Breves glosas en torno de la lingüística computacional*

LUIS FERNANDO NIETO RUIZ*
lufer@uptc.edu.co

Recepción: 01 de abril de 2009
Aprobación: 08 de junio de 2009

_____

* Este artículo pertenece a la línea de investigación Lingüística Teórica y su desarrollo en Colombia, de la Maestría en Lingüística, UPTC.

RESUMEN

La lingüística computacional contribuye en áreas propias del saber lingüísico, como por ejemplo en las gramáticas generativa, cognitiva y la funcional, en la semántica, en la pragmática, en la lingüística textual, en el discurso, en la sociolingüística, mediante diversos programas como: analizadores básicos relacionados con la morfología, el léxico, la sintaxis, recursos que permiten establecer relaciones léxicas entre palabras o relaciones sintácticas entre los constituyentes de una oración.

Palabras clave: corpus, niveles, analistas, buscadores, comandos.

ABSTRACT

Computational linguistics contributes to different areas that belong to linguistic knowledge, for instance generative, cognitive and functional grammar, semantics, pragmatics, text linguistics, discourse analysis and sociolinguistics, by means of software such as: basic analyzers of morphology, lexicon and syntax. These computing devices allow the researcher to establish lexical relationships among words or syntactic relationships among the components of the sentence.

Key words: corpus, labels, analyzers, search engines, drivers.

Vivimos en la era de la información en la "tercera ola". La primera fue la de la revolución agrícola; la segunda fue la industrial y con ella, la migración del campo a la ciudad. La tercera ola es la de la revolución de la información y del conocimiento, el paso del músculo a la inteligencia.

Toffler, A., 1996.

A lo largo de la historia de la humanidad, el hecho lingüístico ha sido uno de los interrogantes que viene despertando día a día un interés constante. Baste el recorrido diciendo que desde los primeros filósofos, hasta los investigadores actuales, dedicados a la inteligencia artificial, así, como filólogos y lingüistas, han tratado de entender la complejidad tan enorme que se evidencia en el lenguaje.

Los estudios sobre el lenguaje y las lenguas han ido progresando, cambiando de orientación, de derrotero, diversificándose, en función de los estadios disímiles del desarrollo científico e intelectual de los objetivos que los propios investigadores se proponían alcanzar y de las necesidades que el desarrollo social exigía.

Lo anterior evidencia que el lenguaje se convierte en un factor decisivo para el desarrollo del ser humano. La importancia del crecimiento, el interés por fomentar el control racional de los actos humanos y la necesidad de buscar acuerdos obligan a prestar atención al fenómeno del lenguaje, máxime cuando en el transcurso de este nuevo milenio ha sido laboriosa la tarea de estudio y avances notables en teoría y aplicaciones en diferentes campos de la ciencia y del arte, pero sobre todo en la técnica.

Es oportuno comentar que en los últimos años se ha despertado un gran interés por el desarrollo de las nuevas tecnologías del lenguaje; dicho con otros términos, se han sembrado los nuevos cimientos de la sociedad de la información del futuro; por tanto, a medida que las interacciones, entre los humanos se hace más frecuente, cualquier paso hacia adelante, será el principio de la construcción de nuevos mundos encauzados en una sociedad de la información. Es aquí, en donde cobran relevancia las interacciones puestas en escena, mediante procesos comunicativos, que tienen como herramienta angular al lenguaje, como generador de entidades artificiales, mediante el uso de medios naturales.

Lo anterior, permite afirmar que se está frente a una empresa común en la que participan comunidades científicas diversas, ciertamente con perspectivas y visiones diferentes que permiten vislumbrar la comunicación entre las máquinas y los seres humanos. Por tanto, se hace latente un enfoque basado en métodos de ingeniería y enmarcado básicamente de cara al desarrollo de diversas aplicaciones, cuyo principal objetivo es el fortalecimiento de sistemas que funcionen y que sean capaces de ofrecer un servicio eficiente, de gran ayuda para los usuarios. Este enfoque se caracteriza por ofrecer estrategias que centren su atención en problemas muy concretos, mediante aplicaciones inmediatas y eficaces.

