarles que esa no es toda la verdad. En primer lugar, me permito mencionar —y espero que disculpen mi atrevimiento— que dos de las cinco tareas generales que Allan Turing consideraba que podrían llevar a cabo los computadores que él estaba concibiendo tienen que ver con las lenguas: traducción automática y aprendizaje de lenguas. En segundo término, con algo ya realmente atingente a lo que hoy nos ocupa, puedo decirles que corre entre los lingüistas una visión según la cual las computadoras parecen haber sido pensadas para su uso en Lexicografía. En efecto, es muy sencillo apreciar las enormes ventajas que el formato electrónico produce en los diccionarios. Tengan ustedes en cuenta que la Lexicografía tradicional desarrolló unos procedimientos realmente refinados mediante los cuales se podía aspirar a cumplir el ideal de que cada entrada del diccionario contuviera todo lo que una persona necesita saber para entender y utilizar correctamente una palabra de una lengua determinada. Para ello, los lexicógrafos desarrollaron unas técnicas de codificación de la información que con demasiada frecuencia llevaron a formulaciones muy cerradas, incomprensibles para la mayor parte de las personas que consultan los diccionarios, que habitualmente se quedan con la acepción que les interesa para comprender el texto que quieren entender o producir y poco más. De todas formas, el gran problema de la Lexicografía preinformática radicaba en la necesidad de presentar la información organizada de una forma determinada. El orden alfabético permite localizar con cierta rapidez la palabra en la que estamos interesados si sabemos cuál es (para, por ejemplo, conocer su significado), pero no sirve de mucho si queremos conocer cuáles son las palabras españolas que proceden del árabe, cuáles los verbos transitivos en construcción pronominal, los verbos derivados en *–izar* o cómo se llama la cadena que sirve para colgar los relojes de bolsillo. Para decirlo en pocas palabras, esa dependencia lleva en muchas ocasiones a producir diferentes tipos de diccionarios en función de los objetivos perseguidos. Es evidente que los diccionarios electrónicos han desmontado la Lexicografía tradicional y han hecho desaparecer las fronteras entre los diversos tipos de diccionarios.

El gran problema de la Lexicografía preinformática radicaba en la necesidad de presentar una información muy compleja organizada con relación a un único factor

El trabajo ha sido largo y, en realidad, estamos todavía en ello puesto que hemos necesitado pasar de lo que es la versión electrónica de diccionarios concebidos para ser publicados en papel a auténticos diccionarios electrónicos, esto es, bases de datos lexicográficos que suponen una aproximación muy distinta a la tradicional. Volveré sobre esta cuestión dentro de un rato.

Creo que resulta muy útil considerar la interacción de Informática y Lingüística como algo que se produce a tres niveles distintos. En el más elemental, las computadoras son utilizadas como máquinas que permiten llevar a cabo con mayor comodidad las tareas que antes había que realizar a mano o mediante máquinas menos capaces o menos eficientes. Los procesadores de texto, por ejemplo, para poner el caso más evidente. Por supuesto, han de enfrentarse con todos los problemas derivados de la manipulación de un conjunto muy amplio de caracteres gráficos que, como todos ustedes saben mejor que yo, ha producido y produce todavía ciertas dificultades cuando el conjunto de caracteres manejado no coincide exactamente con el propio del inglés estadounidense, pero que puede quedar superado en muy poco tiempo.

Los diccionarios electrónicos han desmontado la Lexicografía tradicional y han hecho desaparecer las fronteras entre los diversos tipos de diccionarios

En el segundo nivel está la utilización de las computadoras como máquinas que permiten a los lingüistas manejar con comodidad grandes masas de datos. Para citar la zona más evidente, hasta hace unos años los lingüistas tenían que dedicar grandes cantidades de tiempo a almacenar miles de fichas resultantes del despojo de textos realizado con una determinada finalidad, es decir, en busca de casos de fenómenos de un determinado tipo. En lugar de esa tarea, escasamente gratificante y que había que repetir para cada nueva investigación, disponemos ahora de corpus electrónicos, con tamaños de cientos o incluso miles de millones de formas, en los que podemos recuperar con comodidad y rapidez información acerca de si una forma se documenta o no, con qué frecuencia, en qué países, épocas, tipos de textos o áreas temáticas, cuáles pueden ser las acepciones que presenta, etc. Consultar un corpus no equivale a tratar de localizar documentación sobre una palabra a través de algún buscador, puesto que el corpus ha sido construido con una determinada finalidad y permite

recuperar la información de forma selectiva, pero sí es claro que gracias a la existencia de la red todos los investigadores pueden beneficiarse, sin dejar su despacho, de los recursos que un equipo ha puesto a disposición de todos los demás en una página electrónica.

