RLA. Revista de Lingüística Teórica y Aplicada
Concepción (Chile), 46 (1), I Sem. 2008, pp. 93-119.

CL ISSN 0033 - 698X

LINGÜISTICA DE CORPUS: UNA INTRODUCCION AL AMBITO*

CORPUS LINGUISTICS: AN INTRODUCTION TO THE AREA

GIOVANNI PARODI[1]
Pontificia Universidad Católica de Valparaíso
gparodi@ucv.cl

RESUMEN

Este artículo aborda críticamente un modo relativamente reciente de hacer lingüística, esto es, la denominada *Lingüística de Corpus* (LC). Para ello se presenta una breve panorámica de su (re)surgimiento y se revisa un encuadre teórico que justifica su independencia como una metodología de investigación en lingüística pero con poderosos principios reguladores. También propongo mi propia definición de LC. Se discute el posible estatus de la LC como una teoría. Hacia el final del artículo se pone en relación a la LC y algunas investigaciones en lengua española.

Palabras claves: Lingüística de Corpus, metodología, texto, corpus.

ABSTRACT

This article focuses critically on a relatively new way of doing linguistics, that is, *Corpus Linguistics* (CL). To do so, a general overview of the (re)emergence is presented and a theoretical background is analyzed in order to justify its independence as a linguistic research methodology but with robust principles. Also, my own definition of CL is proposed. A discussion of the possible status of CL as a theory is focused. To the end of the article, CL and Spanish language research is related.

Keywords: Corpus Linguistics, methodology, text, corpus.

Recibido: 15-11-2007. Aceptado: 17-03-2008.

* Este artículo se ha elaborado en el marco del Proyecto Fondecyt Nº 1060440.

[1] Todas las traducciones son responsabilidad del autor, con el fin de poner a disposición de un público más amplio trabajos exclusivamente en lengua inglesa.

1. INTRODUCCION

ES MUY cierto que en lingüística y en sus interdisciplinas se suele enfrentar algunas complejidades no siempre explicitadas totalmente para un novato que se acerca por primera vez a esta área científica. Seguramente ello acontece también en muchas otras ciencias, aunque no es razón suficiente para que suceda en la nuestra. Cuestiones terminológicas no resueltas, vaguedades conceptuales, supuestos no declarados, coexistencia de diversos enfoques alternativos, pero con sutiles divergencias, perspectivas más teóricas en oposición a otras más aplicadas y utilitarias, etc. Todo esto se encuentra en el ámbito de las ciencias del lenguaje y cuando enfrentamos la lectura de un texto o la comprensión de una nueva área, muchas veces, algunas de estas cuestiones no se declaran abiertamente. Ahora bien, no es propósito de este artículo hacerse cargo de estos asuntos. Por el contrario, frente a esta problemática, se busca aportar y delimitar un terreno relativamente novedoso y tratar de que los problemas apuntados no sean temas aquí también recurrentes. Por ello, en este artículo se busca entregar una aproximación a la hoy denominada *Lingüística de Corpus,* pero intentando advertir y entregar al lector algunas pistas explícitas de diversa índole. A modo de inicio, hacemos nuestra la siguiente idea de Malinowski (1935:9):

> La negación de lo obvio ha –a menudo– resultado fatal para el desarrollo del pensamiento científico. La falsa concepción del lenguaje como un medio de transfusión de ideas desde la cabeza del hablante hacia la del oyente ha viciado ampliamente, en mi opinión, el enfoque filosófico del lenguaje. La opción propuesta aquí no es exclusivamente académica: nos impele a, como veremos, a correlacionar otras actividades, a interpretar el significado (texto); y esto quiere decir un nuevo escenario para el manejo de la evidencia lingüística. También nos empuja a definir el significado en términos de experiencia y situación.

Las palabras de Malinowski, expresadas más arriba, enfocan un cambio de mirada que se consideraba por ese entonces necesario. Ellas nos sirven unos cuantos años después para dar inicio a este artículo y nos proporcionan un marco para los temas que defendemos aquí.

