doi: 10.15446/fyf.v28n1.51970

LA LINGÜÍSTICA DE CORPUS: PERSPECTIVAS PARA LA INVESTIGACIÓN LINGÜÍSTICA CONTEMPORÁNEA*

Sergio Bolaños Cuéllar **
Universidad Nacional de Colombia, Bogotá – Colombia

Resumen

La Lingüística de Corpus (LC) constituye una de las áreas de investigación de mayor desarrollo en los estudios del lenguaje. En este artículo se hace una sucinta revisión histórica al uso de los corpus, desde su confección manual en la lexicografía del s. XVII, pasando por los estudios bíblicos y gramaticales en los s. XVIII y XIX, hasta llegar a la primera generación de corpus electrónicos en la década de los sesenta y las generaciones posteriores. Se discute la definición del término *corpus*, su tipología y sus principales características. Así mismo, se muestra el potencial metodológico de la lingüística de corpus para la investigación en diversas áreas de la lingüística contemporánea, entre otras, la revisión de la traducción; la elaboración de materiales para la enseñanza de las lenguas; la diferenciación léxico-gramatical de variedades dialectales de una lengua; la elaboración de gramáticas descriptivas y el análisis sociolingüístico del discurso.

Palabras clave: *lingüística de corpus, definición de* corpus*, tipología de corpus, aplicaciones de la lingüística de corpus, megacorpus.*

Cómo citar este artículo:
Bolaños, S. (2015). La lingüística de corpus: perspectivas para la investigación lingüística contemporánea. *Forma y Función, 28*(1), 31-54. doi: 10.15446/fyf.v28n1.51970

Artículo de investigación. Recibido: 22-09-2014, aceptado: 11-12-2014.

---

* Una versión preliminar de este trabajo se presentó en el "XXVII Congreso Nacional de Lingüística, Semiótica y Literatura y II Congreso Internacional de Lingüística" (Bogotá, 11-13 septiembre de 2014, Departamento de Lingüística, Universidad Nacional de Colombia). El autor es líder del grupo de investigación *LINGUAE-Comunicación, Bilingüismo y Traducción*. El artículo se inscribe como producto del proyecto *Métodos de la investigación lingüística: La lingüística de corpus*.

** Doctor en Lingüística por la Universidad de Hamburgo. Profesor del Seminario de Lingüística Inglesa y el curso Historia de la Lengua Alemana del Programa de Filología e Idiomas de la Universidad Nacional de Colombia. Investigador en teoría y práctica de la traducción y traductor e intérprete oficial. sbolanosc@unal.edu.co

CORPUS LINGUISTICS: APPROACHES FOR CONTEMPORARY LINGUISTIC RESEARCH

Abstract

Nowadays Corpus Linguistics (CL) is one of the most developed research areas in linguistic studies. This paper presents an abridged historical review of the use of corpora from the hand-made lexicography of the 17th c., through biblical and grammatical studies of 18th and 19th c., to the first generation of electronic corpora in the 60s and later. A 'corpus' definition as well as its typology and main characteristics are also discussed. Likewise the methodological potential of CL for research in contemporary linguistics is also depicted, e.g. in translation critique, materials design for language teaching, lexical and grammatical differentiation of dialectal varieties, the making of descriptive grammars, and sociolinguistic discourse analysis.

Keywords: *corpus linguistics, definition of* Corpus*, corpus typology, applications of CL, magacorpora.*

A LINGUÍSTICA DE CORPUS: PERSPECTIVAS PARA A PESQUISA LINGUÍSTICA CONTEMPORÂNEA

Resumo

A Linguística de Corpus (LC) constitui uma das áreas de pesquisa de maior desenvolvimento nos estudos da língua. Neste artigo, faz-se uma sucinta revisão histórica do uso dos corpus, desde sua confecção manual na lexicografia do século XVII, passando pelos estudos bíblicos e gramaticais nos séculos XVIII e XIX, até chegar à primeira geração de corpus eletrônicos na década de 1960 e às gerações posteriores. Discute-se a definição do termo *corpus*, sua tipologia e suas principais características. Também se mostra o potencial metodológico da linguística de corpus para a pesquisa em diversas áreas da linguística contemporânea, entre outras, a revisão da tradução; a elaboração de materiais para o ensino de línguas; a diferença léxico-gramatical de variedades dialetais de uma língua; a elaboração de gramáticas descritivas e a análise sociolinguística do discurso.

Palavras-chave: *linguística de corpus, definição de* corpus*, tipologia de corpus, aplicações da linguística de corpus, megacorpus.*

Introducción

Sin duda, cuando se encuentra por primera vez el término *lingüística de corpus* surgen inmediatamente algunos interrogantes: ¿Se trata acaso de la misma lingüística computacional? ¿Es una nueva aproximación conceptual a los fenómenos del lenguaje? ¿Se trata, más bien, de una metodología?

En primer lugar, vale la pena aclarar que la lingüística de corpus (LC) no es lo mismo que la lingüística computacional. El objetivo de la lingüística computacional consiste, fundamentalmente, en diseñar herramientas informáticas que permitan hacer un procesamiento óptimo, es decir, comprensión y generación del denominado lenguaje natural (Moreno, 1998, p. 13), de modo que este pueda analizarse de manera confiable, eficiente y, especialmente, automática. Para ello se han desarrollado diversos programas, a partir de una modelación estadística o basada en reglas, que permiten, por ejemplo, la caracterización y posterior búsqueda de categorías léxicas o combinación de palabras en textos, por medio de los etiquetadores (*taggers*), así como la descripción y búsqueda de ciertas estructuras sintácticas, mediante el uso de los analizadores sintácticos (*parsers*). El andamiaje conceptual y las herramientas desarrolladas en la lingüística computacional propenden por ir más allá de las manifestaciones superficiales del lenguaje (palabras y estructuras sintácticas) y, al enfrentar problemas concretos de automatización, como el caso de la traducción automática (*machine translation*), se han visto en la necesidad de realizar el análisis de las dimensiones semántica y, por supuesto, pragmática del lenguaje. Estas últimas tareas han requerido investigación adicional y el diseño de etiquetadores semánticos que, por ejemplo, sean capaces de asignar unidades léxicas a un campo semántico. De igual manera, los etiquetadores pragmáticos deben permitir identificar fenómenos transoracionales como la anáfora o la correferencia con antecedentes textuales implícitos o explícitos (Rocha, 1997). La utilización de etiquetadores y analizadores sintácticos hace parte de un proceso que se conoce como *anotación del corpus*. En general, la lingüística de corpus centra su interés en la utilización de los corpus disponibles o en la confección de corpus mediante el empleo de las herramientas que han diseñado para este fin los lingüistas computacionales. El lingüista de corpus no es necesariamente un programador como sí lo es el lingüista computacional. Se trata, pues, de actividades estrechamente ligadas y con linderos borrosos pero, en principio, podemos señalar que tienen objetivos distintos. A la lingüística computacional le compete el diseño de las herramientas informáticas para un óptimo procesamiento automático del lenguaje natural. Además de la traducción automática, área de trabajo con la que se reconoce que

comenzó la lingüística computacional en la década de los años sesenta del siglo pasado, Jurofsky y Martin (2008) señalan que también son áreas de interés de la lingüística computacional: la generación del lenguaje natural, el reconocimiento del habla, la síntesis texto-habla, la solución de problemas semánticos, como la ambigüedad léxica, y pragmáticos, como la anáfora. Por otra parte, a la lingüística de corpus le interesa realizar análisis lingüísticos léxicos, gramaticales, semánticos y pragmáticos (discursivos) mediante el uso de las herramientas informáticas diseñadas para este fin.