En el tercer nivel, el más interesante sin duda, se pretende ya que las computadoras sean capaces de 'comprender' expresiones formuladas en lenguas naturales o de producirlas; en otras palabras, de interactuar con seres humanos utilizando las lenguas naturales como código de intercambio. El objetivo resulta extraordinariamente difícil de alcanzar por varias razones.

En primer lugar, aunque todos los seres humanos hemos adquirido como mínimo una lengua y lo hemos hecho sin necesidad de pasar por un proceso formal de aprendizaje, la adquisición lingüística es, en realidad, algo extraordinariamente complejo, exclusivo de los seres humanos, y que para muchos autores constituye la mayor hazaña intelectual que realizamos a lo largo de nuestra vida. Como reflejo, pálido, de lo que supone, piénsese en las dificultades que hay que vencer para hacerse con lenguas distintas de la primera o primeras, sobre todo si el proceso tiene lugar en épocas ya avanzadas de la vida. La actividad lingüística es, pues, la más típica y característica de las actividades humanas y presenta, por tanto, dificultades del mismo tipo, pero en grado superior, a todas las que suponen en definitiva convertir la simple información en conocimiento.

El que los primeros modelos de análisis lingüístico fueran muy simples, junto con las características de las computadoras de la época, explica el gran fracaso de los primeros intentos de traducción automática

En segundo lugar, en parte por ese carácter 'natural' al que acabo de aludir y también por factores que tienen que ver con la historia de la disciplina, los modelos de análisis lingüístico habituales en los primeros años de desarrollo de las ciencias de la computación eran extraordinariamente simples, realmente ingenuos. Eso, unido a las características de las computadoras de la época, explica el gran fracaso de los primeros intentos de traducción automática, que puede ser considerada como la piedra de toque de los desarrollos en Lingüística computacional. Nuestro gran problema es, en definitiva, que hemos de formalizar un conocimiento que, por su propia naturaleza, se presta a ello con muchas dificultades, a lo que hemos de añadir la evidente falta de tradición en este punto, derivada sin duda de la adscripción tradicional a los estudios humanísticos.

Por otro lado, una buena parte de lo que utilizamos para cifrar nuestro pensamiento en secuencias lingüísticas y para descifrar el pensamiento de los demás depende no solo de nuestra competencia lingüística, sino de nuestro conocimiento del mundo, oculto en esas expresiones. Como una indicación rápida de lo que está implicado en este punto, piénsese en que el conocimiento necesario para traducir al inglés la secuencia *El abuelo sentó a la niña en sus rodillas* es más amplio —y de otro tipo— que el que se refiere a las normas de concordancia de los posesivos con el poseedor o lo poseído.

Gran parte de lo que utilizamos para cifrar nuestro pensamiento en secuencias lingüísticas y para descifrar el de los demás depende no solo de nuestra competencia lingüística,sino también de nuestro conocimiento del mundo

Por último, aunque no sea este el factor más importante desde el punto de vista teórico, la facultad lingüística humana —única, como el género humano— se materializa en un conjunto de cuyo tamaño da idea la estimación de que hoy se hablan en el mundo unas seis mil lenguas distintas, a lo que hay que añadir el hecho de que todas ellas, al menos todas las que son habladas por un número mínimo de personas, muestran una diferenciación interna que da lugar a un gran número de variedades distintas, entre las que en muchos casos la intercomprensión no es fácil. La aparente claridad de la existencia de lenguas distintas es más el resultado de la conformación cultural propia de cada comunidad humana que la estructuración de la propia realidad con la que han de trabajar los lingüistas. La importancia de este rasgo para lo que aquí nos ocupa viene del hecho, evidente, de que es necesario desarrollar herramientas y recursos lingüísticos para cada lengua. Los programas de traducción automática, para ir al punto más llamativo, trabajan por pares, que, además, son direccionales, de modo que necesitamos un desarrollo para la traducción del alemán al español y otro para traducir del español al alemán. Una parte —solo una parte— de lo desarrollado aquí servirá para el par español-inglés, español-francés, español-portugués, etc., pero en cada par será necesario un módulo propio y también ajustes específicos para los componentes generales.