Los anteriores acápites permiten evidenciar el procesamiento del lenguaje natural desde una perspectiva computacional. Es pertinente aclarar que al hacer alusión a perspectiva computacional, no se hace referencia únicamente a ordenadores, como dispositivos, específicamente diseñados para realizar determinadas tareas, pues existen otros elementos que también son punto de conexiones computacionales; a manera de ejemplo, el cerebro, los ábacos e incluso, la misma mano cargada de lápiz y papel. Una cualidad muy interesante de los procesos computacionales es que se pueden analizar desde una perspectiva que es totalmente independiente del mecanismo que debe efectuarlos. Con ello, no se quiere afirmar que la implementación sea irrelevante; por el contrario, se puede aprender mucho de un programa computacional sin tener en cuenta los detalles relacionados con la ejecución.

En el apartado anterior, se tocó un aspecto importante: el lápiz, el papel y la mano; esto ¿a dónde conduce?, pues ni más ni menos que al proceso escritural. Bien vale la pena citar a Platón, evocado por (Ong, 2001, p. 84), quien consideraba la escritura como una tecnología externa y ajena, lo mismo que muchas personas piensan hoy de la computadora. En la actualidad, la escritura, ya ha sido interiorizada, de una manera muy profunda y se ha hecho, de ella, una parte tan importante en la vida cotidiana del ser humano; tanto así, que al día de hoy, ya no se puede convivir sin ella, aunque parezca difícil considerarla como una tecnología, algo parecido a la imprenta y al computador.

Sin embargo, la escritura constituye una tecnología que necesita de un equipo de trabajo, relacionado con herramientas y otros mecanismos como estilos, pinceles, superficies cuidadosamente, preparados como el papel, las tablas de madera, los tintes, las pinturas, entre muchos. Por tanto, se puede afirmar que la escritura es la transformación del habla oral, la separación de la palabra del presente vivo, es completamente artificial; mas esto, no significa que deba ser condenada; por el contrario, debe ser elogiada, pues tiene un valor inestimable y, de hecho, esencial que permite la realización de diversas aptitudes humanas plenas e interiores. Las tecnologías, no son únicamente, recursos externos, sino

también transformaciones internas de la consciencia y mucho más cuando afectan a la palabra. Dichas alternativas pueden resultar benéficas y estimulantes, pues la escritura debe dar vigor a la conciencia.

La escritura, al tratarse de una tecnología, cumple una función primordial cual es la de almacenar información fijándola en un soporte externo a la memoria. Es natural que la escritura, a partir de su invención, produjera cambios ligados a las prácticas, basadas en el intercambio, la manipulación y el almacenamiento de determinada información. De este modo y, desde la producción del conocimiento, hasta las diferentes transacciones comerciales se ven afectadas por el progresivo traspaso de la oralidad, mas no quiere decir que ésta no sea importante, hacia el dominio de la escritura.

La escritura, entonces, hizo posible la diversidad de registros de acontecimientos y, a partir de esto, las diferentes prácticas, asociadas a estudios escriturales, a exámenes de hechos pasados, a la observación sistemática del mundo, se pudo evidenciar más la realidad del mundo circundante. También, fue posible la revisión de los enunciados que deberían ser comunicados y de los cuales no habían conocimientos epistemológicos pertinentes que pudieran redundar en beneficio de los distintos quehaceres. Todo esto redundó en la construcción del conocimiento científico y en la actividad académica, basada principalmente, en prácticas discursivas como la argumentación y la refutación.

La escritura, como producto tecnológico, acompaña al hombre en su evolución histórica y, por tanto, social; de un modo adecuado, lo moldea y lo transforma profundamente. Pero este fenómeno, también es recíproco; es decir, el ser humano tiene una tarea primordial, no sólo con la escritura, sino con la oralidad. Dicho en otros términos, el ser humano se encuentra frente a una diversidad de textos orales y escritos plasmados en documentos que los contienen, debidamente compilados y organizados; más aún, tienden a formar los llamados corpus.

Un corpus es definido como un conjunto de textos en formato electrónico que se ha construido a partir de una selección, realizada, según unos criterios y objetivos concretos. A manera de ejemplo: el análisis de la obra de un autor; el estudio de la lengua o, bien la observación de unos determinados aspectos de la misma. Los corpus se convierten en una fuente de información confiable y verificable, en cuanto al desarrollo de los recursos tecnológicos y, para el caso, lingüísticos, que se pueden utilizar como bancos de pruebas, para la investigación en lingüística teórica y computacional.

La vinculación del corpus está relacionada con una serie de procesos, a saber: la codificación, etiquetados, análisis lingüísticos y herramientas.

