ISSN: 1886-9300

Aproximación a la Lingüística Computacional y sus herramientas para el trabajo con *corpus*

Adela González Fernández
Universidad de Córdoba
adela.gonzalez@uco.es
https://dx.doi.org.10.12795/futhark.2018.i13.02

Fecha de recepción: 18.07.2018
Fecha de aceptación: 2.09.2018

Resumen: Desde que la Lingüística Computacional surgiera a mediados del siglo XX por motivos que poco tenían que ver con la investigación lingüística, esta joven disciplina ha pasado por distintas etapas, en las que ha habido tanto épocas de éxito como de duras críticas. Sin embargo, la aparición de la *World Wide Web* en los años 90 supuso el punto de inflexión definitivo para que esta disciplina contara con los apoyos necesarios. A partir de entonces, se produjo la expansión de la Lingüística Computacional y, con ella, grandes avances en distintas áreas de la Lingüística, entre las que destacan el análisis morfológico y sintáctico, las técnicas de reconocimiento de voz, los sistemas de diálogo o la traducción automática. Naturalmente, la Lingüística de Corpus también se ha beneficiado de estos logros y se ha sumado al progreso científico con herramientas informáticas que no solo trabajan sobre la web como corpus, sino también con técnicas de *big data* y que suponen un gran avance en este tipo de investigación. En este trabajo pretendemos realizar un recorrido por la historia de la lingüística computacional y ofrecer una aproximación a esta disciplina, así como a sus aplicaciones y a las herramientas disponibles hoy en día para el trabajo con corpus lingüísticos.

Palabras clave: Lingüística computacional; Lingüística de corpus; Big data; Web como corpus; Herramientas computacionales.

Approaches to Computational Linguistics and its Tools for Corpus Research

Abstract: Computational Linguistics emerged by the middle of the 20th Century for reasons quite different to those regarding linguistic research. Since then, it has

gone through different periods, enjoying substantial success, but also being subject of criticism. However, the advent of the World Wide Web in the 90's was the biggest turning point for this discipline to obtain the necessary support. In that moment, the expansion of Computational Linguistics took place and, with it, the development of several linguistic areas, such as morphological tagging, parsing, speech recognition techniques, dialogue systems or automatic translation. Logically, Corpus Linguistics has also joined scientific progress with computational tools that work not only with the web as corpus, but also with big data techniques which represent a great step forward in this kind of research. Our aim with this project is to provide an overview of the history of computational linguistics and an approach to this discipline, as well as to its applications and the available tools for the work with linguistic corpora.

Key words: Computational linguistics; Corpus linguistics; Big data; Web as corpus; Computational tools.

Introducción

La afirmación de que el estudio del lenguaje, tanto desde el punto de vista empírico como desde el intuitivo, ha sido una de las actividades más antiguas de la civilización no alberga ningún tipo de duda ni de discusión al respecto. A lo largo de la historia hemos explorado la naturaleza del lenguaje para comprender el papel que este juega en la mente y en la comunicación humanas. Con el paso de los siglos, esta rama de conocimiento ha estado unida a otras áreas con las que ha establecido lazos conceptuales y procedimentales y ha sido en los últimos tiempos cuando la aparición de la Informática ha hecho que la investigación sobre el lenguaje evolucione y explore nuevas metodologías que han añadido otra dimensión a los estudios lingüísticos. De hecho, la mayoría de los autores más representativos de la lingüística de corpus no conciben esta disciplina sin su intrínseca unión con los ordenadores. Esto ha sido posible, entre otras razones, gracias al apoyo de técnicas y herramientas que permiten almacenar ejemplos reales de usos lingüísticos y analizar la información desde nuevas perspectivas. Esta relación entre los lingüistas y los informáticos, condenados a entenderse, es lo que Henry Kučera denomia "la pareja peculiar" (Kučera, 1991: 401).

La introducción de este nuevo enfoque ha contribuido de dos formas básicas en el campo de la Lingüística (Sekhar, 2008):

a) Ha permitido verificar viejas teorías sobre el lenguaje y la conveniencia o no de mantenerlas.

b) Ha ampliado el alcance del uso directo de evidencias e información lingüísticas en los trabajos lingüísticos y en las tecnologías del lenguaje.

Estas dos ideas, aunque quizá algo básicas y sencillas, hablan de una nueva dimensión añadida al campo de la Lingüística, gracias a la aparición y al avance de la tecnología informática, lo que ha dado como resultado el surgimiento de la llamada Lingüística Computacional. Esta Lingüística Computacional se entiende como una de las áreas de la Inteligencia Artificial, cuyo objetivo es el tratamiento del lenguaje como el instrumento fundamental de la comunicación humana, directamente ligado a la cognición.