3. El papel de las Academias

Esa es la primera parte del cuadro en el que tenemos que situarnos. La segunda es la constituida por la propia Academia y el papel que ha venido desempeñando tradicionalmente en nuestra comunidad lingüística. Desde su fundación a comienzos del siglo xviii, siguiendo de cerca el modelo de la Academia francesa, la Real Academia Española se fijó como objetivo fundamental de su actividad la elaboración y edición de los que se consideran habitualmente, en la perspectiva tradicional, los tres grandes códigos de la lengua: la gramática, el diccionario y la ortografía. Siguiendo el plan trazado, en los primeros años de actividad los académicos fundadores publicaron un magnífico diccionario (el conocido habitualmente como *Diccionario de Autoridades*), luego una ortografía y, ya en el último cuarto del siglo, una gramática bastante aceptable para los estándares de la época. Desde entonces, ciertamente con épocas de actividad considerablemente menor, la Academia ha revisado en numerosas ocasiones estas tres obras y, con diferente fortuna, ha tratado de ir adecuándolas a los requisitos técnicos y sociales de cada época. Para decirlo rápidamente, la Ortografía ha sido casi siempre el referente fundamental en todo el mundo hispánico, la Gramática fue durante mucho tiempo texto obligatorio de la enseñanza en España y las sucesivas ediciones del Diccionario han constituido la que sin duda es la columna vertebral de la lexicografía hispánica. A ellas, la Academia añadió algunas otras empresas, entre las que me interesa destacar ahora el segundo intento de elaboración de un diccionario histórico del español, ambicioso proyecto acometido después de la guerra civil y que, por su misma amplitud tuvo que ser suspendido en 1998 para ser reformulado sobre nuevas bases, tarea cuya etapa final la Academia acaba de acometer.

Desde su fundación, la Real Academia Española se fijó como objetivo fundamental la elaboración y edición de los tres grandes códigos de la lengua: la gramática, el diccionario y la ortografía

El papel de institución responsable de la fijación de la forma correcta del español, asumido tradicionalmente por la RAE, ha ido adquiriendo características diferentes con el paso del tiempo. Como resultado final de la actuación de un conjunto de factores que la escasez de tiempo me impide desarrollar aquí, hace ya casi cincuenta años que Dámaso Alonso señaló con toda claridad que el viejo objetivo de 'Limpia, fija y da esplendor' que figura en el emblema de la Academia tendría que ser actualizado en un empeño real, bien fundamentado técnicamente, de garantizar la estabilidad y homogeneidad relativas del español, lengua oficial en una veintena de países y hablada hoy por unos 350 millones de personas.

No es tarea fácil tratar de fijar las opciones lingüísticas preferidas para una lengua tan extendida y con las tendencias evolutivas internas esperables en un complejo lingüístico de estas características. Desarrollar esa labor con cierta esperanza de éxito requiere establecer con toda claridad dos aspectos diferentes, pero relacionados. El primero de ellos, perteneciente a la esfera de la política lingüística, supone eliminar la consideración tradicional, más o menos disimulada, según la cual el español de España es el modelo que deben seguir todos los hablantes de esta lengua. Frente a esta idea —difícil de concretar por otro lado—, hay que reconocer sin reticencias que el español es una lengua policéntrica, una lengua en la que caben varias normas nacionales distintas, ninguna de las cuales es superior a otra. Los diversos modos que presentan el español de Logroño, el de Buenos Aires, el de Caracas, Tegucigalpa, Cartagena de Indias o Las Palmas de Gran Canaria están todos igualmente justificados desde el punto de vista histórico y deben tener la misma consideración no solo para el lingüista profesional, sino también para el hablante hispano de cultura media.