De este modo, a la luz de estas ideas, sostengo que los avances en las ciencias del lenguaje y sus interdisciplinas deben beneficiarse del uso adecuado de las evidencias empíricas provenientes de diversas fuentes (protocolos de verbalización, textos originales, elicitación de datos, técnicas estadísticas, mecanismos introspectivos, etc.); aún más, mayor robustez se conseguirá si se emplea más de un medio de aproximación al fenómeno en indagación. La información concurrente recolectada así fortalece y provee resultados certeros que justifican el desarrollo acumulativo del conocimiento científico. Desde esta perspectiva, es altamente relevante señalar que el empleo de los corpus como fuente de evidencias no es necesariamente in-

compatible con ningún tipo de teoría. Asuntos, éstos, que elaboraremos más adelante, pero que resultan altamente significativos para un anclaje de arranque.

Cabe puntualizar que, en este artículo, abordo asuntos relativos tanto a los fundamentos de la LC como a sus posibilidades metodológicas y al modo en que estos cambios han afectado el devenir de los estudios lingüísticos y sus interdisciplinas. Con esta perspectiva en mente, paso revista a algunos temas centrales para la LC desde diversas escuelas de pensamiento, se perfilan las aplicaciones prácticas y se entregan definiciones operacionales tanto de la LC, de los corpus y sus características, así como también se enfrentan discusiones no necesariamente resueltas y se busca evaluar estos aportes en el marco de los desarrollos en curso.

2. LINGÜISTICA DE CORPUS: ¿UNA O MUCHAS DEFINICIONES?

El problema de definir la LC y decidir si es una teoría o una metodología ha sido debatido desde diversas aproximaciones. Se ha argumentado en uno y otro flanco. Existe amplia bibliografía que aborda este asunto (e.g. Svartvik, 1992; McEnery & Wilson, 1996; Kennedy, 1998; Stubbs, 1996, 2001; Tognini-Bonelli, 2001). Su asociación con las tecnologías informáticas ha sido una fortaleza, pero también –para otros– una debilidad como argumento para una mirada más ambiciosa de corte teórico (De Kock, 2001). Otros afirman que la LC va mucho más allá de un exclusivo rol metodológico (Tognini-Bonelli, 2001). Sin importar el eje en que se cargue la balanza, un aporte fundamental es el decidido enfoque empírico que la LC trae consigo al focalizar datos observables a modo de evidencia científica y que se almacenan como corpus electrónicos.

Ahora bien, de partida, afirmo que la LC en su versión actual constituye un enfoque metodológico para el estudio de las lenguas, el cual revela oportunidades revolucionarias para la descripción, análisis, y enseñanza de discursos de todo tipo. También brinda una base empírica para el desarrollo de materiales educativos y metodológicos de diversa índole así como para la construcción de gramáticas, diccionarios y otros, tanto de discursos generales como especializados, orales y escritos. Desde esta óptica, sostengo que la LC constituye un conjunto o colección de principios metodológicos para estudiar cualquier dominio lingüístico y que se caracteriza por brindar sustento a la investigación de la lengua en uso a partir de corpus lingüísticos con sustrato en tecnología computacional y programas informáticos *ad hoc*.

En este sentido, en mi opinión, la LC no se entiende como una rama o un área de la lingüística tal como son la fonología, la semántica, la sintaxis, sino que como un método de investigación que puede ser empleado en todas las ramas o áreas de la lingüística, en todos los niveles de la lengua y desde enfoques teóricos diferentes. Sus aplicaciones son múltiples y no limitan las posibilidades de indagación. Todo

ello implica, por una parte, que la LC no opera como un enfoque metodológico extremadamente restrictivo, pues de ser así, se impediría cierta diversidad de opciones en el estudio de las lenguas particulares. Sin embargo, y como veremos en el desarrollo de este artículo, adscribir a la LC también involucra un cierto modo de aproximación específica a los datos lingüísticos, ya que subyacen a este enfoque determinados principios fundamentales que lo tiñen de un grado de singularidad.