En cuanto al segundo interrogante sobre si la lingüística de corpus constituye una nueva aproximación conceptual al estudio del leguaje, podemos señalar lo siguiente. El giro en la lingüística que se llevó a cabo en los años setenta (Helbig, 1986, p. 13), que marca una transición de los estudios sistémico-teóricos al énfasis en lo comunicativo-pragmático[1], se concretó en la consolidación y el desarrollo de nuevas orientaciones teóricas como la textolingüística, la teoría de actos de habla, el análisis conversacional, la sociolingüística y la psicolingüística. Todas estas perspectivas tienen en común el hecho de que estudian el lenguaje en contexto, o si se quiere, utilizando la dicotomía chomskiana, se puede afirmar que prestan atención, en primer lugar, a la actuación (*performance*), no a la competencia (*competence*), aunque hoy día reconocemos que no se trata de una oposición tajante, sino de una diferenciación conceptual y de enfoque[2].

En otras palabras, el interés de todas estas disciplinas se centra la actuación chomskiana (o también la *parole* saussureana). Lo que resulta en realidad curioso es que, cuando se habla del giro en la lingüística de los años setenta, generalmente no se menciona la lingüística de corpus que, al igual que las disciplinas antes referidas, pone énfasis en los corpus, es decir, en la colección de textos o fragmentos textuales

1 „Seit etwa 1970 ist in der Sprachwissenschaft international eine ‚kommunikativ-pragmatische Wende' zu beobachten, d.h. eine Abwendung von einer system-orientierten bzw. –zentrierten Linguistik und eine Zuwendung zu einer kommunikationsorientierten Linguistik". (Helbig, 1986, p. 13).

2 No es posible concebir que haya 'actuación' lingüística o 'uso' del lenguaje, independientemente del sistema de reglas gramaticales que hace parte del andamiaje cognitivo del hablante. El sistema de reglas gramaticales se modifica y 'reorganiza' conforme al uso real del lenguaje por parte de los hablantes. Así mismo, tampoco es viable concebir un sistema de reglas gramaticales sin usuarios que lo utilicen para llevar a cabo procesos cognitivos de pensamiento y para interactuar comunicativamente con otros miembros de su comunidad lingüística.

que reflejan el uso real de una lengua en forma oral o escrita por parte de hablantes reales. Esto se puede deber al hecho de que algunos de los partidarios de Chomsky desde un comienzo señalaron que el estudio de corpus era "una empresa inútil e insensata", puesto que "la única fuente legítima del conocimiento gramatical de una lengua era la intuición del hablante nativo, la cual no podía obtenerse a partir de un corpus" (Meyer, 2004, p. 1). Queda claro, pues, que la lingüística de corpus se inscribe en el mismo paradigma de las disciplinas lingüísticas que se dedican a estudiar las diferentes manifestaciones del uso del lenguaje en contextos reales de interacción comunicativa, lo que, en otras palabras, correspondería a un enfoque lingüístico funcional.

En cuanto al tercer interrogante sobre la naturaleza misma de la lingüística de corpus, parecería haber consenso en que se trata de una aproximación metodológica más que de un paradigma teórico distinto (Meyer 2004, p. xi). Por su parte, Biber, Conrad y Reppen (1998/2006) amplían la caracterización de los análisis basados en corpus y señalan que el elemento fundamental es la existencia de los "patrones de asociación", es decir, "la forma sistemática en que se utilizan los rasgos lingüísticos en asociación con otros rasgos lingüísticos y no lingüísticos" (Biber et al., 1998/2006, p. 5). Así mismo, presentan las características fundamentales de los análisis basados en corpus: 1. Son empíricos: analizan los verdaderos patrones de uso en textos naturales; 2. Utilizan un conjunto de textos naturales, organizados según principios, en forma de corpus, para realizar el análisis; 3. Utilizan los computadores, en gran medida, mediante técnicas automáticas e interactivas; y 4. Dependen de técnicas de análisis cualitativas y cuantitativas (Biber et al., 1998/2006, p. 4). Aquí vale la pena subrayar que la lingüística de corpus no solo constituye un nuevo enfoque metodológico, que cada vez empieza a conocerse más y mejor, sino que implica, además, una perspectiva conceptual diferenciada de los fenómenos lingüísticos. La hipótesis subyacente es que en el uso del lenguaje siempre se presentan patrones de asociaciones de diversa índole y con distinto grado de complejidad que permiten descubrir o corroborar intuiciones lingüísticas y presupuestos teóricos de los enfoques tradicionales, estructurales, pragmáticos, discursivos o interdisciplinarios. Esta transversalidad disciplinaria, aunada a la perspectiva de análisis cualitativo y cuantitativo de manera integrada, es lo que caracteriza a la lingüística de corpus[3].

3 En todo caso, vale la pena señalar que, por lo general, los estudios de LC han sido de focalización local, es decir, se centran en unidades léxicas o sintácticas y no se preocupan por aprehender la totalidad textual, de modo que fenómenos como la superestructura textual (Van Dijk, 1980), por definición holística, quedarían por fuera del análisis de corpus tradicional.

Definición y características de un corpus

En cuanto a la definición del término *corpus*, por lo general, los autores presentan algunos rasgos o características definitorias del mismo. Por ejemplo, McEnery, Xiao y Tono (2006/2008) señalan que cada vez hay más consenso en definir el corpus como una recopilación de textos 1) legibles por computadora, 2) auténticos, 3) que son escogidos por muestreo y 4) son representativos. Enseguida, aclaran los autores, parece haber acuerdo sobre las primeras dos características, pero hay controversia sobre cómo realizar el muestreo y a qué se refiere la representatividad del corpus (McEnery et al., 2006/2008, p. 5).

Por su parte, en el ámbito germanófono, Lemnitzer y Zinsmeister (2006/2010, p. 40) proponen una definición del término *corpus*:

> Un corpus es una recopilación de expresiones escritas u orales en una o varias lenguas. Los datos del corpus están digitalizados, es decir, son almacenados y pueden ser leídos por computadora. Los componentes del corpus, las secuencias de textos o expresiones, constan de los datos mismos y de metadatos, que describen estos datos, y de anotaciones lingüísticas que los clasifican[4].