No es tarea fácil tratar de fijar las opciones lingüísticas preferidas para una lengua tan extendida y con las tendencias evolutivas internas esperables en un complejo lingüístico como el español

Trabajar en esa dirección tiene como efecto inmediato una consecuencia técnica: la Academia, las Academias de la lengua española, deben disponer de los materiales en los que sea posible estudiar los usos reales, tal como se dan en este momento, del español en toda su extensión. Solo de esa forma, con el conocimiento de lo que sucede realmente, será posible tomar decisiones con cierta posibilidad de resultar acertadas.

Precisamente en este punto, el que se refiere a la recolección de los datos, es donde se produce la inflexión que da lugar al cambio de rumbo en el modo de trabajar de las Academias. Por supuesto, la

necesidad de reunir grandes cantidades de datos no ha sido ajena a las tareas tradicionales de la Real Academia Española. Por no citar más que un caso bien conocido, los ficheros de papeletas preparados para la redacción del *Diccionario histórico* contienen unos trece millones de fichas léxicas y lexicográficas procedentes de textos del español de todos los tiempos y todos los países. Sin embargo, esta forma de proceder no se aplicaba, en cambio, en otras tareas académicas, más basadas en la intuición lingüística y los conocimientos personales de un grupo reducido de personas.

En 1995, la Academia española toma la decisión de construir un gran corpus del español actual en el que los técnicos primero y las comisiones académicas después puedan encontrar todo lo necesario para mantener al día el DRAE, añadiendo palabras y acepciones que se van introduciendo, aceptando las decisiones de la comunidad lingüística en los casos de duda, etc. Entra así la Academia en uno de los marcos metodológicos que mejor caracterizan la Lingüística contemporánea: la llamada 'Lingüística de corpus'.

Un corpus es un conjunto de textos, producidos en condiciones naturales, y almacenados en formato electrónico con el propósito de obtener de él toda la información relevante para el estudio de los fenómenos lingüísticos más variados. Como el concepto de 'corpus' se ha utilizado también en la lingüística tradicional (y en la historia, el derecho, etc.), me interesa hacer notar que la alusión a que los textos han de estar almacenados en formato electrónico no es una simple guinda tecnológica que, para presumir de avanzados, ponemos en el pastel de siempre. Por el contrario, el formato electrónico es la única forma en que los lingüistas podemos recuperar de forma cómoda y rápida todo lo que necesitamos de conjuntos textuales que comenzaron con tamaños de un millón de formas y están ahora situados, en general, en cientos de millones.

Fuimos los primeros en optar por el modelo de explotación que todos los demás han seguido después: corpus en red y consulta a través de cualquier navegador no demasiado antiguo, es decir, desde cualquier máquina, con cualquier sistema operativo

El primero de los dos corpus proyectados por la Academia, el CREA (*Corpus de Referencia del Español Actual*), tiene ahora algo más de 150 millones de formas correspondientes a textos de todos los países hispánicos publicados o producidos entre 1975 y 2004. Los buenos resultados alcanzados ya en los primeros meses de este proyecto animaron a la Academia española a llevar a cabo algo parecido para los estudios históricos. Surgió así el CORDE (*Corpus Diacrónico del Español*), que, tras la ampliación derivada de su conversión en la fuente básica de materiales para la nueva versión del *Diccionario Histórico*, consta actualmente de algo más de trescientos millones de formas procedentes de textos desde los orígenes del idioma hasta 1974. En total, pues, más de cuatrocientos cincuenta millones de formas.

Estos dos conjuntos textuales han estado a disposición de todos los interesados desde noviembre de 1998, momento en el que la Academia, en una decisión realmente sorprendente por su valentía y generosidad (y puedo calificarla así porque yo no era académico en aquel momento) decidió ponerlos en la red. Dado que estamos en unas jornadas técnicas organizadas por RedIRIS, me interesa especialmente destacar que, si bien llegamos a la construcción de corpus con cierto retraso, fuimos los primeros en optar por el modelo de explotación que todos los demás han seguido después: corpus en red y consulta a través de cualquier navegador no demasiado antiguo, es decir, desde cualquier máquina, con cualquier sistema operativo. Ahí están desde entonces, con un más que notable número de sesiones diarias, en su mayor parte de lingüistas profesionales que los utilizan para su trabajo.