Con el transcurrir de los años y rodando por el mundo, es evidente el deseo de la exploración y la creación de diversidad de corpus, por cuanto evidencian datos reales y exhaustivos que permiten reproducir, con la mayor fidelidad, la variedad de características que posee un referente manifiesto en un corpus, recursos que facilitan la búsqueda de soluciones a algunos problemas tradicionales que, sobre todo, se evidencian en la lingüística computacional.

Con estos cortos acápites, no se pretende hacer un recorrido sobre las diferentes clasificaciones de los corpus, ni mucho menos hablar de sus diseños y criterios específicos; evidenciando, eso sí, la importancia que cada uno de estos apartados merece; pero el enfoque se manifiesta, básicamente, en apreciar las bondades que puede brindar, especialmente, en el ámbito lingüístico, el poder tener acceso a recursos de primera mano que contribuyan a dar soluciones a diversos problemas investigativos.

Uno de los aspectos que más trunca un proceso investigativo es el no tener un horizonte claro, que permita dar pasos seguros y contundentes y, esto se debe a la falta de un derrotero, de un faro, de una brújula o estrella polar, que está repleta de remembranzas marinas, pero sirve de luz y de guía para las naves, de tal manera, que impide que los seres humanos nos perdamos en la bruma de la noche, en las tinieblas, entre la fuerza de las olas y en el sin rumbo de la oscuridad. El corpus brújula, sirve de orientación, de punto de referencia, de signo, de flecha, de ruta por seguir. Son aquellos materiales, ya sean transcripciones, grabaciones o recipientes en los cuales se puede recoger la lengua, tanto escrita, como hablada.

Hoy, Existe mucha tipología de corpus que se establece de acuerdo con su función, su diseño, las características formales o de los métodos utilizados para su construcción. Los aspectos más generales de los diferentes tipos de corpus, de acuerdo con Martí (2003, p. 18) determinan el empleo de hipótesis, sobre los niveles de representatividad, y el tipo de función que quiere dársele al corpus.

Baste también, con decir, que hay diversidad de corpus, dependiendo de los criterios generales o los específicos; entonces, se puede hablar de corpus grandes, equilibrados, piramidales, monitores, paralelos, comparables, multilingües, oportunistas, generales, especializados, genéricos, cronológicos, diacrónicos, entre muchos. Lo importante aquí,

es saber que existe una gama de corpus que permite hacer aplicaciones más adecuadas en el momento de entrar a recolectar información de acuerdo con los objetivos concretos, propuestos para determinado fin.

En términos generales, se distingue entre los corpus en bruto y los corpus anotados. En cuanto a los primeros se refiere, contienen el texto original, sin ningún tipo de información añadida, se evidencia, tal y como lo ha producido el autor. Este tipo de corpus constituye una buena fuente de información para obtener coocurrencias, colocaciones y datos estadísticos sobre ocurrencias absolutas y relativas de una palabra o de las palabras de un texto.

Los corpus anotados, con información morfosintáctica y semántica y los corpus anotados sintácticamente, proporcionan información valiosa sobre la estructura argumental y las restricciones selectivas de nombres, verbos y adjetivos. Constituyen a sí mismo un componente esencial para mejorar los sistemas de etiquetado automático o desambiguadores en el nivel morfosintáctico. Cabe aquí un ejemplo de la importancia que han adquirido, es el papel nuclear que juegan en la evaluación de los sistemas de análisis del lenguaje y de la recuperación de la información. Éstos últimos han impulsado la celebración de las MUC (Message understanding conference), que se caracterizan por desarrollar una tecnología, mediante una metodología de evaluación, basada en corpus de prueba y corpus de contrastación, previamente, seleccionados.

En este apartado es pertinente recalcar que la lingüística, basada en un corpus, es decir, la "lingüística del corpus" evidencia una línea de actividades demasiado importantes en los ámbitos funcionalistas actuales, gracias a su metodología, pues tiene un carácter empírico, ya que permite realizar las investigaciones sobre la base de colecciones extensas de textos naturales, mediante el empleo intensivo de diversos programas computacionales; por consiguiente, se realiza un estudio lingüístico que se destaca por el empleo de las nuevas tecnologías de la información.