En términos generales, se puede señalar que los metadatos, es decir, la información que acompaña a cada una de las expresiones o los textos incluidos en el corpus, ayuda fundamentalmente a contextualizar aspectos relevantes de dicho corpus, por ejemplo, cuándo se produjo o se publicó, dónde, quién es el autor, etc. Con el fin de facilitar el intercambio y la confección de corpus parciales se acude a estándares de descripción de estos metadatos. En otras palabras, se busca que haya mayor acceso a la información sobre los corpus recolectados y que esta pueda ser consultada en el ámbito internacional por cualquier investigador. Estos intentos de unificación o estandarización se conocen como *iniciativas*. Así, por ejemplo, se registra la *Dublin Core Metadata Initiative*, que permite caracterizar diferentes tipos de objetos digitales: imágenes, sonido y texto. Además, para describir cada recurso informativo se utilizan las categorías: título, generador, objeto, descripción, autores,

4 „Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen in einer oder mehreren Sprachen. Die Daten des Korpus sind digitalisiert, d.h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus, die Texte oder Äußerungsfolgen, bestehen aus den Daten selbst, sowie Möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind". (Lemnitzer & Zinsmeister, 2006/2010, p. 40).

editorial, derechos y fecha. Otra iniciativa importante es la *ISLE Metadata Initiative* (*IMDI*), que en principio puede utilizarse para recursos lingüísticos de todo tipo, pero se emplea sobre todo para corpus orales o multimodales. Puede mencionarse finalmente la *Text Encoding Initiative* (*TEI*) que se emplea en un espectro amplio de textos y corpus.

Veamos un ejemplo de metadatos, tomado de Nelson (1996, p. 51) y citado por Meyer (2004, p. 83), que corresponde al encabezado de un texto incluido en el *Corpus Internacional del Inglés*:

```
1. <text.info>
2. <file.description>
3. <textcode>ICE-GB-W2A-008</textcode>
4. <number.of.subtexts>1</number.of.subtexts>
5. <category>printed;informational:popular:humanities</category>
6. <wordcount>2101</wordcount>
7. <version>tagged by TOSCA tagger using the ICE tagset</version>
8. <free.comments> </free.comments>
9. </file.description>
```

Tal como se puede observar, en la primera línea se identifica la información textual (<*text.info*>); luego, en la segunda línea se abre la instrucción sobre la descripción del archivo (<*file.description*>) y se cierra en la línea 9 (</*file.description*). El cierre de la instrucción está indicada por la barra oblicua (/). En la línea 3 se incluye la instrucción sobre la codificación del texto (<*textcode*>); en este caso se trata del *International Corpus of English* (*ICE*) y la variedad del inglés británico (*GB*). Ya internamente en el corpus se identifica el texto como *W2A-008.* En la línea 4 se registra el número de subtextos, en este caso 1 (<*number.of.subtexts*>). La línea 5 presenta la tipología textual (<*category*>), en este caso un texto impreso, informativo, popular y del área de las humanidades. En la línea 6 se indica el número de palabras del texto (<*wordcount*>), es decir, 2101. La línea 7 presenta la información sobre el etiquetador (*tagger*) utilizado para etiquetar el texto (<*version*>). Y, finalmente, en la línea 8 pueden insertarse comentarios libres del texto (<*free.comments*>).

El segundo grupo de metadatos, conocidos como la *anotación del corpus*, contiene la información lingüística de los textos incluidos en el corpus, la cual ha sido ingresada utilizando los etiquetadores (*taggers*), los analizadores sintácticos (*parsers*) y demás herramientas diseñadas para este fin. Para Lemnitzer y Zinsmeister (2006/2010, p. 64), los principales niveles de la anotación son: morfosintáctico

(categoría léxica, *Part of Speech*/*POS*), morfológico (morfología inflexiva, formas de base), sintáctico (constituyentes o dependencias, a menudo con funciones sintácticas), semántico (nombres propios, tipos de lectura/*Word Senses*, marcos temáticos/*Frames*), pragmático (correferencia, estructura informativa, estructura discursiva), otros (estructura textual, ortografía, anotación de errores, rasgos fonéticos y prosódicos, rasgos gestuales y mímica). A este respecto, McEnery et al. (2006/2008, p. 34) precisan aún más los diferentes tipos de anotación del corpus: a nivel fonológico puede anotarse en el corpus la frontera silábica (anotación fonética/fonémica) o rasgos prosódicos (anotación prosódica); en el nivel morfológico pueden distinguirse prefijos, sufijos y raíces (anotación morfológica); a nivel léxico pueden distinguirse las partes de la oración (etiquetado *POS*), lemas (lematización) y campos semánticos (anotación semántica); en el nivel sintáctico puede anotarse el análisis sintáctico (analizadores/*parsers*; diagramas arbóreos/*treebanking*; constituyentes/*bracketing*); en el nivel discursivo los corpus pueden mostrar relaciones anafóricas (anotación correferencial), información pragmática como actos de habla (anotación pragmática) o rasgos estilísticos como presentación del registro de habla y pensamiento, especialmente en textos literarios (anotación estilística).

Veamos un ejemplo de anotación lingüística, tomado de Meyer (2004, p. 93):

```
1.  The det:>2 @DN> DET SG/PL
2.  child child subj:>3 @SUBJ N NOM SG
3.  broke break main:>0 @+FMAINV V PAST
4.  his he attr:>5 @A> PRON PERS MASC GEN SG3
5.  arm arm obj:>3 @OBJ N NOM SG
6.  and and @CC CC
7.  his he attr:>8 @A> PRON PERS MASC GEN SG3
8.  wrist wrist @SUBJ N NOM SG @OBJ N NOM SG @PCOMPL-S N NOM SG
@A>N NOM SG

9.  and and cc:>3 @CC CC
10. his he attr:>11 @A> PRON PERS MASC GEN SG3
11. mother mother subj:>12 @SUBJ N NOM SG
12. called call cc:>3 @+FMAINV V PAST
13. a a det:>14 @DN> DET SG
14. doctor doctor obj:>12 @OBJ N NOM SG

EngFDG Parser
```

Lo primero que hemos hecho es resaltar en negrita la primera palabra de cada línea, de modo que se pueda leer verticalmente, de arriba abajo, la oración del texto que ha sido anotado:

The child broke his arm and his wrist and his mother called a doctor.

A continuación de cada una de las palabras en negrita aparece el lema correspondiente, es decir, la palabra que se toma como base y a partir de la cual la lengua presenta variaciones morfológicas[5]. Así, por ejemplo, el determinante solo tiene una forma de lema *the*, para masculino femenino singular y plural. En este caso se marca *SG/PL*. *Child* es el lema de un nombre (*N*), nominativo (*NOM*), singular (*SG*) que cumple la función de sujeto de la oración (*SUBJ*). En el caso de la forma verbal *broke* se observa que está en pasado (*PAST*) y el lema es la forma verbal principal *break.* Por su parte, *his* es la forma genitiva (*GEN*) o posesiva, de la tercera persona singular (*SG3*) del pronombre personal masculino (*PRON PERS MASC*) en función atributiva (*attr*). Debido a que este etiquetado se hace de manera automática mediante un programa diseñado para tal fin, llama la atención que al encontrarse con el nombre *wrist*, el programa ha sido incapaz de determinar si es sujeto (*SUBJ*) u objeto de complemento (*OBJ*). Esta ambivalencia sintáctica se debe al hecho de que la conjunción *and* puede ser leída como parte del sintagma nominal regido por el verbo *broke* (*broke his arm AND his wrist*) o como una conjunción de una la primera cláusula con otra cláusula introducida por la conjunción (*The child broke his arm AND [otra cláusula]*). Queda claro, pues, que el etiquetado sintáctico requiere una cuidadosa revisión manual para corregir este tipo de errores.

