# Los avances tecnológicos y la ciencia del lenguaje

Mª Antònia MARTÍ
Raquel G. ALHAMA
Marta RECASENS
Universitat de Barcelona
CLiC – Centre de Llenguatge i Computació

## 1. INTRODUCCIÓN

La ciencia moderna nace de la conjunción entre postulados teóricos y el desarrollo de una infraestructura tecnológica que permite observar los hechos de manera adecuada, realizar experimentos y verificar las hipótesis. Desde Galileo, ciencia y tecnología han avanzado conjuntamente. En el mundo occidental, la ciencia ha evolucionado desde propuestas puramente especulativas (basadas en postulados apriorísticos) hasta el uso de métodos experimentales y estadísticos para explicar mejor nuestras observaciones.

La tecnología se hermana con la ciencia facilitando al investigador una aproximación adecuada a los hechos que pretende explicar. Así, Galileo, para observar los cuerpos celestes, mejoró el utillaje óptico, lo que le permitió un acercamiento más preciso al objeto de estudio y, en consecuencia, unos fundamentos más sólidos para su propuesta teórica. De modo similar, actualmente el desarrollo tecnológico digital ha posibilitado la extracción masiva de datos y el análisis estadístico de éstos para verificar las hipótesis de partida: la lingüística no ha podido dar el paso desde la pura especulación hacia el análisis estadístico de los hechos hasta la aparición de las tecnologías digitales.

Desde la lingüística, el análisis de los datos tiene una larga tradición, que se inicia durante el siglo XIX con los estudios historicistas sobre el cambio y evolución de las lenguas y con los estudios de tipología lingüística. Con todo, estos estudios de base empírica tenían serias limitaciones en lo que se refiere a la recopilación, almacenamiento y análisis de los datos. No existía una tecnología adecuada para el estudio del lenguaje humano. Más adelante, ya en el siglo XX, el distribucionalismo americano, con figuras de relieve como Bloomfield (1933) y, muy especialmente, Z. Harris (1954), propone un método de análisis, el método distribucional, que parte de la observación y cuantificación en términos de frecuencia de las producciones lingüísticas con el objetivo de desvelar su estructura interna. Sin embargo, este método adolece del mismo tipo de limitaciones que los estudios del s. XIX: la imposibilidad de acceder a recopilaciones de datos lingüísticos cuantitativa y cualitativamente significativas.

Ante la imposibilidad de observar las producciones lingüísticas de manera sistemática en toda su variedad y complejidad, Chomsky (1957, 1965) propone una teoría de base deductiva, con la intuición y la introspección como recurso para la validación empírica del modelo. Este planteamiento da lugar, a la larga, a un modelo autocontenido y autorreferencial, con escasa relación con los hechos que se quieren explicar (Sampson 2001, Bybee & Hopper 2001). A su vez, el recurso a la intuición como método de verificación plantea serios problemas: no tiene en cuenta la variación entre hablantes, los registros, ni los factores contextuales, y su carácter subjetivo no concuerda con los postulados de objetividad que se exigen al conocimiento científico (Wasow & Arnold 2004). Contrariamente a lo que se pre-

*Cum corde et in nova grammatica*, 543-553.

tendía, este modelo no favoreció que la lingüística alcanzara el estatus de ciencia en el sentido *popperiano* del término.

Frente a estas teorías y modelos de análisis basados estrictamente en la estructura del lenguaje, en los años 70 del siglo pasado aparecen propuestas teóricas que estudian la relación de la estructura lingüística con facultades cognitivas de carácter general (p. ej., la categorización y la analogía), el contexto pragmático y social y principios funcionales como la iconicidad. Esta corriente, conocida actualmente como Lingüística Cognitiva, integra diferentes propuestas teóricas y líneas de investigación como la tipología lingüística y los estudios diacrónicos, y se hermana con disciplinas como la psicología del desarrollo iniciada con Piaget. En su conjunto se caracterizan por considerar que la estructura lingüística emerge a partir del uso, que el lenguaje es una función cognitiva sin principios específicos interconectada con los sistemas cognitivos no lingüísticos, que el aprendizaje constituye un factor decisivo en el proceso de adquisición y que el análisis de los datos figura en el centro de la teoría. Se trata, en definitiva, de una visión holística del lenguaje, en la que no se considera la dicotomía sintaxis-semántica, sino que ambos niveles de análisis se hallan íntimamente imbricados.