La Lingüística de Corpus está, por tanto, íntimamente ligada a la Lingüística Computacional, pues aporta grandes cantidades de ejemplos reales del lenguaje almacenados en bases de datos en forma de corpus de manera sistemática. Por otra parte, la Lingüística Computacional también nos ofrece también herramientas sofisticadas que analizan estos corpus para extraer la información lingüística.

La fuerte motivación lingüística y cognitiva que nos ha llevado siempre a investigar la forma de comunicarnos a través del tiempo y del espacio, junto con la motivación técnica que dirige la construcción de sistemas informáticos inteligentes capaces de realizar una interacción lingüística eficiente con los humanos, se han unido para desarrollar sistemas como traducción automática, extracción de información, generación y comprensión del lenguaje, etc. Pero para el diseño y el desarrollo de estos sistemas, necesitamos entender de forma empírica el lenguaje natural y sus características. Aquí es donde los corpus se convierten en indispensables.

Por tanto, a pesar de que los trabajos pioneros en Lingüística de Corpus de hace cincuenta años tuvieron unos inicios difíciles y encontraron muchas trabas para hacerse un hueco en la predominante gramática generativa, poco a poco, esta orientación ha ido ganando adeptos gracias a lo que se presenta como una nueva metodología y un nuevo enfoque ayudados por los ordenadores. Sin olvidar, claro está, que es al lingüista a quien corresponde el último análisis de la información aportada por las máquinas.

Debido a la corta historia de la Lingüística Computacional, no es de extrañar la variedad terminológica a la hora de nombrarla y también la diversidad de definiciones. La literatura al respecto abunda (Meya y Huber, 1986; Moreno Fernández, 1990; Gómez Guinovart, 1998; Johnson y Johnson, 1998; Crystal, 2000; Martí y Castelló, 2000; Domínguez, 2002, Pérez y Moreno, 2009, etc.). Algunos de los otros términos que le han puesto nombre a esta disciplina son: lingüística

informática (Moreno Fernández, 1990), procesamiento del lenguaje natural (Meya y Huber, 1986), procesamiento de datos lingüísticos (Meya y Huber, 1986) o ingeniería lingüística (Meya y Huber, 1986; Gómez Guinovart, 1998; Pérez y Moreno, 2009).

En líneas generales, nos parece acertada la definición que aportan Pérez y Moreno (2009: 68) y que dice así:

> La Lingüística Computacional constituye un campo científico de carácter interdisciplinar, vinculado a la lingüística y a la informática, cuyo fin fundamental es la elaboración de modelos computacionales que reproduzcan distintos aspectos del lenguaje humano y que faciliten el tratamiento informatizado de las lenguas.

Gómez Guinovart (1998), además, la considera una subdisciplina de la Inteligencia Artificial que, a su vez, es una parte de la Informática. El objetivo de la Inteligencia Artificial es la construcción de ordenadores que simulen un comportamiento inteligente. Minsky (1968: 2) la define como "la ciencia de conseguir que las máquinas hagan cosas que requerirían de la inteligencia si fueran hechas por humanos" y, en una línea similar, Kurzweil (1990 en Russell y Norvig, 1995: 3), como: "el arte de crear máquinas que lleven a cabo funciones que requieren la inteligencia cuando están hechas por humanos".

Guillermo Rojo (2006) determina que la interacción entre la Informática y la Lingüística se efectúa en tres niveles distintos:

a) El primero es aquel en el que los ordenadores se utilizan como herramienta que contribuye al desarrollo del trabajo lingüístico –que, hasta la aparición de estos, se llevaba a cabo de forma manual. Es decir, la función de los ordenadores se limita a facilitar y aligerar la tarea del lingüista, ahorrándole tiempo y esfuerzo. Aquí encontramos, por ejemplo, los procesadores de textos.

b) En el siguiente nivel que determina Rojo, los ordenadores ayudan al lingüista a manejar grandes cantidades de datos y a sistematizar el trabajo. La construcción y el manejo de los grandes corpus informatizados se encuentran en este nivel.

Por último, el autor señala un tercer nivel, el más interesante para él, en el que los ordenadores interactúan con el ser humano, comprenden las lenguas naturales y son capaces de reproducirlas. En este último nivel es donde tiene cabida la Lingüística Computacional.

1. Recorrido histórico por la Lingüística Computacional

La Lingüística Computacional encuentra sus orígenes a mediados del siglo XX. La mayoría de los autores coinciden en situar su nacimiento en los últimos años de la Segunda Guerra Mundial (Domínguez, 2002; Pérez y Moreno, 2009; Villayandre, 2010), cuando Estados Unidos y la Unión Soviética comenzaron a trabajar en proyectos cuyo objetivo era la elaboración de programas de traducción entre el inglés y el ruso. Los organismos oficiales, entre los que se encontraban los servicios de inteligencia y las fuerzas armadas, fueron los impulsores de estos proyectos y los principales inversores.