Confección de corpus pre-electrónicos

La realización de estudios lingüísticos basados en corpus sin apoyo de computadores tiene una larga tradición. Kennedy (1998) recoge y sistematiza estas primeras experiencias de trabajo lingüístico con apoyo de corpus pre-electrónicos y los divide en cinco campos de indagación: 1) Los estudios bíblicos y literarios,

5 La lematización es un proceso de marcación del corpus que permite tener acceso rápido a las diversas formas de una palabra, denominada lema. Por ejemplo, para el caso de los verbos en inglés *goes, gone, went, going,* etc., son variantes del lema *go*. En español, voy, vas, fui, iremos, etc., son variantes del lema *ir*. Por lo general, el lema en los verbos coincide con la forma inflexiva del infinitivo.

2. La lexicografía, 3. Los estudios dialectales, 4. Los estudios de formación lingüística y 5. Los estudios gramaticales (Kennedy, 1998, p. 13).

Una de las primeras áreas de interés tiene que ver con la utilización de la Biblia para realizar comentarios o exégesis. Ya desde el siglo XVIII, anota Kennedy, se elaboraban listas y concordancias de las palabras que se empleaban en la Biblia con el fin de poder determinar si estas eran consistentes. Se menciona, por ejemplo, el trabajo de Alexander Cruden, quien publicó sus Concordancias (*Concordance*) en 1736, con base en la versión autorizada de la Biblia del rey Jacobo (*King James Version/Authorized Version*, 1611). Kennedy muestra en un cuadro un ejemplo de la concordancia que estableció Cruden (1769) entre las palabras *dry* y *ground*:

dry *ground*

behold the face of the *ground* was *d.*	*Gen* 8:13
Israel shall go on *d. ground* in the sea.	*Ex* 14:16,22
stood firm on *d. ground* in Jordan. Israel passed on *d. ground.*	*Josh* 3:17
Elijah and Elisha went over on *d. ground.*	*2 Ki* 2:8
he turneth water-springs into *d. ground.*	*Ps* 107:33
he turneth *d. ground* into water-springs.	35
I will pour floods upon the *d. ground.*	*Is* 44:3
He shall grow as a root out of a *d. ground.*	53:2
She is planted in a *d.* and thirsty *ground.*	*Eze* 19:13

El objetivo fundamental que se perseguía con la búsqueda de concordancias en la Biblia era poder asegurar que hubiera consistencia léxica y semántica en las formas traducidas, es decir, que un mismo término del original hebreo, griego o latino se tradujera en todos los casos de la misma manera. Hoy día este ejercicio de consistencia léxica en la traducción se sigue llevando a cabo y, gracias al advenimiento del computador, esta verificación de concordancias se realiza de manera casi inmediata.

En cuanto al trabajo lexicográfico, ya en el siglo XVII, Samuel Johnson había recolectado manualmente, con sus colaboradores, 150 000 citas de autores connotados con el fin de ilustrar el significado y el uso de las palabras inglesas para su *Dictionary of the English Language*. En época más reciente el *Oxford English Dictionary*, publicado inicialmente en 1928, acudió a un corpus del inglés literario escrito, recopilado durante 71 años. De igual manera, en el siglo XIX, Noah Webster se valió de un monumental corpus de citas para su obra *An American Dictionary of the English Language*, publicada

inicialmente en 1828 y la cual, para su tercera edición como *New International Dictionary,* contaba con un corpus de más de 10 millones de citas. En nuestro contexto latinoamericano resulta pertinente recordar la ardua labor lexicográfica que llevó a cabo a finales del siglo XIX y comienzos del XX el lingüista colombiano Rufino José Cuervo, mediante el acopio de múltiples citas de los escritores más connotados de la lengua española hasta ese entonces, que se concretó en los primeros tomos del *Diccionario de construcción y régimen de la lengua castellana.* En esta misma línea del trabajo lexicográfico se inscribe también el interés por la variación dialectal en el uso de las palabras con su ortografía y pronunciación, el cual se concreta en la obra *English Dialect Dictionary* (Wright, 1998-1905) y *The Existing Phonology of English Dialects* (Ellis, 1989) (Kennedy, 1989, pp. 15-16).

En el área de formación lingüística, concretamente la enseñanza del inglés para hablantes nativos, Kennedy reseña el trabajo de Thorndike (1921), quien compiló un corpus de 4.5 millones de palabras para confeccionar una lista de frecuencia de uso que sirviera de guía para los materiales didácticos correspondientes. La principal fuente del corpus era la Biblia y otra parte correspondía a las obras de ficción del siglo XIX, especialmente de Charles Dickens. Otras experiencias de compilación de corpus manuales que vale la pena mencionar son el corpus de 11 millones de palabras elaborado bajo la coordinación de J. W. Kaeding en Alemania en la década de 1890 y el corpus de medio millón de palabras que compiló H. V. George sobre el inglés literario y de periódicos en la India a mediados del siglo XX. Así mismo, a comienzos del s. XX se empezaron a confeccionar las primeras gramáticas descriptivas a partir de corpus recolectados y no de la simple introspección. A este respecto, son muy conocidas las obras de Jespersen (1909-1949) (*Modern English Grammar on Historical Principles*) y de Fries (*American English Grammar*). Este último registró, por ejemplo, que era común decir en el inglés de la época *they sung* o *it sunk*, en vez de las formas canónicas *they sang* y *it sank,* y que el subjuntivo había prácticamente desaparecido, afirmación, esta última, que está por corroborar en el inglés contemporáneo.

Corpus electrónicos

Según Kennedy (1998), los corpus electrónicos pueden clasificarse desde diferentes puntos de vista. En una primera aproximación se distingue si el corpus incluye una amplia diversidad de tipos de texto (o géneros/dominios), en cuyo caso se hablará de corpus generales o balanceados o si se trata, más bien, de corpus

especializados. Como ejemplo de corpus general se puede mencionar el *Survey of English Usage* (SEU), confeccionado por Randolph Quirk a comienzos de la década de los sesenta a partir de 200 muestras de 5000 palabras cada una, que representaban el inglés británico hablado y escrito de la época. Los textos escritos incluían textos impresos, no impresos y con guion escrito para ser hablados; los textos orales comprendían monólogos y diálogos. Por otra parte, los corpus especializados recogen textos con un propósito específico, por ejemplo, el inglés utilizado en la exploración geológica, la perforación y la refinería de la industria del petróleo; otros corpus especializados pueden incluir dialectos, variantes regionales, formas no estándar de la lengua o incluso pueden describir las diversas formas lingüísticas que utilizan los aprendices de una segunda lengua o lengua extranjera, comúnmente conocidos como corpus de aprendices. Probablemente uno de los corpus de aprendizaje de lenguas mejor conocido es CHILDES (*Child Language Data Exchange System*). Este corpus, creado en 1984 por Brian MacWinney y Catherine Snow, incluye muestras en veinte lenguas de transcripciones, audio y video de niños en proceso de aprendizaje monolingüe y bilingüe.