Así, desde una perspectiva general, a lo largo del siglo XX la Lingüística ha propuesto, por un lado, modelos de base empírica pero para los cuales la tecnología existente no permitía desarrollar el aparato experimental adecuado para captar las características del objeto de estudio y, por otro, modelos deductivos sin una base empírica adecuada y, por tanto, alejados del objeto de estudio.

En este artículo proponemos una reflexión sobre la incidencia de las tecnologías digitales en el desarrollo de los estudios sobre el lenguaje y las interconexiones con disciplinas relacionadas como la neurociencia y la psicología del desarrollo. Los avances en tecnología digital, en el tratamiento de textos tanto orales como escritos, y los avances en el tratamiento de la neuroimagen han definido nuevos métodos de investigación —que incluyen métodos estadísticos aplicados a la lingüística—, y han permitido proponer modelos teóricos del lenguaje acordes con las últimas formulaciones sobre el comportamiento y el funcionamiento del cerebro.

## 2. LAS NUEVAS TECNOLOGÍAS Y EL ESTUDIO DE LAS LENGUAS

Con la aparición de los ordenadores personales y la difusión de Internet como plataforma de comunicación e interacción se han modificado los comportamientos individuales y sociales que parecían arraigados y casi naturales de la vida humana en sociedad. Los cambios en la transmisión y tratamiento de la información derivados de los avances tecnológicos en telecomunicaciones y en informática han afectado considerablemente las actividades lingüísticas en nuestra sociedad. El desarrollo tecnológico está modificando nuestra manera de entender el mundo y el modo de actuar en él (Badia 2009).

Como resultado de la actividad comunicativa humana que se desarrolla en soporte digital y de la capacidad de almacenamiento de información que significa tanto el uso de los ordenadores personales como de Internet, se ha generado una gran cantidad de textos de origen oral y escrito en soporte electrónico. Este material lingüístico en soporte digital ha propiciado, a su vez, el desarrollo de una infraestructura tecnológica sofisticada orientada a su tratamiento, explotación y estudio (Gries 2010). Parece que por primera vez la lingüís-

tica dispone de su propio "telescopio", es decir, de una tecnología adecuada que nos permite aproximarnos al objeto de estudio sobre una base empírica fiable.

De la capacidad de almacenamiento de la tecnología digital nace el concepto de corpus lingüístico, entendido como una colección de producciones lingüísticas reales, escritas u orales (normalmente transcritas), anotadas o sin anotar, generalmente de gran tamaño, incluyendo textos de distintos registros, provenientes de diversas fuentes.

En este contexto, cabe destacar el *British National Corpus*[1], con más de 100 millones de palabras etiquetadas con información morfológica, y que constituye uno de los corpus más representativos del inglés. Fue también el inglés la primera lengua para la que se desarrollaron corpus anotados sintácticamente, siendo el más destacado el *Penn TreeBank*[2], que ha servido de referencia para la anotación de corpus para otras lenguas. Este mismo corpus, anotado con estructura argumental y papeles temáticos, ha dado lugar al corpus *PropBank*[3]. El español, a su vez, cuenta con el corpus de referencia de la Real Academia Española[4], de más de 200 millones de palabras. Al igual que para el inglés, diversos grupos de investigación en lingüística computacional han recopilado y etiquetado corpus. Cabe destacar el corpus *AnCora*[5], que contiene información sobre el lema, la categoría morfológica, la estructura argumental, los papeles temáticos, la clase semántica verbal, el tipo denotativo de nombres deverbales, los sentidos de *WordNet* para los nombres, las entidades nombradas y las relaciones de correferencia; también para el español, tenemos el corpus *Adesse*[6] estructurado como una base de datos con información sobre la categoría sintáctica, el tipo semántico de los núcleos argumentales y los roles temáticos (animado, concreto, abstracto, …). Todos ellos constituyen una fuente de información de gran valor para los estudios lingüísticos y, a su vez, un banco de pruebas para la verificación de hipótesis.

Han empezado a aparecer formas alternativas de almacenaje de datos lingüísticos que abren un amplio abanico de aplicaciones no sólo en el ámbito de la Lingüística, sino también en el de la sociología, la cultura, etc. El proyecto *Culturomics*, que utiliza la colección de libros de Google Books del inglés, francés, alemán, español, hebreo, ruso y chino, forma asimismo un corpus textual del cual puede consultarse la frecuencia de aparición de términos y n-gramas en un espacio de dos siglos (del 1800 al 2000) mediante la aplicación *Google n-grams*[7].