Un corpus textual no es simplemente una colección de textos en versión electrónica, sino algo bastante más elaborado

Un corpus textual no es simplemente una colección de textos en versión electrónica, sino algo bastante más elaborado, equilibrado según los fines perseguidos, que puede añadir información gramatical, semántica y pragmática sobre las unidades presentes en los textos y que, como mínimo, incorpora un sistema de codificación que nos permite recuperar la información que buscamos de forma selectiva. Eso es lo que da a los corpus el valor técnico que poseen y que hacen que los prefiramos a la simple consulta de todo lo que hay en la red mediante alguno de los buscadores masivos existentes. Por poner una ilustración rápida, si consultan ustedes el CREA podrán recuperar,

por ejemplo, todos los casos de una cierta forma o expresión aparecidos en textos pertenecientes a prensa venezolana, publicados entre 1995 y 2003 y pertenecientes a, por ejemplo, el área de comercio y finanzas.

Este enorme conjunto textual, que añade alrededor de siete millones y medio de formas para cada nuevo año transcurrido, se ha convertido en la fuente de datos de la que viven todas las obras académicas y una parte considerable de las investigaciones lingüísticas sobre el español. La última edición del DRAE, publicada en 2001, se benefició ya de la ventaja que tiene el poder disponer de esos materiales. Desde entonces, el *Diccionario del Estudiante*, que acaba de aparecer, el *Diccionario Panhispánico de Dudas*, que estará en las librerías dentro de unos días, el nuevo *Diccionario esencial*, de publicación prevista para finales del año próximo, la *Nueva Gramática Académica de la Lengua Española*... En fin, todos los proyectos desarrollados por la Asociación de Academias de la Lengua y los que vendrán en el futuro inmediato, como el *Nuevo Diccionario Histórico de la Lengua Española*, utilizan ampliamente los datos proporcionados por el CREA y el CORDE.

Los corpus, como recordarán de lo visto hace un rato, pertenecen al segundo de los niveles de relación entre Lingüística e Informática. También aquí podemos situar la conversión del viejo DRAE, de casi 300 años de antigüedad, en una base de datos lexicográficos de naturaleza electrónica. Tal cambio ha supuesto notables modificaciones en la concepción y desarrollo del diccionario, lo cual tiene especial significación en este caso, ya que el hecho de que sea resultado del acuerdo de todas las Academias de la Lengua lo hace especialmente complicado en su aspecto administrativo. En una dimensión más evidente y no menos importante, el formato electrónico y las posibilidades de la red mundial nos han permitido tomar la decisión de poner en la página electrónica la última edición del DRAE. Es un servicio que tiene una enorme cantidad de consultas diarias y que nos ha enseñado algunas cosas acerca de la transformación que supone el cambio de soporte. En primer lugar, y es una interesante novedad desde el punto de vista técnico, el formato electrónico nos ha permitido dar una solución elegante al problema de la distancia entre ediciones sucesivas del DRAE. Dado su carácter normativo, el DRAE tiene que sufrir continuamente la tensión producida entre el deseo de que esté siempre al día y la imposibilidad de proceder a ediciones cada poco tiempo. Se nos ocurrió entonces la posibilidad intermedia, consistente en ir incorporando a la versión que tenemos en la página electrónica el conjunto de modificaciones (supresiones, enmiendas y adiciones) que las Academias van aprobando. En este momento tenemos ya unas 12 000 modificaciones producidas con posterioridad a la publicación de la edición de 2001. Es una forma realmente interesante de mantener diferenciados el documento básico (la edición publicada en papel) y las modificaciones que se le van introduciendo, con lo que la presión para proceder cuanto antes a una nueva edición desaparece o, al menos, se ve muy reducida.