Muy pocos años después, los científicos Alan Turing, a quien se le considera uno de los padres de la informática –y quien fue contratado por el gobierno británico para construir máquinas capaces de descifrar mensajes clave– y Claude Shannon fueron determinantes para la evolución de la Inteligencia Artificial y de la Lingüística Computacional. El primero de ellos (que en su conocido artículo *Intelligent Machinery* (Turing, 1996) ya había comenzado a contemplar la posibilidad de construir máquinas capaces de pensar) fue quien desarrolló la *Teoría de los Autómatas*. Shannon, por su parte, también contribuyó a la construcción de autómatas aplicando una teoría de probabilidad basada en el modelo de Markov en 1948 (Shannon y Weaver, 1998), un modelo estocástico de probabilidad en el que la probabilidad de que ocurra un evento depende del evento anterior. Estos trabajos sentaron las bases de la líneas científicas que surgirían en los años siguientes, donde son protagonistas Chomsky – centrado en el análisis sintáctico–, por un lado, y Minsky, Shannon y Weaver[1], por el otro, más volcados en el ámbito de la inteligencia artificial. Estos dos últimos, matemáticos ambos, acabarían formulando en 1949 la *Teoría de la Información y de la Comunicación*, tan influyente, todavía hoy, en el campo de la Lingüística y también en el de la Informática. (Shannon y Weaver, 1998). Según esta teoría, para que la información que parte de una fuente de comunicación llegue al receptor, este tiene que descodificarla, y esta descodificación se basa en un proceso probabilístico que determina en parte la probabilidad de la siguiente descodificación (Cerny, 2006).

Como apunta Villayandre (2010), la información codificada en forma de símbolos puede ser procesada tanto por los ordenadores –en el caso de que sean símbolos numéricos–, como por la mente humana –en el caso de que no sean numéricos. Fue Shannon quien dio un paso más allá y contempló la posibilidad de que los ordenadores también fueran capaces de descifrar símbolos no numéricos en su artículo *A Chess Playing Machine* (1950), en el que sugería que la flexibilidad de las máquinas las posibilitaba para entender el lenguaje (Villayandre, 2010).

[1] Warren Weaver planteó la posibilidad de aprovechar la velocidad, la capacidad y la flexibilidad lógica que habían adquirido los ordenadores para utilizarlos en la traducción 1949 (Shannon y Weaver, 1998).

Estos años fueron bastante prolíficos, debido al interés que la Inteligencia Artificial comenzó a despertar en los ámbitos científicos y a los avances que vieron la luz durante este tiempo, como el método Bayes de reconocimiento de caracteres, para determinar la autoría de los documentos o las técnicas de reconocimiento de voz, entre otros. A partir de aquí, dada la posibilidad de que los ordenadores simularan el pensamiento humano, surgió el Procesamiento del Lenguaje Natural (PNL), denominación que tardaría tiempo en consensuarse y asentarse, pero que atrajo la atención de las investigaciones y cuya aplicación más importante fue la traducción automática.

El auge que experimentara durante estos años la Inteligencia Artificial y, más concretamente, la traducción automática, sufrió un duro revés en la década de los 60, cuando la escasez de resultados, debida a la complejidad del lenguaje –y, según Villayandre (2010), a la ausencia y la falta de adecuación de las teorías lingüísticas– se apoderó de la situación y dio lugar a un informe negativo de la *National Academy of Science* (Domínguez, 2002). El conocido informe ALPAC (*Automatic Language Processing Advisory Committe,* 1964, en Hutchins, 2003) admitía la falta de logros y daba comienzo a la drástica reducción de la financiación y al cambio de rumbo en las investigaciones en el área de la Inteligencia Artificial.

Mientras tanto, la irrupción de Noam Chomsky en este panorama con *Syntactic Structures* (1957) y *Aspects of a Theory of Language* (1965) resultó, una vez más, determinante. Su naturaleza y su tradición lingüísticas no impidieron que su gramática generativa y su teoría de los lenguajes formales fueran unas de las más influyentes en el campo de la Informática y de la Lingüística Computacional.