Veamos un ejemplo de un niño bilingüe español-inglés:

0 @Loc: Biling/Perez/john1.cha
1 @PID: 11312/c-00001825-1
2 @Begin
3 @Languages: spa , eng
4 @Participants: CHI John Target_Child , MOT Mother
5 @ID: spa , eng|perez|CHI|2;0.||||Target_Child|||
6 @ID: spa , eng|perez|MOT|||||Mother|||
7 @Date: 10-MAY-2000
8 @Time Start: 20:30
9 @Time Duration: 20:30-20:53
10 @Situation: at home , after bath and before bed
11 @Activities: reading books
12 @Location: Ann Arbor , MI
13 @Tape Location:tape1 , side b , 310-635
14 @Comment: background noises
15 *CHI: night time .
16 *CHI: tea xxx .
17 *CHI: xxx .
18 *CHI: it's duck .

19 *CHI: this xxx .
20 *CHI: xxx .
21 *CHI: my xxx .
22 *CHI: doggie .
23 *CHI: right here .
24 *CHI: ne baby .
25 *MOT: it's a nene , no?
26 *CHI: nene doggie .
27 *CHI: doggie training .
28 *CHI: doggie got the book .
29 *MOT: six nenes .
30 *MOT: look what's he doing?
31 *CHI: p(l)aying grandpa .
32 *MOT: yeah he's playing grandpa , he's doing what granpa does (.) what's that?
33 *CHI: eh .
34 *MOT: what's that John , what is it?
35 *CHI: this this tortuga .
36 %xpho: in English
37 %eng: turtle
38 *MOT: sí , tortuga , tortuga .
39 *MOT: tortuga .

En las primeras catorce líneas se hace una descripción del corpus: ubicación, lenguas, participantes, fecha de recolección de la muestra, situación y actividades. De las líneas 15 a 24 el niño usa algunas expresiones en inglés. En la línea 25 la madre pregunta, utilizando la alternancia de inglés y español: *it's a nene, no?* El niño responde también con alternancia, esta vez con términos en las dos lenguas pero manteniendo la sintaxis inglesa: *nene doggie.* Después en la línea 34 la madre pregunta en inglés: *What's that John, What is it?* A lo que el niño responde alternando las dos lenguas pero sin utilizar el verbo copulativo (*be*): *this tortuga*. La madre le pide que utilice la palabra en inglés y el niño responde *turtle*. Este simple ejemplo sirve para mostrar el gran potencial de investigación sobre adquisición de la lengua que puede desarrollarse a partir del material registrado en este corpus.

Otra clasificación importante de corpus tiene que ver con el hecho de que incluya solamente textos escritos de la lengua o si también comprende registros

orales. Es claro que resulta mucho más expedito el trabajo con corpus escritos que con corpus orales, puesto que el procesamiento de la información oral, con todas las marcaciones de énfasis, acento, entonación, pausa, etc., constituye un reto para el lingüista computacional. De una parte se debe resolver el problema de la transcripción de todos estos fenómenos peculiares del habla y, de otra parte, se deben diseñar herramientas que permitan recuperar fragmentos específicos según el interés de los investigadores usuarios del corpus. A este respecto, hay algunos problemas que no se han podido solucionar satisfactoriamente. Por ejemplo, determinar cuál es la mejor manera de mostrar en la ventana de trabajo que hay una superposición del registro de dos interlocutores que hablan simultáneamente.

Existe otro tipo de corpus que se caracteriza por la enorme cantidad de textos que recopila, los cuales no necesariamente son balanceados. Se trata de los corpus dinámicos o monitor, que se caracterizan por permanecer abiertos a seguir aumentando de tamaño conforme se ingresen más textos. Un ejemplo de estos corpus son aquellos que se conforman a partir de periódicos o revistas, que incrementan su tamaño con el paso del tiempo. Tienen la ventaja de ser dinámicos, pero se paga un alto precio por esta apertura constante a la modificación: no hay forma de balancear la información que se ingresa. En otras palabras, habrá léxico y determinadas estructuras que están mejor representadas que otras en el corpus.

Ahora bien, hay consenso en señalar que los corpus electrónicos comienzan con el corpus Brown, confeccionado por Nelson Francis y Henry Kučera hace cinco décadas (en 1964). Se trata de un corpus sincrónico que consta de 500 muestras, cada una de 2000 palabras, con aproximadamente un millón de palabras del inglés escrito en Estados Unidos. El corpus incluye prosa informativa (prensa, religión, habilidades y pasatiempos, cultura popular, biografía y ciencias) y prosa de ficción (general, misterio y detectivesca, ciencia ficción, aventura, romance y humor) (cf. Kennedy, 1998, pp. 24-26). La contraparte británica del corpus Brown, confeccionada entre 1970 y 1978, se conoce como el corpus Lancaster-Oslo/Bergen (LOB). También contiene 500 textos, cada uno de unas 2000 palabras. La estructura de estos dos corpus se replicó en la elaboración de los corpus de las variedades del inglés de la India (*Kolhapur Corpus of Indian English*, Shastri, 1988), Nueva Zelandia (*Wellington Corpus of Written New Zealand English*) y Australia (*Australian Corpus of English/Macquarie Corpus of Written Australian English*).

Dentro de las principales herramientas que se pueden aplicar a los corpus electrónicos se encuentran los concordadores (*concordance tools*) que desde la confección de los primeros corpus electrónicos permiten realizar búsquedas de unidades o

combinaciones léxicas muy rápidamente. En estos casos, el concordador muestra la(s) palabra(s) de la búsqueda alineada en la parte central de la pantalla, es decir, la palabra clave en contexto; en inglés se denomina *Key Word in Context* (*KWIC*).

Veamos un ejemplo tomado de Tiedemann (2012, p. 2218):

1023298	want something done ...do it yourself - Come on ,	**honey**	. You can't die . Wake up! Where are the stones
3857140	s him . Where are we going , Mama ? It's a game ,	**honey**	. Like hide and seek ? Yes . Like hide and seek .
4652049	. Is he still alive ? Barely , but dont'y worry ,	**honey**	, I think I can save him . You're a very brave y
5772257	y . Oh ...And ...Aw ! Well , actually , I ate my	**honey**	. - But it mde me do it . - Humph . I was asleep
6560977	- Lau . - Yeah ? - What do you think of this one ,	**honey**	? - Please ! It needs a little swag . A scooch of
7687480	where reality es over there , somewhere ... - oh ,	**honey**	, don't , and we hide from it over here and pret
7886828	honey , and if you want honey , you'd just buy	**honey**	instead of ... apricots . Um , but nevertheless ,
8987992	r downtown. Come on. Heren they come ! This way ,	**honey**	. Oh , come on . It? s a shame to hide such a bea
9055714	oesn't interfere with couples' bowling , right ,	**honey**	? You ever bowl with Rusty ? It's a god thing .
9798978	good day . What do you want , Bruiser ? Bruiser ,	**honey**	, come on . We hace to go . We' re late . We have
10002428	he school nurse . - It's against the law . - Oh ,	**honey**	. It's a girl's best frind . - A certain kind

Figura 1. Concordancia de la palabra *honey*.

En el ejemplo anterior (Figura 1) se observan dos usos diferenciados de la palabra *honey*. De una parte, se utiliza como nombre apelativo y forma de tratamiento cariñosa ('querida', 'querido'), por ejemplo: *Come on,* ***honey***; *Oh,* ***honey***, *don't*; *This way,* ***honey***). El otro uso corresponde a la definición denotativa del término, es decir, 'miel'; por ejemplo: *I ate my* ***honey***; *you'd just buy* ***honey***.