La lingüística computacional consta de un componente teórico, de ciencia básica y una parcela más aplicada y tecnológica, que, en la mayoría de los casos, recibe el nombre de *ingeniería lingüística (language engineering)*. El nombre de Lingüística Computacional, hace alusión a una perspectiva más teórica, especialmente a la centrada con la utilización de ordenadores, con el objeto de poder tratar aspectos de las lenguas naturales, que, tradicionalmente, competen a aspectos como: morfología, sintaxis y semántica, niveles propiamente, lingüísticos, tal como se puede evidenciar en la siguiente gráfica:

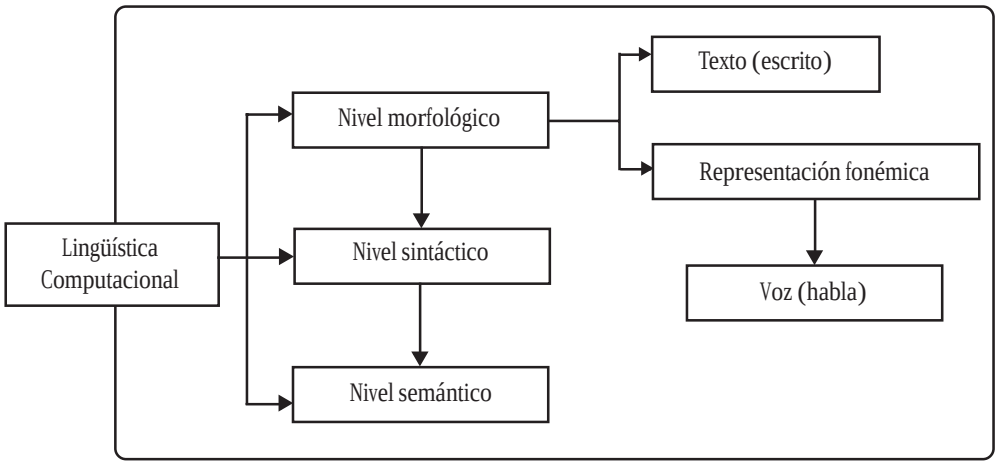

Figura No. 1. Niveles Lingüísticos.

Cada uno de los constituyentes de la figura uno, es un módulo que recibe un tipo de información específica como son los datos de entrada (in Put), su procesamiento, con base en determinados algoritmos y, finalmente, como resultado (out Put), se tiene una producción de una estructura tipológica que pasa a ser la entrada del módulo siguiente.

La lingüística computacional, orientada a estudios textuales, intenta identificar y analizar patrones de uso, es decir, estructuras y rasgos lingüísticos, que pueden ser correlacionados con diversidad de variables lingüísticas, previamente determinadas. Para ello, se debe emplear técnicas de análisis, tanto cuantitativas, como cualitativas. En cuanto a las primeras se refiere, se pueden elaborar programas de concordancia, estadísticas, búsqueda de listas, índices, representaciones de rangos totalizadores de esquemas basados en la concurrencia en la lengua; a manera de ejemplo, la identificación de determinados rasgos lingüísticos que se asocian funcionalmente y pueden dar origen a la cuantificación de las categorías gramaticales, ya sean marcadores discursivos, verbos artículos, adjetivos, adverbios, entre muchos. En segunda instancia, se puede desarrollar programas computacionales que permitan identificar las ocurrencias o desinencias de cada rasgo lingüístico en un texto determinado.

La lingüística del corpus, actualmente, ha desarrollado, de nuevo, el interés por los métodos estadísticos, de modo tal que se ha vuelto a revivir, la controversia, entre la perspectiva empirista y racionalista, en cuanto al análisis del lenguaje, se refiere y entre las técnicas estadísticas, frente a las simbolistas. En la actualidad, se está adoptando una estrategia sincrética que se encarga de la combinación de los principales logros, tanto en una, como en otra dirección.

La creación de programas de extracción y de recuperación de información, el desarrollo de analizadores, la corrección automática de textos y, en general, todo tipo de proceso que implica una información textual, requiere de la disposición de un tipo de recurso especial como lo es un corpus, ya sea de tipo bruto o anotado. A largo plazo se podrán derivar gramáticas y léxicos de amplia cobertura, a partir de la evidencia que puedan proporcionar los mismos.

La construcción de gramáticas generales de una lengua, requiere disponer de información sobre los usos reales de la misma, por parte de los hablantes. Las reglas gramaticales y los léxicos, que componen una lengua, deben procesar textos reales, lo cual evidencia que el recurso del hablante, aunque sea un experto en el tema, es insuficiente. Así, en el ámbito de las tecnologías de la lengua, los corpus se utilizan como fuente primaria de información para la elaboración de léxico y de gramáticas computacionales, puesto que proporcionan ejemplos sobre cómo los hablantes utilizan las palabras, cómo las combinan y les asignan significados.