En los años que siguieron al informe ALPAC, la Inteligencia Artificial se centró en otras áreas del procesamiento automático del lenguaje, especialmente en la elaboración de corpus, coincidiendo con la aparición del Brown Corpus. Se desarrollaron programas informáticos como *ELIZA*[2], capaces de mantener conversaciones con los usuarios y que demostraron las teorías de Turing (Villayandre, 2010) y también técnicas de recuperación de información. Durante los años 70, empresas como IBM y AT&T Lab Research investigaron sobre las técnicas de tratamiento y reconocimiento del habla que siguieron consolidándose en la década siguiente, donde apareció, entre otros avances, *PROLOG*, un lenguaje de programación que utilizaba la lógica como base creado por Colmerauer en 1970 (Llorens y Castel, 1996-2001). Los esfuerzos durante esta época se centraron en la comprensión y en la simulación de los procesos que subyacen al lenguaje (Meya y Huber, 1986). En la década de los 80, la traducción automática vuelve a resurgir

[2] *ELIZA* fue desarrollado por Joseph Weizenbaum en el MIT, entre 1964 y 1966. En el programa, *ELIZA* emula ser un psicoterapeuta y los usuarios interactúan con él. En el siguiente enlace, se puede dialogar con *ELIZA*: http://www-ai.ijs.si/eliza/eliza.html

con fuerza, al igual que también se recupera la Lingüística de Corpus, y siguen apareciendo nuevos trabajos sobre generación del lenguaje.

El momento más decisivo, en cambio, y más influyente para la Lingüística Computacional –y nos atrevemos a decir que para todas las áreas del conocimiento– llegó en los años 90 con la aparición de la *World Wide Web*. La introducción masiva de la Informática y de Internet en el mundo trajo consigo cambios paradigmáticos en la ciencia y una enorme ampliación de sus posibilidades. A partir de este momento, la expansión del conocimiento y la velocidad a la que la investigación ha avanzado han dado impulso a logros científicos sin precedentes en todos los ámbitos del saber, y la Lingüística Computacional, naturalmente, también se ha beneficiado de ellos. Las traducciones automáticas, los sistemas de diálogo, la recuperación de información, y el almacenamiento y análisis de cantidades masivas de datos han experimentado transformaciones que han tomado nuevas direcciones para las que ya no hay marcha atrás. Incluso han aparecido materias nuevas, fruto de la combinación de ámbitos más tradicionales con las nuevas tecnologías, como la *terminótica* (Cabré, 1993: 359), que se encarga "en general, de las relaciones entre la informática y la terminología; y, en particular, que trata de la aplicación de la informática al trabajo terminológico".

Desde el punto de vista de esta autora, la relación entre la Informática y la Lingüística ha traído consigo nuevas aplicaciones "que se pueden clasificar según el grado de complejidad creciente del tratamiento informático que requieren sus objetivos" (Cabré, 1993: 356). Así, distingue cuatro niveles, que abarcan desde las aplicaciones que no manipulan ni analizan los datos (como los sistemas de tratamiento de textos) hasta los sistemas inteligentes que pretenden realizar tareas propias del ser humano (como pueden ser los sistemas de vaciado automático de términos, los sistemas de traducción automática o los generadores de texto). En la zona intermedia entre estos dos extremos sitúa Cabré las herramientas lingüísticas automatizadas, como los diccionarios automatizados, y en un nivel informático mayor, los sistemas automáticos que manipulan los datos, entre los que se encuentran los analizadores, lematizadores, programas de tratamiento estadístico, etcétera.

2. Áreas de trabajo de la Lingüística Computacional

Cualquier aspecto del lenguaje humano susceptible de ser trabajado con los ordenadores puede despertar el interés de la Lingüística Computacional. La creación de programas informáticos que emulen de la forma más fiel posible el comportamiento humano es la última meta de esta disciplina.

Domínguez (2002) enumera en seis las áreas de trabajo de la Lingüística Computacional:

1. *Tagging* o etiquetamiento morfológico: se trata del etiquetamiento de las palabras de forma aislada que nos da información acerca de su morfología.
2. *Parsing* o análisis sintáctico: consiste en el proceso de análisis de oraciones según su sintaxis. Dentro del campo de la Inteligencia Artificial, se incluyen procedimientos de interpretación semántica. Los algoritmos necesarios para esta tarea siguen dos tipos de procedimientos (Meya y Huber 1986, Grishman, 1991, Domínguez, 2002; Villayandre, 2010): *bottom-up*, si se parte de símbolos para llegar a estructuras más complejas; y *top-down*, que parte oraciones dadas para hacer hipótesis sobre cómo están constituidas. Domínguez hace también alusión al llamado análisis superficial o *shallow parsing*, que analiza algunos componentes de la oración sin llegar a ser exhaustivo.
3. Técnicas de reconocimiento de voz y conversión de texto a voz: las técnicas de reconocimiento de voz se hacen a través de sistemas automáticos (*automatic speech recognition*) que transcriben la voz humana en datos procesables por un ordenador. Mediante un sistema de probabilidades y basándose en la teoría de la comunicación de Shannon, identifican las oraciones con más probabilidades de ser la que parte del transmisor de la señal acústica para decodificarla. El objetivo de la conversión de texto a voz, por otro lado, es generar automáticamente los sonidos que un ser humano produciría al leer un texto.
4. Recuperación inteligente de información o *information retrieval*: este campo incluye todos los sistemas automáticos de obtención y análisis de información para su utilización posterior por los usuarios. Estos sistemas son los que se utilizan en los corpus actuales y en los buscadores de Internet.
5. Sistemas de diálogo y sistemas expertos: consisten en la transmisión de información entre los usuarios y el ordenador, gracias al almacenamiento digital previo del conocimiento de expertos en un área determinada.
6. Traducción automática: no cabe duda de que estos sistemas no han alcanzado la perfección y sus limitaciones provocan que no sean capaces de sustituir a los profesionales, pero han experimentado grandes avances en tres líneas distintas: la traducción palabra por palabra, la traducción por transferencia y la traducción por medio de una interlingua (generalmente, a través del inglés o del esperanto).