En una segunda generación de corpus electrónicos, también conocidos como megacorpus, se distinguen, de una parte, los proyectos de corpus con una perspectiva comercial, entre los cuales cabe mencionar el proyecto *Cobuild*, también conocido como *Collins Birmingham University International Language Database*, el *Longman Corpus Network* y el *International Corpus of English* (Kennedy, 1998, p. 46). De otra parte, desde una perspectiva no comercial, de libre acceso a los corpus, debe mencionarse el proyecto de Mark Davies, de la Brigham Young University (*corpus.byu.edu*), el cual incluye los siguientes corpus: *Global Web-based English* (*GloWbE*) 1.9 mil millones/palabras; *Corpus of Contemporary American English* (*COCA*) 450 millones/palabras; *Corpus of Historical American English* (*COHA*), 400 millones/palabras; *TIME Magazine Corpus*, 100 millones/palabras; *Corpus of American Soap Operas*, 100 millones/palabras; *British National Corpus* (*BYU-BNC*), 100 millones/palabras; *Strathy Corpus* (*Canada*), 50 millones/palabras; *Corpus del Español*, 100 millones/palabras; *Corpus do Português*, 45 millones/palabras.

Así mismo, debido al multilingüismo predominante, en Europa se han confeccionado corpus multilingües que permiten, las más de las veces, acceso gratuito

a estos recursos, bases de datos y corpus. Cabe mencionar, por ejemplo: *European Language Resources Association* (*ELRA*), *Europarl*, *OPUS-Korpora* y *Oslo Multilingual Corpus* (*OMC*).

Veamos un ejemplo del corpus multilingüe OPUS, tomado de Tiedemann (2012, p. 2216):

183554493	Hey , **honey** , we're coming !
es	Oye , querida, ya vamos !
pt	Querida , estamos a chegar !
sv	Vi kommer nu !
18615253	You okay , **honey** ?
es	¿ Estas bien , cariño?
pt	Estás bem querida ?
sv	Är du okej , älskling ?
18615685	It's okay , **honey** .
es	Está bien , cariño .
pt	Está tudo bem , querida .
sv	- Det blir bra , det blir bra , älskling .

Figura 2. Diversos usos de *honey* en un corpus multilingüe.

Este ejemplo (Figura 2) corresponde a un corpus de subtítulos originalmente en inglés, con sus respectivas traducciones alineadas en español, portugués y sueco. Resulta interesante analizar la forma como se han establecido las traducciones. En la primera expresión del original (*Hey, **honey**, we're coming!*) se llama la atención del interlocutor a través de una interjección y un nombre apelativo (*Hey, honey*), lo cual solo se reproduce en español (*Oye, querida*), en tanto que en portugués se pierde la interjección (*Ø Querida*) y en sueco no aparece ni la interjección ni el nombre apelativo (*ØØ Vi kommer nu!*). Desde el punto de vista pragmático, en el original se concreta la intención de 'llamar la atención del interlocutor' mediante un acto de habla explícito (*Hey, **honey***), mientras que en las traducciones de portugués y sueco queda 'abierto' este llamado de atención al no haber un interlocutor claramente reconocible. En la segunda expresión resulta de interés la agramaticalidad estructural del original que carece de forma verbal (*You okay, **honey***), que se explica por tratarse de un registro conversacional que permite en inglés, en casos como este, la omisión del verbo. No obstante, las tres traducciones han reproducido la forma verbal necesaria, incluso en el registro conversacional de estas lenguas meta ('*Estás*', '*Estás', 'Är'*). En la tercera expresión (*It's okay, honey*), las traducciones conservan el

nombre apelativo (*cariño*, *querida*, *älskling*); sin embargo, en la traducción al sueco se introduce el equivalente idiomático de *It´s okay*, o sea, *Det blir bra*, en forma reiterada: *Det blir bra, det blir bra*.

Por otra parte, actualmente también contamos con corpus diseñados para el estudio de la lengua española. Entre ellos vale la pena destacar: *CORPES XXI*, *CDH*, *CREA* y *CORDE*. El proyecto *CORPES XXI* es una iniciativa ambiciosa, en la cual participan las diferentes academias de la lengua española y cuyo propósito es recopilar un corpus del español del siglo XXI, abarcando, en la primera fase, un periodo que se extiende de 2001 a 2012 y que aspira a completar 300 millones de palabras en esta primera etapa del proyecto. Se señala en la página electrónica de la Real Academia Española (*www.rae.es*) que se trata de un corpus general o de referencia, es decir, que incluye una amplia gama de tipos textuales, orales y escritos, con diversos registros o grados de formalidad; es un corpus anotado morfosintácticamente con lematización y tiene una distribución diatópica de entradas con 70 % procedentes de América y 30 % de España. Previamente a esta iniciativa, en 1995 la Real Academia Española había puesto en marcha la confección del *Corpus de Referencia del Español Actual* (*CREA*) y el proyecto de construcción del *Corpus Diacrónico del Español* (*CORDE*). Así mismo, también se desarrolló el proyecto del *Nuevo Diccionario Histórico del Español* (*DHE*), en dos fases: del s. XII a 1975 y de 1975 al año 2000.

Otras aplicaciones en la investigación lingüística

El carácter transversal de la lingüística de corpus permite que esta aproximación metodológica y conceptual se emplee actualmente para realizar investigación en diversas áreas de la lingüística contemporánea. Por ejemplo, McEnery et al. (2006/2008, p. 80) mencionan la aplicación de la lingüística de corpus en los estudios lexicográficos y léxicos, específicamente en la confección de diccionarios a partir de la década de los noventa del siglo pasado. Es decir, los diccionarios se basan en datos tomados de corpus. De este modo, estos diccionarios en formato electrónico pueden extraer los diferentes ejemplos para ilustrar el uso de las palabras en cuestión de segundos. Además, estas fuentes lexicográficas, como los diccionarios *Cobuild* (1995) y *Longman* (1995), presentan la frecuencia de uso de determinadas combinaciones de palabras (colocaciones). Por otra parte, el uso de corpus puede ayudar a confirmar o refutar intuiciones de los lexicógrafos que, generalmente, no son confiables.