Tal como propongo, la LC se define, *strictu sensu*, como una metodología para la investigación de las lenguas y del lenguaje, la cual permite llevar a cabo investigaciones empíricas en contextos auténticos y que se constituye en torno a ciertos principios reguladores poderosos. Desde este enfoque, se estudia información lingüística original y completa, compilada a través de corpus, dado que desde la LC no se apoya la indagación de datos fragmentados, inconexos o de textos incompletos, sino que de unidades de sentido y con propósitos comunicativos específicos.

Como se dijo, desde esta opción metodológica se puede explorar cualquier área o dominio de la lingüística y/o de los niveles del sistema de la lengua, pero desde una concepción particular de corpus (la cual abordaremos un poco más adelante). En este sentido, la LC aporta al estudio de corpus textuales digitales preferentemente de tamaño amplio y con soporte en tecnologías computacionales de variada índole, con énfasis en una aproximación empírica, basada en amplios conjuntos de datos reales y mayoritaria, pero no exclusivamente, con apoyo de técnicas estadísticas.

De lo dicho hasta aquí, una cuestión se detecta como de alta relevancia. Aunque tengo claro que la LC no reúne requisitos fundamentales como para constituir plenamente una teoría del lenguaje en sí misma, cabe señalar que el concepto de lenguaje que detente cada investigador dará sustento epistemológico a la versión más específica de LC a la que se adhiera. Si bien es cierto que sostengo que la LC es un enfoque metodológico, lo es para el estudio de un objeto cuya naturaleza se vincula directamente con la metodología empleada. Por ello, mi propia visión de la LC la hace de suyo interdisciplinaria pues asumo una postura cognitiva, mentalista y socioconstructivista del lenguaje y, por ende, el estudio de una lengua particular (como el español) se enmarca en esta opción.

La visión que defiendo acerca de la LC estimo posee un carácter original dado que se enfoca desde una concepción interdisciplinaria del lenguaje humano como es la desarrollada por los miembros de la Escuela Lingüística de Valparaíso: www.linguistica.cl (Peronard & Gómez, 1985; Peronard, Gómez, Parodi & Núñez, 1998; Parodi, 2003, 2005a, 2007a). En parte, a través de esta opción, busco explícitamente deslindar la nuestra de otras visiones excesivamente descriptivistas e inmanentistas (en especial de aquéllas con sesgos conductistas) y también de otras demasiado idealizadas del lenguaje humano. Todo ello con el fin de hacer sentir de modo certero el interés por los textos reales en uso y la variabilidad inherente a ellos y a las situaciones y contextos de su producción. Algunos de estos aspectos

resultaron descuidados desde los estrechos límites del estructuralismo saussureano y del generativismo chomskiano, debido –en parte– a que el uso de la lengua (*parole* o actuación, según corresponda) era considerado demasiado cambiante e impredecible y, por consiguiente, inadecuado como objeto de ciencia. Desde la LC, con el despuntar del medio siglo XX, son muchos los lingüistas que anhelan indagar el uso lingüístico, tal como es producido, comunicado y comprendido entre hablantes/escribientes y oyentes/lectores reales y en situaciones concretas y particulares.

Esta dimensión interdisciplinaria y vanguardista que propongo no será necesariamente compartida por todos los adherentes a la LC, ya que existen quienes propugnan una postura empiricista extremadamente radical en que los corpus sólo deben ser objeto de análisis en sí mismos, desligados de sus productores y comprendedores, no permitiendo así el uso de categorías provenientes de otras esferas del conocimiento. A este tipo de LC es justamente a la que aludía en los párrafos precedentes. Tal es el caso de Teubert (2005:5), defensor de una LC, en mi opinión, muy radical y antimentalista:

> Los conceptos y categorías derivadas del estudio introspectivo del lenguaje o de modelos provenientes de otras disciplinas (por ejemplo, computación) pueden no ser apropiados para la descripción de la información lingüística auténtica.