Prácticamente las mismas áreas de trabajo son las que menciona Gómez Guinovart (1998), aunque con una clasificación un poco más particular. Él divide en tres los campos fundamentales en los que la Lingüística Computacional encuentra su aplicación y los ordena partiendo del más ligado a la Lingüística para finalizar con el más relacionado con la Informática. Así, establece las siguientes líneas:

1. La informática aplicada a la investigación lingüística: aquí incluye los etiquetados morfológicos y sintácticos.
2. La implementación de teorías lingüísticas: según el autor, esta línea posee tres objetivos: a) la elaboración de teorías o modelos lingüísticos, b) descripción de fenómenos lingüísticos concretos enmarcados en estas teorías y c)comprobación de una forma automatizada de la consistencia de una teoría lingüística o de sus predicciones. En este grupo incluye los sistemas de planificación lingüística o formalismos lingüísticos, diseñados para representar conocimientos lingüísticos, entendidos por los ordenadores y que sirven para la comprobación de las teorías. Para ellos, entre otras, se usa el programa *PROLOG*, que hemos mencionado anteriormente.
3. Las aplicaciones lingüísticas de la informática: esta línea está centrada en el PNL y la comprensión y creación de lenguajes naturales. Aquí tienen cabida por tanto, las tecnologías del habla (reconocimiento del habla y síntesis), la traducción automática (Gómez distingue entre traducción totalmente automática y traducción asistida por ordenador) y la extracción de información.

Villayandre (2010), por último, en su trabajo de tesis sobre Lingüística Computacional, clasifica en tres grandes grupos las aplicaciones que considera más importantes:

1. Traducción automática.
2. Interacción en lenguaje natural, donde incluye las interfaces y los sistemas de diálogo.
3. Recuperación y extracción de información.

Aparte de estas tres, también nombra las herramientas de ayuda a la escritura (como los correctores ortográficos o sintácticos), la creación automática de resúmenes, la extracción de terminología, la indexación automática, la síntesis y el reconocimiento del habla y *data mining* (o minería de datos).

3. Herramientas computacionales para el análisis de corpus

Las herramientas informáticas actuales con las que trabaja la Lingüística de Corpus se mueven fundamentalmente en los ámbitos de la anotación textual y el etiquetado. La Lingüística Computacional ha desarrollado programas de fácil manejo para los investigadores, diseñados para extraer información de los corpus de una forma mucho más rápida y segura que la manual.

En este sentido, existen numerosas herramientas de software que permiten descifrar los corpus en términos de información morfológica (tagging), sintáctica (parsing) y semántica. Sin embargo, el elevado número de programas disponibles no nos garantiza variedad en los sistemas de análisis, pues las funciones que cumplen unos y otros son casi idénticas.

Entre estas funciones, se encuentran el conocido part of speech tagging, o POS, que realiza funciones de etiquetado morfológico, con distintos tipos de estrategias. Por ejemplo, TAGGIT fue el primer etiquetador automático que se aplicó a corpus de gran tamaño, entre ellos, el Brown Corpus. El programa se basaba en reglas adquiridas de forma semiautomática a partir de un diccionario de tres mil entradas más una lista de sufijos (Dipper, 2008).

En los años 80, la Universidad de Lancaster desarrolló otro tipo de anotación, basado en TAGGIT, conocido como CLAWS (the Constituent Likelihood Automatic Word-tagging System) (Garside, 1987). Este programa heredó de TAGGIT aspectos como las entradas del diccionario, los sufijos o las reglas entre palabras. Sin embargo, introdujo un programa de frecuencias estadísticas como novedad que permitió más precisión a la hora de desambiguar aquellas palabras que lo necesitaran (aunque es importante señalar que la desambiguación, todavía hoy, no está del todo conseguida). Corpus como el BNC han sido anotados con este sistema (Leech, Garside y Bryant, 1994).