Otra de las áreas destacadas de la aplicación de la lingüística de corpus es la confección de gramáticas. Cabe mencionar, por ejemplo, la *Longman Grammar of Spoken and Written English* (LSWE) (Biber, Johansson, Leech, Conrad, & Finegan,

1999), que se basa en el corpus Longman de inglés hablado y escrito de 40 millones de palabras e incluye una "descripción detallada de la gramática inglesa, que se ilustra con ejemplos reales del corpus y que presta igual atención a la forma como hablantes y escritores realmente usan estos recursos lingüísticos" (Biber et al., 1999, p. 40). Esta gramática se basa en un corpus que contiene cuatro registros o tipos de textos: textos académicos (ACAD), transcripción de conversaciones (CONV), textos de ficción (FICT) y textos de noticias (NEWS). Incluye, así mismo, dos variedades dialectales: inglés estadounidense (AME) e inglés británico (BRE). El contenido de la gramática también se ilustra mediante ejemplos del uso real de las variedades estándar y no estándar (vernacular) del inglés. Por ejemplo:

> They were by the pub **what** we stayed in (CONV)
> I **ain't** done **nothing** (CONV)

Esta gramática sirvió de base para la *Longman Student Grammar of Spoken and Written English* (2002), escrita por Biber, Conrad y Leech, que contiene como aspecto sobresaliente la inclusión de un capítulo final sobre la gramática conversacional (*The Grammar of Conversation*). Allí se describen las "circunstancias discursivas" y la "realización" de la conversación. Hay que destacar que, gracias al uso del registro conversacional del inglés, los aprendices de esta lengua como lengua extranjera o segunda lengua pueden comprender las peculiaridades de esta gramática conversacional que se caracteriza, entre otras cosas, por la presencia de unidades de análisis no tradicionales como las no cláusulas, la elipsis, las agrupaciones léxicas y los insertos. Esta última categoría de insertos incluye aquellas manifestaciones discursivas que en la pragmática inglesa también se conocen como nexos (*hedges*), es decir, aquellas expresiones como *uh*, *um*, *well*, *you know*, etc., que emplean los participantes en la conversación para asegurar el contacto con el interlocutor, expresar incertidumbre, duda, seguridad, etc.

McEnery et al. (2006/2008) mencionan otro campo de aplicación de la lingüística de corpus, que corresponde a la variación de registros y el análisis de géneros. A este respecto, se destaca sobre todo el trabajo de Biber (1988), quien, utilizando el análisis factorial para procesar sesenta y siete rasgos lingüísticos, logró identificar cinco dimensiones relevantes para realizar la clasificación o la tipología textual/de géneros:

1. Producción informativa o participativa
2. Discurso narrativo vs. no narrativo

3. Referencia elaborada o dependiente de la situación
4. Expresión explícita de argumentación
5. Estilo impersonal vs. no impersonal

En Biber, Conrad y Reppen (1998/2006, p. 150) encontramos un ejemplo de los rasgos positivos de la primera dimensión, es decir, producción participativa en una conversación formal:

B: come in . come in – ah good morning
A: good morning
B: you're Mrs Finney
A: Yes I am
B: how are you – my names Hart and this is Mr Mortlake
C: how are you
A: how do you do .
B: won't you sit down
A: thank you –
B: mm well you are proposing . taking on . quite something Mrs Finney aren't you
A: yes I am
B: mm
A: I should like to anyhow
B: you know what you'd be going into
A: yes I do

En esta muestra de conversación formal se activan los principales rasgos positivos de esta dimensión, es decir, hay un eje de participación interactiva, cara a cara, con verbos de la esfera personal (por ejemplo, *think*, *feel*, *know*), verbos en tiempo presente, *do* como proverbo, omisión de la conjunción *that*, uso de contracciones, y pronombres de primera y segunda persona que corresponden a los interlocutores. En oposición a estos rasgos, en la misma dimensión aparece la prosa académica (Biber et al., 1998/2006, p. 149):

> Apart from these very general group related aspects, there are also individual aspects that need to be considered. Empirical data show that similar processes can be guided quite differently by users with different views on the purpose of the communication.

Los rasgos discursivos más sobresalientes que aquí se activan son el uso de nombres, frases preposicionales, adjetivos atributivos y estructuras de voz pasiva sin agente. La interpretación funcional de cada dimensión se basa en 1. El análisis de la función comunicativa más ampliamente compartida por el conjunto de rasgos copresentes que definen cada dimensión y 2. El análisis de las semejanzas y diferencias entre registros respecto de cada dimensión.

La segunda dimensión sirve para distinguir el discurso narrativo del no narrativo. El discurso narrativo, especialmente de ficción, presenta entre sus rasgos más destacados el uso frecuente del tiempo pasado, así como los referentes de tercera persona y el discurso indirecto. Por su parte, el discurso no narrativo, que se caracteriza por ser expositivo, descriptivo o conversacional, se caracteriza por la referencia frecuente al no pasado. En cuanto a la tercera dimensión, que trata sobre la referencia elaborada o dependiente de la situación, se caracteriza por la referencia muy explícita, independiente del contexto y el uso de oraciones relativas introducidas por *wh-*. En el caso de la referencia dependiente de la situación se emplean los adverbios de tiempo y lugar. La cuarta dimensión sobre expresión explícita de argumentación indica el grado en que la persuasión se realiza de manera explícita, es decir, de qué modo el punto de vista del hablante intenta persuadir al interlocutor. Finalmente, la quinta dimensión sobre el estilo impersonal se caracteriza por la copresencia de construcciones en voz pasiva y cláusulas de participio pasado en inglés. En esta dimensión predomina el estilo impersonal en los artículos de investigación sobre ecología e historia y presentan rasgos menos impersonales la ficción en general y la conversación cara a cara (Biber et al., 1998/2006, pp. 153-155).

Por otra parte, a partir de los corpus multilingües se pueden llevar a cabo estudios contrastivos y de traducción. En otro lugar (Bolaños, 2007) presento algunas de las posibilidades que ofrecen los corpus paralelos al traductor. Básicamente, el traductor puede buscar textos semejantes al que está traduciendo en la lengua de llegada con el fin de verificar el uso de la terminología correspondiente y, además, puede corroborar cuál es la estructura textual habitual de dicho tipo de texto. Además, puede valerse de traducciones en la lengua de llegada que confirmarán o no las características que se presentan en los textos semejantes, no traducciones, escritas en esta lengua. En todo caso, esta sería una estrategia posible para contrarrestar el predominio de un 'lenguaje' especial de la traducción (*translationese*), que denota peculiaridades lingüísticas que no existen en los textos escritos por hablantes maternos de la lengua origen. Un ejemplo interesante, y muy diciente del trabajo de investigación en traducción que puede realizarse

a partir de un corpus multilingüe, lo presenta Mark Shuttleworth (2014), quien investiga de qué manera se traducen, en los textos científicos, las imágenes elaboradas que aparecen en el original de la revista especializada *Scientific American,* al francés, italiano, alemán, ruso y polaco. Shuttelworth llega a la conclusión de que, en general, en estas traducciones predomina la tendencia a traducir una imagen elaborada por una menos elaborada, con la consiguiente pérdida de especificidad e intensidad del original (Shuttelworth, 2014, p. 49).

Por ejemplo:

> English: as animals move up the **evolutionary ladder**.
> German: Je **höher** in der Evolution Tiere stehen
> [The **higher** animals stand in evolution]
> (Fields 2004a, p. 61; 2004b, p. 56) (Shuttelworth, 2014, p. 39).

La expresión metafórica del original, que se basa en la metáfora de la evolución concebida como una escalera por la cual se sube, se pierde en la traducción al alemán donde simplemente se hace referencia a la posición de los animales en la evolución.

Veamos otro ejemplo:

> English: researchers **mapped** the regions of the brain
> Russian: исследователям удалось идентифицировать области головного мозга
> [researchers managed to **identify** regions of the brain.]
> (Nestler and Malenka 2004a, p. 81; 2004b, p. 52) (Shuttelworth, 2014, p. 42).