La lingüística computacional, manifiesta en sus diferentes parcelas, por ejemplo, en el caso de estudios de análisis morfológicos, la tarea primordial consiste en reducir las diversas formas léxicas que se pueden encontrar en un texto de entrada en su forma básica o lema. Como complemento, el análisis morfológico, puede adjuntar al lema, en forma normalizada, la información gramatical transmitida por los afijos. A guisa de información y para el caso del español, los morfemas -s, y -es, podrían reducirse ambos, bajo un rasgo o corpus {núm.=plural} Como lo evidencia el siguiente ejemplo: 2

Casa/s ⟶ casa + plural
Jarron/es ⟶ Jarrón + plural

El análisis morfológico presupone que el texto de entrada se ha segmentado en palabras, una tarea que es relativamente, trivial, pues las palabras aparecen separadas por espacios en blanco o por signos de puntuación, pero el problema se podría evidenciar si se tratare de una lengua ideográfica, como el chino o el japonés, en donde, convencionalmente, no se separan las palabras dentro de la frase. En determinado momento, se complejiza la situación, pues, se puede evidenciar ciertas contracciones, como en el caso de el = del, en determinadas frases como: *la llave del carro* o, de formas clíticas, como los pronombres personales en el infinitivo y gerundio de los verbos, como en: *dárselos o prestándomelo.* Bastante más problemática se hace la existencia de las locuciones compuestas por una secuencia de dos o más palabras ortográficas que pueden componer una unidad semántica, a saber: *no obstante, sin embargo o a menudo*, pues, en ocasiones, pueden ser interpretadas

de manera composicional, en donde cada una de las palabras, que la conforman, conservan su sentido original, como se evidencia a continuación:

1. *Juan viene a menudo*
2. *¡A menudo sitio has venido a preguntar!*

En cuanto al primer componente, a menudo, forma una locución de tipo adverbial; mientras que en el segundo caso se trata de dos palabras diferentes, una preposición (a) y un adjetivo (menudo). Es aquí, en este tipo de oraciones, en donde se torna evidente, la necesidad de un sistema computacional que contribuya, mediante un programa específico, orientado por la lingüística computacional, que permita hacer transcripciones de este corte.

Los errores de ortografía, que las personas solemos cometer, en el momento de la escritura de un documento, mediante un ordenador, en este caso, el computador, pueden ser motivados por un desconocimiento de la norma o bien por un descuido. En cuanto al primer caso, cuando la persona no sabe cuál es la escritura correcta de determinada palabra, se habla entonces, de errores de competencia; éstos, por lo general, son de carácter individual, puesto que no todas las personas poseen el mismo nivel de conocimiento en cuanto a las reglas de ortografía de determinada lengua; sin embargo, existen ciertos factores lingüísticos que de una u otra manera, permiten la aparición de este fenómeno, como el grado de disparidad entre la ortografía y la fonética de determinada palabra: ombre/hombre, selebre/celebre, vaca/baca, voy/boy, entre muchos. La interferencia con otras normativas, también es otro factor que influye en este apartado como la nasalización de determinados fonemas: Inmenso/immenso, hinno/himno y, por último, la más usual, que es la baja frecuencia en cuanto al uso de determinada palabra.

Para el segundo caso, cuando la persona conoce cómo debe escribirse determinada palabra, pero por alguna causa, comete determinado error de escritura, se habla entonces, de errores de actuación. A manera de ejemplo, se puede evidenciar por intercambio de rasgos fonológicos de localización entre las consonantes líquidas desmolalizar por desmoralizar. También pueden provenir de descuidos visuales como probablemente por probablemente. O pueden ser producto de un desacierto mecanográfico: escritutra por escritura, estructutra por estructura.

Los verificadores ortográficos, son programas informáticos, que se dedican a la revisión de la ortografía de un texto. La tarea de estos programas se bifurca en dos actividades relevantes, a saber: por una parte la identificación de las palabras del texto, que pueden evidenciar algún error disortográfico (césar); mas la tarea se evidencia en la contraparte, que es la

segunda actividad, en hacer el respectivo correctivo en forma directa, inclusive, la mayoría de las veces, sin que el usuario se dé cuenta del error cometido; ahora, si esto no es posible, además de mostrar el error, el corpus sugiere, al usuario, la forma correcta, de acuerdo con su intención comunicativa.

Ahora, en cuanto a la gramática se refiere, concretamente, con los errores gramaticales, siempre se debe evidenciar si una secuencia de caracteres respeta o no las reglas instituidas por la normativa de determinada lengua. No es fácil, decidir si una secuencia de palabras, ortográficamente, correctas contiene errores gramaticales o no. A guisa de ejemplo:

A yo me parece que no tienes sentimientos.