También existe la posibilidad de lematizar las palabras del corpus, esto es, asignarles su lema para que todas aquellas palabras que compartan el mismo lema puedan ser incluidas en una sola búsqueda.

Pero la cuestión de la anotación tampoco se escapa de la polémica en cuanto a su utilidad a la hora de analizar la información textual contenida en los corpus. El argumento más utilizado en su contra es que la anotación "contamina" la información original y dificulta la visualización de los patrones lingüísticos. Esta visión es apoyada fundamentalmente por los investigadores que defienden el enfoque de corpus-driven linguistics, puesto que ven el corpus como el punto de partida para el análisis. Puesto que su objetivo es observar los patrones lingüísticos contenidos en el corpus, la anotación no aporta ninguna ventaja a la investigación y puede generar "ruido". Así lo expresa Sinclair, para quien la anotación pudo haber sido útil hace cincuenta años, cuando los primeros sistemas operativos y software no eran capaces de procesar texto:

La inclusión de etiquetas en los textos es una tarea arriesgada porque el texto pierde su integridad, independientemente de lo cuidadoso que se sea, porque el texto original no puede recuperarse de manera completamente fiable… En lingüística de corpus no se usan textos pre-etiquetados, sino que se procesa directamente texto original para que luego se puedan observar los patrones de este texto que no está contaminado. (Sinclair, 2004: 191)

Desde su punto de vista, mientras sigamos confiando en las etiquetas, estaremos centrando nuestra atención en los modelos antiguos de investigación que se basaban en corpus pequeños. Sin embargo, los corpus anotados no se ajustan a las necesidades de la sociedad de la información porque no son lo suficientemente "sensibles". Se ha demostrado que no son adecuados para lo que él denomina "textos abiertos" (open texts)[3], que constituyen una parte esencial para entender el lenguaje humano.

Por otro lado, los defensores de la anotación la consideran un paso necesario para probar una teoría lingüística concreta (Anthony, 2013).

En cualquier caso, como hemos señalado unas líneas más arriba, existen diversas herramientas utilizadas para estos propósitos, muy parecidas entre sí y que resultan útiles para ciertas tareas del lingüista.

Para elaborar concordancias y listas de frecuencias con el sistema de KWIC (Key Word In Context), algunas de las más conocidas son: WordSmith Tools (Scott, 2012), MonoConc Pro (Barlow, 2000), AntConc (Anthony, 2012) o The Sketch Engine, desarrollada por Kilgarriff y su equipo en 2004 (Kilgarriff, Rychly, Smrz y Tugwell, 2004).

Por otro lado, en cuanto a los programas más utilizados para la anotación morfológica y sintáctica, también existe una amplia oferta de software que ofrecen a los lingüístas la posibilidad de llevar a cabo estas tareas de forma semiautomática. Por nombrar algunos de ellos, encontramos SALTO, UAM CorpusTool o WordStat, para la anotación morfológica; o CLaRK o NooJ, para la sintáctica. No obstante, casi todos llevan a cabo diversas funciones de anotación y etiquetado, aunque la mayoría de ellos solo trabajan con el inglés.

[3] "Con 'textos abiertos' me refiero a textos sin restricciones, cualquier texto sin restricciones, cualquier texto que conforme una muestra razonable del uso de una lengua particular; de hecho, por la forma en la que la investigación lingüística se ha desarrollado, hay una clara evidencia de que el estudio de texto abierto se suele evitar." (Sinclair, 2004: 186).

4. Nuevas tendencias en el trabajo con corpus

Como ya apuntábamos unas líneas más arriba, en los últimos tiempos lingüística de corpus ha visto cómo la aparición de internet ha dado un vuelco a su perspectiva de estudio y al trabajo relacionado con ella. La aparición del concepto de la *web como corpus* (Kilgarriff, 2001) es una de las consecuencias de esta nueva situación. Numerosos han sido los autores desde entonces que han trabajado siguiendo esta metodología y beneficiándose de las ventajas que aporta al trabajo tradicional de corpus, entre las que se encuentran el tamaño, enumera el tamaño, el amplio espectro que cubre, la constante actualización y la multimodalidad, según Fletcher (2012). Entre estos autores, por nombrar solo algunos, destacan Keller, Lapata y Ourioupina (2002), Rigau, Magnini, Agirre y Carroll (2002), Volk (2001), Villaseñor Pineda, Montes Gómez, Pérez Coutiño y Vaufreydaz (2003), Zheng (2002), Agirre, Ansa, Hovi y Martínez (2000), Varantola (2002) o Baroni, Bernardini, Ferraresi y Zanchetta (2009).

Sin embargo, la inmensidad que está alcanzando la web, unida a la enorme variedad y velocidad de los datos que la componen, es decir, lo que conocemos como *big data* (Chen, Mau y Liu, 2014) está generando la necesidad de nuevos enfoques y herramientas específicas para poder trabajar con esta nueva metodología.