Otros proyectos han explotado la información contenida en corpus para relacionar unidades léxicas con su uso en contexto. Del proyecto *CPA*[8] (*Corpus Pattern Analysis*) nace el *Pattern Dictionary of English Verbs*[9], un recopilatorio que unifica verbos con los

[1] http://www.natcorp.ox.ac.uk/.
[2] http://www.cis.upenn.edu/~treebank/.
[3] http://verbs.colorado.edu/propbank/.
[4] http://corpus.rae.es/creanet.html.
[5] http://clic.ub.edu/corpus/ancora.
[6] http://adesse.uvigo.es/ADESSE/.
[7] http://books.google.com/ngrams.
[8] http://nlp.fi.muni.cz/projects/cpa/.
[9] http://deb.fi.muni.cz/pdev/.

patrones prototípicos de su uso sintagmático. La base de datos de *StringNet*[10] contiene más de un millón de n-gramas con distintintos niveles de representación (léxico, morfológico y sintáctico), y su portal web permite buscar el uso de un término en su contexto, expresado en estos distintos niveles de abstracción.

El uso de corpus en los estudios sobre las diferentes lenguas del mundo ha permitido a los especialistas en tipología lingüística ampliar el abanico de lenguas a analizar, mejorando el enfoque contrastivo de la disciplina. El proyecto *Valency Classes in the World's Languages*[11] está realizando un estudio contrastivo a gran escala de las clases de valencias en varios idiomas (Haspelmath *et al.* 2005).

La explotación de todas estas grandes recopilaciones de datos requiere de técnicas especializadas para el análisis y la selección de la información. El poder disponer de corpus ha hecho posible que la lingüística, al igual que han hecho las otras ciencias, haya incorporado los modelos estadísticos que caracterizan la metodología científica del siglo XX. El conjunto de comandos de Unix es una herramienta de gran utilizad para el manejo de texto, mientras que lenguajes de programación como el *R*[12] se han creado específicamente para el análisis estadístico de grandes cantidades de datos. En relación a la explotación de corpus mediante técnicas estadísticas cabe destacar el trabajo que realiza Stefan Gries tanto en la investigación sobre técnicas y métodos como en la difusión de los mismos[13].

Este conjunto de recursos y herramientas es un ejemplo representativo de cómo la tecnología y la estadística han cambiado el estudio de la lingüística, haciendo posible el acceso al objeto de estudio y permitiendo un análisis masivo de producciones reales para desvelar nuevo conocimiento anteriormente no detectado.

## 3. EL LENGUAJE EN UN MARCO INTERDISCIPLINAR

Hasta hace pocos años, la lingüística se ha circunscrito al estudio de la estructura interna de la lengua, no por deseo propio sino por no disponer de una infraestructura tecnológica adecuada que le permitiera acceder a muestras representativas de las lenguas tanto en formato escrito como en formato oral y en el contexto de uso. Estas limitaciones están desapareciendo y cada vez parece más evidente la posibilidad de desarrollar los estudios sobre el lenguaje en un marco amplio, en estrecha relación con disciplinas relacionadas. El lenguaje, en tanto que capacidad cognitiva humana, constituye una parte fundamental del comportamiento y no se puede ignorar su base neurocognitiva.

En este apartado presentamos brevemente investigaciones que se llevan a cabo en neurociencia y en psicología del desarrollo directamente relacionadas con el lenguaje. Parece que por primera vez en los estudios lingüísticos, se proponen modelos en consonancia con los que se han desarrollado desde áreas de conocimiento directamente relacionadas, definiendo así un marco realmente interdisciplinar con el que habrá que contar a partir de ahora.

[10] http://nav.stringnet.org/.
[11] http://www.eva.mpg.de/lingua/valency/.
[12] http://www.r-project.org/.
[13] http://www.linguistics.ucsb.edu/faculty/stgries/.

### 3.1. Lenguaje y neurociencia

Como es de esperar, el avance tecnológico está teniendo impacto en otras disciplinas científicas, y una de ellas es la neurociencia. Las nuevas tecnologías de neuroimagen permiten tanto la captación y la transformación en formato digital de los procesos cerebrales como su almacenamiento en grandes bases de datos. Disponer de gran cantidad de imágenes cerebrales permite a esta ciencia aproximarse mejor a su objeto de estudio e inducir conocimiento sobre el funcionamiento del cerebro a partir de gran cantidad de ejemplos, así como probar hipótesis o extender conclusiones derivadas de estudios sobre primates.