El león contaba la historia

La puerta, enmudecida, veía todo lo que estaba pasando a su alrededor.

Las anteriores manifestaciones de habla, manifiestan la distinción entre dos aspectos muy importantes: la gramaticalidad y la aceptabilidad. Según Martí (2003, p. 34) la gramaticalidad estaría vinculada a la competencia o el conocimiento del lenguaje; mientras que la aceptabilidad sería un concepto relacionado con el ámbito de la aceptación o el uso del lenguaje. Entonces, se puede decir que en los ejemplos anteriores y con base en las anteriores definiciones que los enunciados aceptables son aquellos que no resultan extraños estilísticamente y cuya comprensión no requiere un gran esfuerzo de concentración o de memoria.

Desde el punto de vista formal, los errores gramaticales, propios de la escritura, mediante ordenador, suelen producirse por la omisión de una palabra (le informó que cantaría por le informó de que cantaría); también, se puede dar por la adición de una palabra (dice de llegará mañana por dice que llegará mañana) o por sustitución de una palabra por otra (Come de tres por come por tres).

Con base en lo anterior, se proponen tres técnicas para el tratamiento automático de la verificación gramatical:

1. La técnica basada en el reconocimiento de patrones. Existe un verificador que reconoce el texto, mediante la búsqueda de secuencias de palabras que siguen unas determinadas pautas de error preestablecidas; éstas pueden estar definidas en un aspecto meramente, gráfico o utilizar información de tipo lingüístico accesible para el verificador; aquí, los patrones de error pueden ampliarse con la correspondiente sugerencia de corrección. Algunos de estos patrones pueden ser los siguientes:

A fin que > a fin de que

A el > al

Se me > se me

Decir de que >decir que

Para que el verificador detecte determinado error, es necesario que el enunciado corresponda con uno de los patrones previamente elaborado; es decir, que para que la verificación sea efectiva, se debe anticipar el tipo de errores que se puedan presentar, pues no todos los errores gramaticales son fáciles de detectar.

2. Loa analizadores o parsers o análisis sintáctico. Esta técnica de verificación gramatical está aún en periodo de prueba y, se basa en los resultados de los programas informáticos que permiten identificar las estructuras sintácticas de las oraciones, de acuerdo con las normas gramaticales de cada lengua. El programa consiste en realizar los análisis sintácticos de las oraciones gramaticalmente incorrectas, atenuando las reglas gramaticales que la oración por analizar no respeta. A manera de ejemplo, en la oración: este niña no tiene cuadernos y, la regla gramatical que se va a analizar es la de concordancia entre determinante y nombre, la tarea del verificador es señalar el error de concordancia.

3. La técnica probabilística de identificación de errores gramaticales. Inicia, mediante el análisis estadístico de un corpus textual utilizado como modelo de uso lingüístico. En el corpus se comienza por establecer la categoría morfosintáctica de cada palabra, posteriormente, se determina la probabilidad de aparición en forma continua de cada combinación posible de dos categorías. La tarea del verificador consiste en detectar, en el texto, las palabras contiguas, con categorías morfosintácticas de baja probabilidad de coaparición en el corpus previo, tomado como modelo. La aplicación del método exige que el verificador sea capaz de descubrir la categoría de cada palabra del texto analizado, una tarea realizada gracias a los buenos resultados de los programas de etiquetadores.

La adquisición de un corpus para un estudio fonético-fonológico, por lo general, se puede realizar en estudios en donde existan entornos acústicamente, controlados con cámaras anecoicas o una sala insonorizada para evitar la influencia de ruidos externos; aunque también, se pueden realizar en ambientes naturales, pero es pertinente tener presente las dificultades para el análisis acústico que esto conlleva.

Para los estudios del tratamiento del habla, mediante la identificación de los elementos, existen sistemas que procesan la voz humana, tanto en aspectos de producción (síntesis de voz o speech synthesis), como de análisis, especialmente, para la identificación de palabras y unidades significativas (reconocimiento de voz o speech recognition). Este tipo de aplicaciones tiene muchos usos; algunos, hoy, ya son una realidad como la lectura de e-mails, la lectura de documentos y libros electrónicos, sistemas de dictado y otros que generan prometedoras perspectivas a futuro, como la invención de máquinas y robots, mediante la voz, el acceso telefónico a bases de datos, controladas sin intervención humana, la interacción con sistemas expertos, sistemas de autoaprendizaje y de tutoría automática, entre muchos más.