Se ha realizado la conversión del viejo DRAE, de casi 300 años de antigüedad, en una base de datos lexicográficos de naturaleza electrónica

El tercer nivel, propio ya de lo que se conoce como Lingüística computacional, es, sin duda, el que realmente resulta decisivo para el papel que una lengua pueda jugar en la llamada sociedad del conocimiento. La red, ustedes lo saben mejor que nadie, está repleta de información. Miles de millones de documentos en formato electrónico en cientos de lenguas. Pero la información necesita ser filtrada y elaborada para que se convierta en auténtico conocimiento. Y para avanzar en este proceso es necesario, sin duda, que los procedimientos que utilizamos para la recuperación tengan mayor capacidad lingüística, es decir, capten e interpreten de modo mucho más inteligente el contenido de los textos, que ha sido codificado en alguna lengua natural. En otras palabras, que "entiendan" las lenguas humanas.

El formato electrónico y las posibilidades de la red mundial nos han permitido tomar la decisión de poner en la página electrónica la última edición del DRAE con la inclusión de las modificaciones aprobadas con posterioridad a su publicación en papel

El proceso necesario para ello pasa por incorporar a los textos, a los corpus, la mayor cantidad posible de conocimiento lingüístico. La parte más sencilla consiste en añadir a cada una de las formas ortográficas que forman un texto la información correspondiente a lo que antes, en la escuela, se llamaba 'análisis morfológico' (es decir, la consideración de que *llegábamos* es primera persona de

plural del pretérito imperfecto de indicativo del verbo *llegar*). Esta tarea, que, como he dicho, es la más sencilla, está plagada de dificultades originadas por muy diversos factores. Para citar solo los más evidentes, de una parte tenemos las discrepancias entre la organización léxica y gramatical de un texto y la organización ortográfica. Por ejemplo, una palabra gráfica como *del* debe remitir a dos palabras gramaticales, la preposición *de* y el artículo *el*. Aunque no es estrictamente necesario hacerlo en los primeros niveles, está claro que tiene ventajas considerar que una expresión como *a causa de*, formada por tres palabras gráficas, es un único elemento lingüístico en español actual. Por fin, ya con implicaciones en otros terrenos, tenemos que ser capaces de analizar una forma gráfica única como *decírselo* en tres palabras gramaticales diferentes —una forma verbal y dos pronombres personales—, con la complicación adicional, no precisamente menor, de que la forma canónica del infinitivo *decir* no lleva tilde y, mucho más difícil de solucionar, que la forma se que aparece enclítica puede remitir a tres formas distintas: *le, les, se (cf. decir algo a él/ella, decir algo a ellos/ellas, decirse algo a uno mismo)*.

El alto número de homografías que existen en el español y que no causan problema de comprensión a los hablantes, presenta en cambio fuertes dificultades para los análisis automáticos

Por otra parte, el alto número de homografías que existen en una lengua como el español, que habitualmente no causan ningún problema de comprensión a los hablantes, presenta en cambio fuertes dificultades para los análisis automáticos. Saber cuándo *casa* es sustantivo o una forma del verbo *casar* requiere una buena cantidad de conocimiento gramatical que, insisto en ello, el hablante añade sin esfuerzo, pero que el programa de desambiguación debe obtener, por vía estadística o mediante reglas, del contexto inmediato en que se encuentra esa forma.

En esta situación, que, como he indicado, no es más que el comienzo de un tan largo como apasionante camino, nos encontramos en la actualidad. Hemos anotado ya varias veces, en cada ocasión con una tasa más baja de error, todos los textos que forman el CREA. Estamos ahora mismo preparando las aplicaciones necesarias para poder poner esos materiales en la página electrónica, tarea complicada por el enorme volumen que suponen. Esperamos que a comienzos de 2006 los investigadores del español y también todas las personas interesadas puedan disponer de esta poderosa herramienta de búsqueda léxica y gramatical. Con ello, la Academia española habrá dado un nuevo paso y habrá consolidado su posición en el lugar que debe ocupar: el órgano encargado de coordinar la normativa lingüística en el mundo hispánico y también un centro de referencia en el que los investigadores puedan encontrar y utilizar los materiales que necesitan para desarrollar su trabajo. Con esa labor conjunta, entre todos podremos conseguir que el español disponga de las herramientas y recursos necesarios para que pueda ser plenamente integrada en esa sociedad del conocimiento hacia la que caminamos.

Guillermo Rojo
(grojo@rae.es)
Secretario de la
Real Academia Española