En esta línea, el mismo Teubert (2005:6), en relación al significado contenido en un texto, apunta que:

> El significado está en el discurso. Una vez que preguntamos por el significado de un segmento textual, sólo encontraremos la respuesta en el discurso, en los segmentos textuales anteriores que ayudan a interpretar este segmento, o en una nueva contribución que responda a nuestra pregunta. *El significado no concierne al mundo fuera del discurso.* No existe relación directa entre el discurso y el 'mundo real'. Depende de cada individuo conectar el segmento textual a sus experiencias en primera persona […] Cómo tal conexión funciona, está fuera del alcance del lingüista de corpus. (La cursiva es nuestra).

Sin lugar a dudas, nuestra concepción de la LC no pretende tal nivel de radicalismo ni empirismo extremo. Tampoco coincidimos con la visión de texto/discurso que sostiene tal propuesta, pues nuestra opción es decididamente interdisciplinaria, cognitivista/mentalista (lo que no implica adherir a un innatismo radical) y desde una mirada psicosociolingüística del discurso (Parodi, 2003, 2005a y b, 2007a). Siguiendo las ideas de Teubert (2005), no parece posible –en mi opinión– aceptar que la LC pueda operar a partir de un objeto de estudio tan restringido y circunscrito como el que este lingüista describe y sobre una distinción entre oralidad y escritura con la que ciertamente no coincidimos:

> Para la lingüística de corpus, el significado de un texto o de un segmento textual es independiente de las intenciones de sus hablantes (su autor). La dislocación del hablante/autor de su texto distingue el lenguaje escrito (grabado) del lenguaje oral. En el lenguaje oral, el hablante está usualmente presente y si existe un fallo de comunicación, preguntamos: "¿Qué quieres decir" y no: "¿Qué significa esto? (Teubert, 2005: 6).

Por su parte, para otros científicos como Leech (1992), la LC no es un campo ni un área de estudio, sino que un terreno determinado por el foco especial en los corpus con base en metodologías radicalmente diferentes producto de la incorporación de los avances tecnológicos y de ciertas categorías prototípicas. Sinclair (1991) y Simpson y Swales (2001) argumentan que la LC es una técnica o una tecnología, cuyo fundamento es el corpus mismo y que sus consecuencias son potencialmente de consideración. La clave está en la construcción adecuada de un corpus representativo; de este modo, los resultados generados a partir de dicho corpus tendrán directa relación con la constitución de la base de datos.

Así las cosas, aunque desde mi definición la LC no constituye una disciplina lingüística ni alcanza el estatus de un nuevo paradigma científico, ella sí cuenta con principios orientadores originales y con desarrollos informáticos específicos imprescindibles y muy sofisticados.

También se debe puntualizar que la manera de entender un corpus ha evolucionado y que la explotación del mismo enfrenta desafíos y proyecciones jamás antes imaginados; sobre todo, en la posibilidad de dar pie para la construcción de nuevas teorías fundadas a partir de los datos de los corpus. Más adelante abordaremos la vertiente que propugna otro estatus para la LC: ella dice relación con la posibilidad de ser efectivamente una teoría y de constituir así un nuevo paradigma dentro de las ciencias del lenguaje y sus interdisciplinas.

Otro aspecto relevante, que buscan los trabajos desde la LC, radica en el interés por el uso y la variabilidad lingüística. Por ello, existe una fuerte tendencia a las indagaciones multirregistros y/o multigéneros en los cuales es posible establecer comparaciones entre variedades de una lengua o incluso entre lenguas (ver Parodi 2005a, 2007a; 2007b).

Una cuestión central radica en qué diferencia a la LC de la década del cincuenta y sesenta del siglo pasado y el actual modo de hacer LC o de si existe o no tal diferencia y, de existir, de qué naturaleza sería. Allí reside la clave. En este contexto, es comprensible y se constata que algunos especialistas argumenten no estar de acuerdo en lo novedoso de este enfoque y ponen de relieve que los principios fundamentales de la hoy llamada LC ya han sido utilizados por la lingüística desde hace cincuenta o más años (Caravedo, 1999). El núcleo de este argumento dice relación con que lo único novedoso de la versión actual de la LC sería el empleo de herramientas y soportes informáticos, y ello, en opinión de Caravedo (1999), sería

asunto pasajero y podría responder a modas ilusorias. En palabras de esta investigadora, la lingüística no puede depender exclusivamente de un modo de almacenar la información para así llegar a defenderse que estamos en presencia de una nueva metodología y de alcances relevantes. Confío en que, en lo ya dicho y en lo que sigue del libro, brindo argumentos que revelan que esta opinión, desde mi mirada, no es correcta.