En la década de los 60 se empezaron a desarrollar tecnologías de neuroimagen como son la Resonancia Magnética (RM), la Electroencefalografía (EEG) o la Tomografía de Emisión de Positrones (TEP), que, a grandes rasgos, presentan la novedad de captar el funcionamiento del cerebro en seres vivos. Con el tiempo estas tecnologías se han ido refinando para ser cada vez más precisas y menos invasivas dando lugar a la Resonancia Magnética funcional (fMRI) y la Estimulación Magnética Transcraneal (EMS). Para ilustrar la importancia de estas nuevas técnicas, nótese que la hipótesis sobre la existencia de las neuronas espejo en el cerebro humano, de gran relevancia para el origen y adquisición del lenguaje, pudo ser probada gracias a las evidencias obtenidas con fMRI y EMS (Rizzolati & Craighero 2004).

Últimamente han aparecido modelos holísticos sobre el funcionamiento del cerebro como la propuesta de Friston (2010) y el *Memory Prediction Framework* del neurocientífico Jeff Hawkins (Hawkins & Blakeslee 2005). Según este autor, las distintas áreas que se han identificado en el cerebro tienen más puntos en común que divergencias, de modo que resulta plausible la propuesta de un único algoritmo general que ejecuta siempre la misma operación aunque trate con información de distinta naturaleza (visual, auditiva, motora o lingüística), idea que ya había sido postulada por Mountcastle en los años 70 (Mountcastle 1978).

Hawkins postula que el cerebro reconoce las regularidades de aquello que observa y forma una representación en el cerebro que no incluye una imagen exacta, sino los patrones relevantes observados, de modo que se puede acceder nuevamente a la representación mediante analogía para reconocer nuevos objetos observados. Así, reconocemos el rostro de un ser conocido a pesar de que los gestos o los cambios de luz hagan que la imagen que percibe nuestra retina sea distinta en cada momento. La frecuencia de observación (o de ejecución de actividades motoras) juega un papel importante, ya que cuanto más frecuentemente se accede a una representación, más se paquetiza en el cerebro, de tal manera que cada vez accedemos a ella más rápidamente y con menos consciencia de cada una de sus partes. Esto permite la rutinización de actividades complejas y el reconocimiento rápido de elementos con los que interactuamos frecuentemente, así como la agrupación de unidades en una unidad mayor que las engloba (*chunk*).

La abstracción de patrones es otro efecto de la frecuencia de observación: el cerebro humano aprende a reconocer patrones cada vez menos específicos, de modo que las representaciones de objetos que previamente se categorizaron por separado pueden asociarse y reconocerse como análogas gracias a la detección de un patrón común más genérico. La capacidad de abstracción y la naturaleza asociativa del cerebro permiten, por tanto, unir en

una misma representación información de distinta índole como la imagen y el tacto de un objeto, o una forma lingüística y el concepto que representa.

En lo que al lenguaje concierne, del modelo de Hawkins se derivan varias conclusiones que coinciden con los principios de la Lingüística Cognitiva. En primer lugar, el lenguaje no requiere de unas capacidades cognitivas específicas, sino que funciona bajo los mismos principios que el resto de las facultades cognitivas y motoras. En segundo lugar, el conocimiento lingüístico se adquiere gracias a la exposición del sujeto a situaciones comunicativas que le permiten captar regularidades o patrones lingüísticos que, con el tiempo, el sujeto podrá abstraer creando patrones cada vez menos específicos, desde un conjunto de construcciones léxicas a una gramática general. Nótese que esta progresión no implica que las construcciones léxicas adquiridas al principio se desechen, sino que éstas se almacenan con distintos niveles de abstracción, dando lugar a una memoria redundante. Finalmente, esta visión holística que integra información de distinta naturaleza, y la inducción de patrones como mecanismo de percepción y aprendizaje, son congruentes con la hipótesis emergentista sobre la estructura del lenguaje y el concepto de construcción como unidad de representación del conocimiento lingüístico propuestas desde la Gramática Cognitiva (Croft & Cruse 2004). Como se verá a continuación, esta confluencia de modelos sobre el funcionamiento del cerebro y del lenguaje son consistentes con las propuestas de adquisición del lenguaje derivadas de la psicología del desarrollo.