Un corpus oral, adaptado a las necesidades de la fonética y de las tecnologías del habla, contiene elementos básicos, a saber: señal sonora registrada; en algunos casos puede estar acompañada de datos articulatorios que facilitan el estudio del proceso de producción del habla. En cuanto a las características lingüísticas se puede tener en cuenta los estilos de habla, entendidos como una serie de dimensiones que varían en relación con la espontaneidad, la formalidad y el grado de preparación o planeación del discurso oral.

El soporte lingüístico de los corpus orales, que se utilizan en fonética y en las tecnologías del habla, toma, desde los sonidos aislados hasta la espontaneidad, incluyendo además, elementos específicos como los logatomes -palabras sin sentido, pero fonológicamente, bien formadas-; o las frases marco, frases de estructura controlada, en las cuales se debe insertar los elementos por analizar, también, conocidas como frases portadoras, cuyo ejemplo típico es el "dijo…. y salió".

Los corpus utilizados para el campo del reconocimiento del habla, pueden incluir las denominadas frases fonéticamente equilibradas. En éstas, la frecuencia de aparición de los sonidos en el corpus debe ser equivalente a la de la lengua en general o, en su efecto, con una representación de todas las combinaciones de los sonidos de dicha lengua: dígitos, números conectados, secuencias alfanuméricas, letras y palabras dichas, fecha y horas, antropónimos y topónimos y palabras relacionadas con la aplicación.

La lexicografía es otro campo en donde la lingüística computacional ha tenido auge. En los últimos años, ha resurgido el interés por el estudio del componente léxico desde varias perspectivas; hoy, no sólo despierta el interés de los lexicógrafos, sino también de todas aquellas personas que trabajan dentro del campo de la ingeniería del lenguaje, la lógica y la representación semántica., En estas parcelas de la lingüística, se trabajan la selección léxica, la desambiguación de sentidos y el tratamiento de la similitud semántica.

En cuanto a la colocación hace alusión a las construcciones semi-idiomáticas, formada por dos construcciones léxicas L1 y L2, en donde L2 es escogida de un modo arbitrario para expresar un determinado sentido o, también, para expresar un papel sintáctico en función de L1. La característica fundamental de las colocaciones es la coocurrencia léxica restringida entre los dos constituyentes de la colocación. De acuerdo con Hausmann (1998, p. 65), la coocurrencia de dos unidades léxicas es restringida si para expresar un significado L2, teniendo en cuenta la unidad léxica L1, la elección de L2, que expresa el significado L2, debe estar léxicamente determinada por L1 y la combinación de L1 y L2, viene a formar una colocación en donde L1 es la base y L2 es el colocativo.

A manera de ejemplo y para aclarar lo anterior, veamos la coocurrencia de amigo y de inseparable, es léxicamente restringida, puesto que para expresar el significado intenso o en alto grado de amigo, la elección de inseparable, que expresa el significado intenso, está léxicamente determinado por amigo. Por tanto, se está ante una colocación formada por la base amigo y el colocativo inseparable. El mismo sentido intenso o en alto grado no puede ser expresado como inseparable, si la base es, por ejemplo, cazador; en este caso, la unidad léxica selecciona el colocativo furtivo. La combinación cazador furtivo, constituye también, una colocación.

Otro ejemplo de colocaciones se encuentra en las construcciones verbales, realizadas con verbos de apoyo o verbos soporte, como por ejemplo, dar un paseo, infligir una derrota, echar un vistazo, tomar una decisión, prestar juramento, entre muchas. Del mismo modo, el significado del verbo causar es expresado, de algún modo, restringido, especialmente, con nombres que designan sentimientos; por ejemplo: despertar curiosidades, dar vergüenza, sembrar el descontento, causar pena, entre muchos.