El trabajo de con lingüística de corpus a través de *big data*, no obstante requiere tener presentes una serie de dificultades a la hora de acometerlo. Una de las principales es la ingente cantidad de datos disponibles, que exigen unas especificaciones técnicas muy complejas, así como su recuperación, almacenamiento y gestión. Ello supone que la lingüística debe trabajar conjuntamente con la informática para la elaboración de herramientas específicas que sean capaces de manejar este tipo de información. Por otro lado, debido no solo al volumen de los datos, sino también a su rica variedad (información estructurada y no estructurada), creemos conveniente acotar el concepto de la web como corpus para los estudios lingüísticos, de manera que se establezcan unos límites para la fuente de obtención de los datos que posteriormente conformarán el corpus lingüístico.

En esta línea, existen algunas investigaciones que utilizan *big data*, en general, o la información textual contenida en las redes sociales, en particular, para la elaboración de los corpus y su posterior análisis lingüístico, como González-Fernández (2015 y 2016). Estos trabajos suponen un avance en la relación entre la lingüística de corpus y la informática desde el momento en el que, gracias a herramientas informáticas de última generación, permiten obtener información a la que hasta el momento no era posible acceder, como información en tiempo real o geolocalizada.

En cualquier caso, ya sea mediante este tipo de herramientas, más centradas en las redes sociales, o mediante otro, parece claro que el futuro de la lingüística de corpus y de la lingüística computacional pasa por la aplicación de nuevas técnicas al trabajo con *big data*, gracias al cual no solo mejora la cantidad y representatividad de los datos, sino que supone un gran avance en la disminución del tiempo invertido para la confección de los corpus y de la posibilidad de error en los análisis.

Referencias bibliográficas

Agirre, E., Olatz, A., Hovy, E. & Martínez, E. (2000). Enriching very large ontologies using the WWW. Trabajo presentado en The Ontology Learning Workshop of the European Conference of AI (ECAI), Berlin, Germany.

Anthony, L. (2012). AntConc (Version 3.3.5) [Computer Software]. Tokyo, Japan: Waseda University http://www.antlab.sci.waseda.ac.jp/

Anthony, L. (2013). A critical look at software tools in corpus linguistics. Linguistic Research, 30(2), 141-161.

Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation, 43(3): 209-226.

Barlow, M. (2000). MonoConc Pro, editado por Athelstan [Computer Software]

Cabré, M. T. (1993). La terminología. Teoría, metodología, aplicaciones, traducción de Carles Tebé. Barcelona: Antártida / Empuries.

Cerny, J. (2006). Historia de la lingüística. Cáceres: Universidad de Extremadura.

Chen, M., Mao, S. & Liu, Y. (2014). Big Data: A Survey. Mobile Networks and Applications, 19(2): 171-209. doi: 10.1007/s11036-013-0489-0

Chomsky, N. (1957). Aspects of the theory of syntax. Cambridge, MA: MIT.

Chomsky, N. (1965). Syntactic Structures. The Hague/Paris: Mouton.

Crystal, D. (2000). Diccionario de lingüística y fonética, traducción y adaptación de Xavier Villalba. Barcelona: Octaedro.

Dipper, S. (2008). Theory-driven and corpus-driven computational linguistics, and the use of corpora. En A. Lüdeling, A. & M. Kytö (Eds.), Corpus Linguistics: an International Handbook, v. 1, editado por Anke Lüdeling, A. y Merja Kytö, (pp. 68-97). Berlin: Mouton de Gruyter.

Domínguez Burgos, A. (2002). Lingüística computacional: un esbozo. Boletín de Lingüística 18: 104-109.

Fletcher, W. H. (2012). Corpus Analysis of the World Wide Web. En C. Chapelle (Ed.), Encyclopedia of Applied Linguistics, (pp. 339-347). London: Wiley-Blackwell.

Gómez Guinovart, J. (1998). Fundamentos de Lingüística Computacional: bases teóricas, líneas de investigación y aplicaciones. En X. Baró i Queralt & P. Cid Leal (Eds.), Anuari SOCADI de Documentació i Informació, (pp. 135-146). Barcelona: Societat Catalana de Documentació i Informació.

González-Fernández, A. (2015). Big Data as a Tool for Linguistic Research: Approaches to Trends in Bilingualism in Ten Latin-American Countries. International Journal of Language and Applied Linguistics (IJLAL), special issue "Bilingual Education", 1: 1-12.

González-Fernández, A. (2016). Análisis de las necesidades traductológicas en Europa a través de big data. Skopos, 7: 45-74.