### 3.2. Adquisición del lenguaje

Los seres humanos están dotados de una capacidad innata que les permite adquirir una o más lenguas. La naturaleza de esta capacidad, sin embargo, es objeto de controversia. Existe una amplia gama de propuestas que se sitúan entre dos polos opuestos: las de quienes consideran que la adquisición del lenguaje está genéticamente determinada por representaciones y mecanismos específicos y las de quienes consideran que el conocimiento lingüístico es de manera total o parcial el resultado de procedimientos de aprendizaje de carácter general (Clark & Lappin 2011).

Para los investigadores que se sitúan en el primer extremo, la problemática de la adquisición del lenguaje queda explicada por la dotación genética, dejando en un plano totalmente secundario la exposición al estímulo lingüístico. Un conjunto de principios y parámetros genéticamente heredados delimita los tipos de gramáticas posibles circunscribiendo el proceso de adquisición a la instanciación de los parámetros dentro de un conjunto delimitado de valores. Para los segundos, la adquisición del lenguaje constituye uno de los focos de estudio en el marco de la psicología del desarrollo. Nos centraremos en estos últimos.

Desde la psicología del desarrollo, Tomasello (2001, 2003, 2008) presenta una teoría sobre la adquisición del lenguaje que tiene como base el aprendizaje imitativo y procesos de analogía y abstracción. La teoría se fundamenta en experimentos sobre el comportamiento lingüístico infantil y en modelos del lenguaje procedentes de la Lingüística Cognitiva (Goldberg 1995) y concuerda con los hallazgos sobre las neuronas espejo propuestos por Rizzolati & Craighero (2004).

Según Tomasello, el niño empieza a comprender a partir del primer año de edad que cuando los adultos le hablan están intentando comunicar algo, es decir, tienen la intención de captar su atención hacia una tercera entidad. Sitúa, por tanto, el aprendizaje de la

lengua en el marco más amplio de la comunicación. El apercibimiento de la voluntad comunicativa de los adultos constituye un hito fundamental en el proceso de adquisición del lenguaje, y por ello, Tomasello postula que la unidad básica de análisis que el niño tiene en este momento es la unidad de habla o producción[14], es decir, una secuencia fónica delimitada por un entorno de entonación y una determinada intención en un contexto comunicativo. Por ello, considera que las primeras producciones del niño son segmentos fónicos con los que trata de imitar una producción completa, aunque no tenga éxito en reproducirla con exactitud. Tomasello se refiere a estos primeros intentos con el nombre de *holofrases*.

Según esta propuesta, la exposición continuada a eventos comunicativos permite al niño aprender a descomponer las producciones que percibe en unidades más pequeñas, captando de este modo la recurrencia de algunos componentes de las mismas, y detectando la variación de tipos que se da en determinadas posiciones dentro de la cadena. El resultado son "esquemas" de producción con elementos fijos y variables que aprende a combinar de manera creativa. Estos esquemas pueden ser secuencias léxicas, construcciones vinculadas a un ítem léxico o construcciones totalmente abstractas. El tipo de elementos que pueden aparecer en las posiciones variables de los esquemas se categoriza progresivamente: se empieza por categorías semánticas relacionadas con la intención comunicativa de la producción, y se acaba en categorías abstractas similares a las categorías morfosintácticas tradicionales. El resultado es una gramática cada vez más refinada.

Nótese que esta propuesta presupone disponer de una gran capacidad de memoria para almacenar un elevado número de ejemplos así como de habilidades asociativas para inducir relaciones y niveles de abstracción organizados jerárquicamente entre los mismos. Esta formulación del proceso de adquisición es congruente, por un lado, con los modelos de inteligencia postulados por Hawkins y Friston, y por el otro, con el concepto de construcción como unidad lingüística propuesto desde la Gramática Cognitiva (Langaker 1987, 1991), la Gramática de Construcciones Radical (Croft 2001) y las distintas versiones de la Gramática de Construcciones (Lakoff 1987, Goldberg 1995, Kay & Fillmore 1999).