Vale la pena consignar que el uso que aquí defiendo del término LC es, en muchos sentidos, equivalente al de Lingüística de Corpus Computacional. No obstante ello, dado que partimos del supuesto de que tanto el soporte y proceso de digitalización de los corpus como el desarrollo y empleo de programas computacionales es parte inherente a la LC, no estimo pertinente utilizar tal adjetivo postmodificador (computacional). Otra cuestión muy diferente es la denominación de Lingüística Computacional de Corpus. Así, debe quedar claro que la adscripción a una "lingüística de corpus (computacional)" no reviste los mismos principios ni compromisos que a una "lingüística computacional (de corpus)". Sin entrar en mayores profundidades, baste apuntar que la primera puede circunscribirse a un trabajo que preferentemente maneje textos digitales y se adhiera a un conjunto de principios metodológicos; mas, en la segunda opción, el centro de la mirada proviene desde la lingüística computacional propiamente dicha y puede que su material de trabajo sean corpus (obviamente digitales), pero su foco está en la construcción de modelos computacionales del lenguaje humano con el objetivo de crear gramáticas que luego puedan implementarse computacionalmente en sistemas automáticos de diversa índole (probablemente para la comprensión y producción del discurso). Por ello, en su versión más aplicada también es conocida como ingeniería lingüística o procesamiento del lenguaje natural.

3. COGNITIVISMO Y CONTEXTUALISMO: DE LA COMPETENCIA AL USO

Tal como la preocupación por el estudio de la lengua en contexto y su correspondiente variación surge de manera simultánea a partir de múltiples vertientes, no resulta aconsejable limitar únicamente la discontinuidad de los estudios de corpus a la irrupción de un movimiento lingüístico como el chomskiano. Sin duda, existe más de una razón para justificar el des-énfasis en los estudios de corpus. No obstante ello, diversos investigadores coinciden en apuntar que la lingüística generativa constituyó una influencia decisiva y hegemónica en el devenir científico de las ciencias del lenguaje, diluyendo o debilitando el desarrollo de posturas que abordaban el estudio del lenguaje desde ópticas diversas; en particular, desde opciones que no coincidían en una definición idealizada del lenguaje ni de metodologías de índole hipotético deductivo (Francis, 1979; Conrad & Biber, 2001; Chafe, 1992;

Sinclair, 1991; Leech, 1991; Kennedy, 1998; McEnery & Wilson, 1996; Moreno, 1998).

El giro racionalista cognitivo que se impone desde el generativismo tiende a opacar de cierto modo el empirismo imperante y, en algunos casos, teñido de influencia conductista. Las bases contextualistas (o también externalistas), enmarcadas en paradigmas socioculturales del lenguaje, proveían un andamiaje para la lingüística de corpus tradicional, la que comienza a enfrentar una oposición desde el nuevo escenario interdisciplinario. Ahora bien, si bien es cierto que el generativismo aportó de manera crucial en materias nucleares acerca de la naturaleza del lenguaje humano, no es menos cierto que –entre otras– la visión idealizada del lenguaje (a saber, el estudio de la competencia lingüística) mantuvo un objeto de estudio casi único y se vieron difuminadas algunas investigaciones focalizadas en el estudio del lenguaje en uso (de la *performance*) y de la investigación de la variabilidad lingüística. Ello produjo una cierta discontinuidad o pérdida de impacto de ciertas líneas de investigaciones en lingüística. Sinclair (1991:1) ilustra con claridad los efectos de lo limitado del enfoque generativista:

> Sedienta por falta de información adecuada, la lingüística languideció –de hecho– se volvió totalmente introvertida. Se hizo una moda mirar hacia adentro de la mente más que hacia la sociedad. La intuición se volvió la clave y se enfatizó la similitud de la estructura del lenguaje y varios modelos formales. El rol comunicativo del lenguaje fue escasamente mencionado.