En la traducción al ruso se pierde el sentido de la metáfora del original en inglés donde los científicos son percibidos como cartógrafos del cerebro. No es lo mismo decir que ellos 'mapean' el cerebro a decir que 'identifican regiones' en él.

Conclusiones

En el sucinto recorrido realizado en este trabajo, hemos podido describir y ejemplificar algunas de las múltiples posibilidades de investigación que ofrece la lingüística de corpus tanto para la lingüística sistémica como para la lingüística aplicada. Así, por ejemplo, la LC resulta de gran utilidad para los docentes y los estudiantes que participan en los programas de lingüística o de aprendizaje de lenguas extranjeras. Se puede estudiar una gama bastante amplia y variopinta de temas de muy diversa naturaleza: desde aspectos gramaticales (por ejemplo, el uso real de los auxiliares *shall* y *will* en textos orales y escritos del inglés estadounidense, británico y de otras

variedades dialectales), pasando por el análisis de la evolución de la adjetivación en español con marcación distintiva de género (por ejemplo, cuáles adjetivos aparecen en la frase nominal acompañando los términos *hombre* y *mujer*), hasta realizar análisis crítico del discurso que permita, por ejemplo, determinar la colocación verbal más frecuente en español y otras lenguas en relación con la situación de los inmigrantes que han sido desplazados de sus países de origen o que buscan mejores oportunidades laborales para ellos y sus familias en otros lugares del orbe; o se puede estudiar la percepción que genera una declaratoria de impago de la deuda externa por parte de un país como Estados Unidos frente a un país en vías de desarrollo.

No sobra resaltar el gran potencial que tiene la lingüística de corpus para el aprendizaje de las lenguas extranjeras. Puede utilizarse como una herramienta valiosa para que, por ejemplo, los estudiantes de inglés como lengua extranjera consulten corpus de aprendices en otras latitudes y puedan contrastar si están enfrentando las mismas dificultades y la forma de superarlas. Al mismo tiempo, otra de las actividades de investigación que pueden realizar los estudiantes es corroborar si las estructuras y el léxico que aparecen en sus libros de texto efectivamente coinciden con el uso real que encuentran en los diversos corpus del inglés estadounidense o de otras variedades dialectales estándar. Es decir, contrastar la teoría gramatical y los ejemplos que la ilustran en los libros de texto con la realidad del uso de la lengua.

Este gran potencial de utilización de la lingüística de corpus en los procesos de investigación en las diversas disciplinas de la lingüística sistémica (léxico, sintaxis, morfología) y las interdisciplinas del lenguaje (análisis del discurso, traducción, enseñanza de lenguas, pragmática, sociolingüística, etc.) que hemos descrito e ilustrado en este trabajo no debe llevarnos a pensar que no tiene limitaciones. McEnery et al. (2006/2008, p. 121), por ejemplo, mencionan cuatro aspectos problemáticos. En primer lugar, el hecho de que los corpus no suministran pruebas negativas; no nos dicen lo que es posible y lo que no es posible. Este reparo puede ser importante y relevante solamente si el corpus no contiene suficientes metadatos, es decir, si no se identifica con claridad el tipo y la procedencia textual, puesto que estos datos arrojan información importante sobre el grado de informalidad discursiva y, según la extensión del corpus, el grado de representatividad que tiene. El segundo reparo es más serio. Se refiere al hecho de que el corpus suministra información pero no explica lo observado. Aquí puede decirse que el investigador o el usuario del corpus parten de suposiciones o hipótesis acerca de lo que van a encontrar en el corpus. Luego, la explicación proviene de la confirmación o refutación de esa base teórica inicial. Si, por el contrario, el investigador no tiene una idea preconcebida

acerca de lo que puede encontrar en el corpus y lo explora 'ingenuamente', lo que allí encuentre requerirá un procesamiento teórico-hipotético posterior de su parte. En cualquier caso, la lingüística de corpus no tiene, en principio, una pretensión explicativa, por lo tanto es inútil buscarla allí. El tercer punto parece más una constatación que un reparo y se refiere al hecho de que el uso de corpus como metodología también define los límites del estudio. Los autores citados reiteran que la utilidad del corpus en los estudios del lenguaje depende de la pregunta que se formule. Esto es cierto. Sin embargo, en el caso de estudios sobre la estructura discursiva, por ejemplo, es indudable que incluir en el corpus solamente una parte y no la totalidad de los textos limita las conclusiones que se extraigan del análisis de dicho corpus. En otras palabras, si, por ejemplo, se excluye sistemáticamente del corpus la parte final de las narrativas donde aparece la coda se perderá información estructural fundamental en dicho estudio.

Tal como lo señalamos al comienzo de este artículo, las áreas más y mejor desarrolladas de la lingüística de corpus tienen que ver con los estudios léxicos, mediante el uso de etiquetadores (*taggers*) y analizadores sintácticos (*parsers*) (focalización local). Queda por desarrollar aún el diseño de herramientas que permitan etiquetar adecuadamente el componente semántico, pragmático y discursivo del corpus, de modo que se pueda aprehender su dimensión holística. A pesar de ello, esperamos haber podido mostrar que con las herramientas y los corpus existentes, de hecho, ya es posible realizar investigación interesante e innovadora en lingüística sistémica y del uso del lenguaje desde sus diversas aproximaciones.

Referencias

Biber, D. (1988). *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Reppen, R. (1998/2006). *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* London: Longman.

Biber, D., Leech, G., & Conrad, S. (2002). *Longman Student Grammar of Spoken and Written English.* London: Longman.

Bolaños, S. (2007). Source Language Text, Parallel Text, and Model Translated Text. A Pilot Study in Teaching Translation. *Forma y Función*, *20*, 225-252.

Helbig, G. (1986). *Entwicklung der Sprachwissenschaft seit 1970*. Leipzig: VEB Bibliographisches Institut Leipzig.

Jurafsky, D. & J. Martin. (2008). *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd ed.). New York: Prentice Hall.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics.* London & New York: Longman.

Lemnitzer, L., & Zinsmeister, H. (2006/2010). *Korpuslinguistik. Eine Einführung. 2. Auflage.* Tübingen: NarrFrancke Attempto Verlag.

McEnery, T., Xiao, R., & Tono, Y. (2006/2008). *Corpus-Based Language Studies. An Advanced Resource Book.* London & New York: Roudledge Applied Linguistics.

Meyer, Ch. (2004). *English Corpus Linguistics. An Introduction*. Cambridge: Cambridge University Press.

Moreno, A. (1998). *La lingüística computacional. Introducción a los modelos simbólicos, estadísticos y biológicos.* Madrid: Editorial Síntesis.

Rocha, M. (1997). Supporting anaphor resolution with a corpus-based probabilistic model. *Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, 54-61, Madrid, Spain.

Shuttelworth, M. (2014). Scientific Rich Images in Translation: A Multilingual Study. *JoSTrans*, *The Journal of Specialized Translation*, *21*, 35-51.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. En *Eighth International Conference on Language Resources and Evaluation* (pp. 2214-2218). May, Istanbul, Turkey.

Van Dijk, T. A. (1980). *Estructuras y funciones del discurso.* Madrid: Siglo XXI.