Las nuevas tecnologías digitales están teniendo un impacto decisivo en la investigación sobre la adquisición del lenguaje. Por un lado, el aparato experimental sobre el que se fundamenta esta investigación aplica la tecnología digital para la grabación y posterior transcripción de las producciones infantiles. A partir de estos datos en soporte electrónico se puede proceder a analizar la estructura interna de dichas producciones (Tomasello 2001) así como a la aplicación de métodos estadísticos para la inferencia de gramáticas infantiles (Bannard *et al.* 2009). Por otro lado, se investiga en el desarrollo de modelos estocásticos, como son las redes neuronales y los modelos bayesianos, para la representación de los procesos de aprendizaje con el objetivo de demostrar la plausibilidad de determinadas propuestas teóricas. Lewis & Ellman (2001) y Clark & Lappin (2011) cuestionan el argumento de la pobreza del estímulo demostrando que los modelos estocásticos que infieren modelos gramaticales simplemente a partir de los datos de entrada, es decir, sin ponderar determinados recorridos en el marco de las hipótesis posibles, son tan viables como los modelos que incorporan conocimiento previo orientado a restringir el número de hipótesis. Los pri-

[14] Del inglés *utterance*.

meros emulan los modelos procedentes de la Gramática Cognitiva y los segundos las propuestas innatistas.

### 3.3. La Lingüística Cognitiva

En este apartado trataremos algunos de los aspectos fundamentales de la Lingüística Cognitiva, comunes a sus diferentes formulaciones, y los pondremos en relación con los postulados sobre el funcionamiento del cerebro y los modelos de adquisición que hemos presentado anteriormente en esta misma sección. Nuestro objetivo es poner de relieve la sintonía de esta familia de modelos lingüísticos con las otras disciplinas y, a su vez, poner de manifiesto los postulados que cohesionan e identifican las diferentes formulaciones de la Lingüística Cognitiva.

En primer lugar, el cognitivismo, contrariamente a los proponentes de la Gramática Universal, considera que el lenguaje se rige por los mismos principios que las otras facultades cognitivas. Esta visión unificada concuerda con las propuestas de Hawkins y Friston, que postulan un único algoritmo cerebral para el procesamiento de todo tipo de información, sea visual, auditiva o motora (Nygren 2011). En segundo lugar, son teorías emergentistas en el sentido de que la estructura lingüística no viene predeterminada, sino que emerge a partir del uso. Las estructuras emergentes, a su vez, son inestables y se manifiestan de manera estocástica, es decir, con una gran variación, siendo la frecuencia un factor determinante tanto para su preservación como para su desaparición. El uso constituye, por tanto, el foco de atención a partir del cual se analizan las producciones lingüísticas, lo que conlleva una ampliación del marco de estudio situando el lenguaje en contexto. En consecuencia, el estudio del lenguaje se realiza a partir de muestras de uso, adoptando de este modo una visión empirista basada en los datos (*usage-based models*), con un especial interés en el estudio del lenguaje en acción, es decir, tomando el acto comunicativo en toda su complejidad, en sintonía con la propuesta desde la psicología del desarrollo sobre adquisición del lenguaje (Diessel 2004).

En tercer lugar, el cognitivismo propone cinco procesos cognitivos básicos que explican la adquisición, uso y cambio de las lenguas: categorización, *chunking*, capacidad de almacenamiento, analogía y capacidad de asociación de información de diferente naturaleza (Bybee 2010). Todos estos procesos coinciden con lo que Hawkins propone como principios básicos del funcionamiento de la inteligencia humana: captación del *input*, paquetización (*chunking*) de la información en unidades de orden superior, almacenamiento de las mismas en memoria, comparación mediante analogía del conocimiento nuevo con el ya disponible y puesta en relación e integración de la información procesada y recibida desde diferentes canales. Así, en el proceso de adquisición del lenguaje el niño asocia las producciones lingüísticas a situaciones comunicativas concretas y desde la Gramática Cognitiva se postula la imbricación de forma y significado en las unidades básicas (construcciones).

En el marco cognitivo, las construcciones[15] —la unión de una forma específica con una función comunicativa o significado que exhibe propiedades generales e idiosincrá-

[15] Otros términos para referirse a las construcciones desde la propia Gramática Cognitiva o desde otras disciplinas relacionadas son *frames*, esquemas, *chunks* o patrones.

ticas— son las unidades básicas de la gramática. Desaparece, por tanto, el concepto de regla y el de léxico como componentes diferenciados de la gramática, y pasa a ocupar su lugar una estructura jerárquica de construcciones de diferentes niveles de abstracción que representa nuestro conocimiento del lenguaje. Este repositorio de construcciones tiene un carácter dinámico, es decir, cambia a lo largo de la vida de los individuos. La naturaleza simbólica de las construcciones gramaticales explica por qué muchos patrones lingüísticos se comportan como prototipo, ya que esto es resultado de la relación entre un esquema y sus instancias. El concepto de construcción se corresponde a la idea de patrón que se propone en el *Memory Prediction Framework*, y con las unidades de aprendizaje de la teoría de adquisición propuesta desde la psicología del desarrollo.