En cuanto a la zona semántica, se consigna una etiqueta semántica y una fórmula proposicional. Aquí, se pone el énfasis en el aspecto combinatorio y no en el explicativo; no se incluye una verdadera lexicografía de cada unidad léxica. En su lugar aparece una etiqueta semántica, que debe representar el significado central de la unidad léxica. Por ejemplo, dentro del campo semántico de los nombres que designan sentimientos, es frecuente encontrar vocablos polisémicos en donde una de las unidades léxicas es etiquetada como sentimiento y otra como hecho y objeto; veamos:

1.

a. Siento pena por la muerte de Juan
b. Juan tiene envidia de su hermano
c. Tengo esperanza de aprobar el examen

d. Hoy he sufrido un gran discurso
e. Siento orgullo por su triunfo

2.

a. Mi mayor pena es que se haya muerto Juan
b. Juan es la envidia de todos
c. Mi esperanza es aprobar el examen
d. La pérdida del documento fue un gran disgusto para ella
e. Tú eres el orgullo de la familia

La fórmula proposicional representa la estructura de argumentos o estructura actancial de la unidad léxica en cuestión. Por ejemplo, para la unidad léxica compasión 1, su formula proposicional dice que se trata de un nombre con dos actantes: compasión de persona X, por individuo Y. Cada uno de los actantes es etiquetado a su vez, por una etiqueta semántica: el que siente compasión debe ser una persona y el objeto de la compasión, para el caso del ejemplo, es un individuo, pero también puede ser una cosa, un animal

a. Siento compasión <por el daño ocasionado a tantas personas> por las víctimas.
b. Siento compasión <por su terrible enfermedad> por tu hermano enfermo.
c. Siento compasión <por su extrema pobreza> por los pobres.
d. Siento compasión <por el maltrato de los animales> por los animales maltratados.

En cuanto a la indexación de documentos se refiere, hoy, existen algunos problemas, que pueden ser resueltos, mediante la utilización de la lingüística computacional. En primer lugar, es recomendable, por motivos de eficiencia, que el índice no contenga las diferentes formas de cada palabra (singular y plural de los nombres, formas verbales, entre muchas), sino una forma básica o "lema", lo cual promueve al estudio de los análisis de formas de afijación, ya sea prefijos, infijos o sufijos, de tiempos verbales, géneros, números, etc. Además, el índice debe incluir, no sólo palabras, sino también términos, expresados mediante expresiones complejas, como por ejemplo: libertad de cátedra, seguro de enfermedad, seguro de vida, seguro contra terceros, entre muchas.

Otro aspecto por resaltar tiene que ver con la búsqueda de información en bases de datos o documentales (information retrieval) que consiste en la extracción de textos de diferentes bases documentales, gracias a las distintas maneras de consultas por parte de los usuarios, y su ordenación, mediante criterios definidos; normalmente, de acuerdo con la pertinencia de las consultas elaboradas. Ejemplos claros de este tipo de aplicación son los diferentes buscadores existentes en internet: Google, Hot mail, (gmail), yahoo, entre otros.

La clave del funcionamiento de estas aplicaciones es la construcción previa de un índice de palabras y conceptos clave que contiene cada documento (indexation). Las consultas son rápidas y eficientes, pues son realizadas sobre los índices.

En suma, se debe mencionar que la lingüística del corpus contribuye en áreas propias del saber lingüísico, como por ejemplo en las gramáticas generativa, cognitiva y la funcional, en la semántica, en la pragmática, en la lingüística textual, en el discurso, en la lingüística funcional, en la sociolingüística, mediante diversos programas de lingüística computacional como: analizadores básicos relacionados con la morfología, el léxico, la sintaxis, recursos que permiten establecer relaciones léxicas entre palabras o relaciones sintácticas entre los constituyentes de una oración; los correctores y traductores automáticos, herramientas que, con mucha frecuencia, están incorporados a los procesadores de texto; los programas de síntesis y reconocimiento de voz que incorporan la dimensión fonética; los programas complejos, que incluyen herramientas útiles para realizar resúmenes de textos y ayudan a realizar documentos altamente, estandarizados.

REFERENCIAS BIBLIOGRÁFICAS

Blecua, J. M.; Clavería, G.; Sánchez, C. y Torruella, J. (1999). *Filología e informática. Nuevas tecnologías en los estudios filológicos.*

Hausmann, F. J. "O diccionario de colocación. Criterios de organización. Santiago de Compostela, Centro Ramón Pineiro, Xunta de Galicia.

Lavid, J. (2005). *Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*. Madrid, España: Cátedra.

Martí, M.A. (2003). *Tecnologías del lenguaje*. Barcelona: Uoc.

Martí, M.A.; Fernández, M. A. y Vásquez, G. G. (2003). *Lexicografía computacional y semántica.* Barcelona, Gráficas Rey.

Ong, W. J. (2004). *Realidad y escritura. Tecnologías de la palabra*. México: Fondo de Cultura Económica.

Toffler, A. (1996) La tercera ola: Barcelona: Plaza y Janés.

SOCIOLINGÜÍSTICA Y AFINES