Grishman, R. (1991). Introducción a la lingüística computacional, traducción de Antonio Moreno Sandoval. Madrid: Visor.

Hutchins, J. (2003). Alpac: the (in)famous report. En S. Niremburg, H. Somers & Y. Wilks (Eds.) Readings in Machine Translation, (pp. 131-136). Cambridge: MIT.

Johnson, K. & Johnson, H. (Eds). (1998). Encyclopedic Dictionary of Applied Linguistics: A Handbook for Language Teaching. Oxford: Blackwell.

Keller, F., Lapata, M. & Ourioupina, O. (2002). Using the Web to Overcome Data Sparseness. Proceedings of the Conference on Empirical Methods in Natural Language Processing, (pp. 230-237). Philadelphia: ACM.

Kilgarriff, A. (2001). Web as corpus. Proceedings of the Corpus Linguistics Conference (CL 2001). University Centre for Computer Research on Language Technical Paper, 13, (pp. 342-344), Special Issue, Lancaster University.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. Trabajo presentado en The Eleventh EURALEX International Congress, (pp. 105-115), Lorient, Francia.

Kučera, H. (1992). The odd couple: The linguist and the software engineer. The struggle for high quality computerized language aids. En J., Svartvik (Ed.) Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991, (pp. 401-420). Berlin/New York: Mouton de Gruyter.

Leech, G., Garside, R. & Bryant, M. (1994). CLAWS4: The tagging of the British National Corpus. Trabajo publicado en Proceedings of the 15th International Conference on Computational Linguistics, (pp. 622-628), Kyoto, Japan.

Llorens Largo, F. & Castel de Haro, M. J. (1996-2001). Prácticas de Lógica, Prolog. Universidad de Alicante. http://www.infor.uva.es/~teodoro/PrologAlicante.pdf

Martí Antonín, M. A. & Castelló Masallels, I. (2000). Lingüística computacional. Barcelona: Universitat de Barcelona.

Meya Llopart, M. & Huber, W. (1896). Lingüística computacional. Barcelona: Editorial Teide.

Moreno Fernández, F. (1990). Lingüística informática e informática lingüística. Lingüística Española Actual 12(1): 5-16.

Minsky, M. (1968). Semantic information processing. Cambridge, Mass.: MIT Press.

Pérez Hernández, C. & Moreno Ortiz, A. (2009). Lingüística computacional y lingüística de corpus. Potencialidades para la investigación textual. En N. Rodríguez Ortega (Ed.), Teoría y literatura artística en la sociedad digital: construcción y aplicabilidad de colecciones textuales informatizadas, (pp. 67-96). Gijón: TREA.

Rigau, G., Magnini, B.,Agirre, E. & Carroll, J. (2002). Meaning: A roadmap to knowledge technologies. En Proceedings of COLING Workshop on A Roadmap for Computational Linguistics. Taipei, Taiwan, 13, (pp. 1-7), Stroudsburg, PA: ACM.

Rojo, G. (2006). Informática y Lingüística: Las lenguas en la sociedad del conocimiento. Boletín de RedIRIS 74-75: 1-8.

Russell, S. & Norvig, P. (1995). Artificial intelligence: A Modern Approach. Englewood Cliffs, N.J.: Prentice Hall.

Scott, M. (2012). WordSmith Tools (Version 5.0) [Computer Software] http://www.lexically.net/software/index.htm

Sekhar, N. (2008). Corpus linguistics: An introduction. Nueva Delhi: Pearson Education.

Shannon, C. E. (1950). A Chess-Playing Machine. Scientific American, 182(2): 48-52.

Shannon, C. E. & Weaver, W. (1998). The mathematical theory of communication. Urbana: University of Illinois.

Sinclair, J. (2004). Trust the text: language, corpus and discourse. London/New York: Routledge.

Turing, A. M. (1996). Intellingent Machinery, A Heretical Theory. Philosohia Mathematica, 3(4): 256-260.

Varantola, K. (2002). Disposable corpora as intelligent tools in translation. Cadernos de Traduçao: Corpora e Traduçao, 1(9): 171-189.

Villaseñor Pineda, L., Montes Gómez, M., Pérez Coutiño, M. & Vaufreydaz, D. (2003). A Corpus Balancing Method for Language Model Construction. Trabajo presentado en Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico.

Villayandre Llamazares, M. (2010). "Aproximación a la lingüística computacional". Tesis de doctorado, Universidad de León.

Volk, M. (2001). Exploiting the WWW as a corpus to resolve PP attachment ambiguities. Trabajo presentado en el congreso de Corpus Linguistics 2001, Lancaster, UK.

Zheng, Z. (2002). AnswerBus Question Answering System. Proceedings of the 2nd International Conference on Human Language Technology Research, California, United States, (pp. 399-404). New York, NY: ACM.