Finalmente, la Gramática Cognitiva propone una concepción gradual del lenguaje basada en el concepto de prototipo, frente al carácter categórico y dicotómico de la lingüística que ha predominado hasta el momento. Así, el lenguaje exhibe todas las características de un sistema probabilístico: las categorías y el concepto de "buena formación" son graduales y en todas partes se manifiestas los efectos de la frecuencia. La probabilidad capta de manera adecuada la noción de gradiencia, situándola en un continuo. Algunos autores (Bod *et al.* 2003) apuntan que este carácter difuso del lenguaje y las lenguas constituye también una característica de la competencia. Esta concepción probabilística encaja con los modelos de cognición de Friston y Hawkins, sobre la interpretación de los datos y su representación en forma de conocimiento y con el modelo de adquisición de Tomasello.

Por último consideramos interesante señalar que en Lingüística Computacional se han abandonado los métodos basados en conocimiento para procesar el lenguaje debido a su poca adecuación empírica y se han adoptado métodos de aprendizaje automático a partir de corpus anotados o sin anotar para derivar procesadores a nivel morfológico, sintáctico y semántico. Se evidencia, por tanto, una coincidencia desde diferentes disciplinas al considerar los datos como fuente primera de información. Ello es así porque por primera vez se dispone de los medios tecnológicos para acceder a ellos y procesarlos de manera adecuada.

## 4. ALGUNOS APUNTES SOBRE LA LINGÜÍSTICA DEL SIGLO XXI

Nuestro objetivo en este artículo ha sido reflexionar sobre la incidencia de las nuevas tecnologías en el desarrollo de los estudios sobre el lenguaje. Por primera vez, desde la lingüística, la neurociencia, la psicología y las ciencias de la computación, se proponen modelos teóricos compatibles entre sí. La tecnología ha permitido que estas ciencias avancen de tal modo que los resultados de sus respectivas investigaciones sean consistentes y complementarios. Probablemente se están sentando las bases de una verdadera interdisciplinaridad.

Gracias a la tecnología digital se dispone de grandes repositorios de información lingüística oral y escrita, lo que nos permite un acercamiento al lenguaje distinto al de la lingüística de siglos anteriores. Se nos abre un amplio abanico de nuevas maneras de trabajar en lingüística: ésta puede tener una base más o menos computacional, estar teóricamente o empíricamente orientada, pero lo que sí es cierto es que la lingüística a partir de ahora tiene que estar guiada (o ratificada) por los datos. Ha llegado el momento de sentar las bases de una metodología rigurosa. Habrá que definir cómo vamos a tratar los datos de que disponemos (Wintner 2009), es decir, cómo vamos a usar nuestro propio "telescopio".

Este escenario permite anticipar algunas de las líneas de investigación de la Lingüística del siglo XXI:

— Al estudio de la estructura interna del lenguaje y las lenguas se le sumará el interés por los aspectos relacionados con la comunicación en su sentido más amplio.

— El predominio del eje sintagmático sobre el paradigmático. La lingüística va a estar fuertemente guiada por los datos, lo que llevará parejo el desarrollo de modelos de base estadística

— Gracias al desarrollo de las tecnologías del habla el estudio de la lengua oral ocupará un lugar predominante abriendo el camino al desarrollo de una gramática de la oralidad.

Probablemente dejen de tener sentido las distinciones entre "lingüística teórica", "lingüística computacional", "lingüística de corpus", etc., ya que todo ello quedará englobado en una Lingüística que metodológicamente va a requerir del uso de la estadística, la computación, el acceso a muestras reales de la lengua y la confrontación con modelos procedentes de otras disciplinas relacionadas.

## REFERENCIAS BIBLIOGRÁFICAS


BADIA, T. (2009): "L'impacte social de les tecnologies de la llengua". *Llengua, Societat i Comunicació* 7, 3-10.

BANNARD, C., E. LIEVEN & M. TOMASELLO (2009): "Modeling children's early grammatical knowledge". *Proceedings of the National Academy of Sciences* 106/41, 17284-17289.

BLOOMFIELD, L. (1933): *Language*. New York: Henry Holt.

BOD, R., J. HAY & S. JANNEDY (2003): *Probabilistic Linguistics.* Cambridge MA: The MIT Press.

BYBEE, J. & P. HOPPER (eds.) (2001): *Frequency and the emergence of Linguistic Structure*. Amsterdam: John Benjamins.

BYBEE, J. (2010): *Language, usage and cognition*. Cambridge: Cambridge University Press

CHOMSKY, N. (1957): *Syntactic structures*. Amsterdam: Walter de Gruyter.

CHOMSKY, N. (1965): *Aspects of the Theory of Syntax*. Cambridge MA: The MIT Press.

CLARK, A. S. & S. LAPPIN (2011): *Linguistic Nativism and the Poverty of the Stimulus*. New York: Wiley-Blackwell.

CROFT, W. (2001): *Radical Construction Grammar: Syntactic Theory in Typological perspective*. Oxford University Press. Oxford, UK.

CROFT, W. & D. A. CRUSE (2004): *Cognitive Linguistics*. Cambridge: Cambridge University Press.

DIESSEL, H. (2004): "A Dynamic Network Model of Grammatical Constructions". En *The acquisition of Complex Sentences*, cap. 2. Cambridge: Cambridge University Press.

FRISTON, K. (2010): "The free-energy principle: a unified brain theory?". *Nature Reviews Neuroscience*, publ. en línea 13/01/2010. <http://swarma.org/thesis/doc/jake_343.pdf>

GOLDBERG, A. (1995): *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

GRIES, S. (2010): "Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics". *The Mental Lexicon* 5/3, 323-346

HARRIS, Z. (1954): "Distributional structure". *Word* 10, 146-162.

HASPELMATH, M., M. S. DRYER, D. GIL, & B. COMRIE (2005): *The World Atlas of Language Structures*. Oxford: Oxford University Press.

HAWKINS, J. & S. BLAKESLEE (2005): *On intelligence: How a new understanding of the brain will lead to the creation of truly intelligent machines.* New York: Henry Holt.

LAKOFF, G. (1987): *Women, fire and dangerous things: what categories reveal about the mind*. Chicago: University of Chicago Press.

LANGAKER, R.W. (1987): *Foundations of Cognitive Grammar*, vol. I. Stanford: Stanford University Press.

LANGAKER, R. W. (1991): *Foundations of Cognitive Grammar*, vol. II. Stanford: Stanford University Press.

KAY, P. & C. J. FILLMORE (1999): "Grammatical Constructions and Linguistic Generalizations: The 'What's X doing Y?'". *Language* 75, 1-33.

LEWIS, J. D. & J. ELMAN (2001): "Learnability and the statistical structure of language: Poverty of stimulus arguments revisited". En B. SKARABELA, S. FISH & A. H.-J. DO (eds.): Pr*oceedings of the 26th Annual Boston University Conference on Language Development*. Boston: Cascadilla Press, vol. 1, 359-370.

MOUNTCASTLE, V. B. (1978): "An Organizing Principle for Cerebral Function: the Unit Model and the Distributed System". En G. M. EDELMAN & V. B MOUNTCASTLE (eds.): *The Mindful Brain*. Cambridge MA: The MIT Press.

NYGREN, TH. I. (2011): "Language Acquisition, Emergentism, and the Brain-changing Norms of Unilateral Interventionism". *TCNJ Journal of Student Scholarship* 13. <http://joss.pages.tcnj.edu/files/2012/04/2011-Nygren.pdf>.

RIZZOLATI, G. & L. CRAIGHERO (2004): "The mirror-neuron system". *Annual Review of Neuroscience* 27, 169-192.

SAMPSON, G. (2001): *Empirical Linguistics*. London: Continuum.

TOMASELLO, M. (2001): First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics* 11-1/2, 61-82.

TOMASELLO, M. (2003): *Constructing a Language. A Usage-based theory of language Acquisition*. Cambridge MA: Harvard University Press.

TOMASELLO, M. (2008): *Origins of Human Communication*. Cambridge MA: The MIT Press.

WASOW, T. & J. ARNOLD (2005): "Intuitions in linguistic argumentation". *Lingua* 115, 1481-1496.

WINTNER, S. (2009): "What science underlies natural language engineering?". *Computational Linguistics* 35, 641-644.