Buscando una explicación a la falta de preocupación por el uso lingüístico, Chafe (1992) arguye que la naturaleza modular de la teoría impulsada por Chomsky, cuyo núcleo se fundamenta en que el sistema lingüístico opera de manera independiente del sistema cognitivo humano, se constituye en un impedimento al estudio del uso lingüístico. Chafe (1992: 81) afirma que:

> Una consecuencia de la visión modular del lenguaje humano es que sus adherentes no están interesados en la observación del uso del lenguaje cotidiano ya que consideran que lo más interesante acerca del lenguaje humano existe independientemente de su uso.

Del mismo modo que la hegemonía generativista desestimó el estudio del lenguaje a través de corpus de textos naturales, también evadió un enfoque de dimensiones probabilísticas.

Enfatizando esta postura, Chomsky (1969: 38) opinaba:

> Se debe reconocer que la noción de "probabilidad de una oración" es completamente inútil, sea cual sea la interpretación de este término.

Este marco histórico diluyó de cierto modo el interés por los estudios basados en corpus. Al parecer, lograron únicamente mantenerse algunos enclaves lingüísticos en ciertas universidades que no seguían los postulados chomskianos pero que, para sobrevivir, vieron reducidos sus recursos económicos y el impacto de sus investigaciones (Kennedy, 1998; McEnery & Wilson, 1996).

La sucesión de estos y otros cambios provocó una nueva manera de enfrentar la investigación científica, revitalizando el interés por los usos de las lenguas naturales y cotidianas y su inherente variabilidad. Esta renovada mirada alternativa nos enfrenta al renacimiento del empirismo, pero no necesariamente bajo la influencia de la lingüística estructural de corte behaviorista ni de la psicología conductista imperantes en los años cincuenta. Desde nuestra opción, propugnamos un empirismo moderado que se vincula con una perspectiva mentalista del lenguaje; hecho que, como ya se ha enfatizado, tampoco implica adherir a un innatismo extremo. Así, la oposición entre métodos basados en el conocimiento (Church & Mercer, 1993) y métodos empiristas, tal como la oposición entre una llamada "lingüística del sillón" versus una "lingüística de corpus" (Fillmore, 1992), son distinciones dicotómicas que ya no tienen cabida ante las visiones inter y transdisciplinarias, en donde se propende hacia integraciones y colaboraciones más eficientes entre los distintos ámbitos de las ciencias.

Todo esto implica que la LC no está exclusivamente comprometida con una aproximación analítica cuantitativa, sino que una mirada cualitativa de los hechos lingüísticos es perfectamente posible y una integración entre ambos tipos de análisis resulta más que saludable y oportuna, siendo muy posiblemente el aporte en su conjunto lo que enriquezca el análisis; obviamente, dependiendo de las decisiones de cada investigador. Por supuesto, todo ello no impide la existencia de posturas extremadamente radicales, por un lado, en uno y otro polo de una opción deductivista o inductivista y, por otro, entre un análisis exclusivamente cuantitativo o cualitativo.

4. EL CORPUS COMO HERRAMIENTA DE INDAGACION: ALGUNAS DEFINICIONES

Como todos sabemos, explicitar una definición operacional de un concepto dado, muchas veces es una tarea compleja. La LC no está exenta de ello. Existen complejidades de diversa índole que tienen que ver con énfasis, variables a considerar, y –por supuesto– opciones epistemológicas. Algunas de estas complejidades residen, por ejemplo, en el criterio de clasificación de los corpus; en si se enfoca un corpus electrónico, un corpus en papel, un corpus diacrónico, un corpus representativo, un corpus oral, un corpus ejemplar, un corpus estratificacional diversificado, un corpus de referencia, un corpus en paralelo, o un corpus incremental, etc.

Una revisión bibliográfica somera permite comprobar la heterogeneidad de aproximaciones. Por ejemplo, Leech (1991, 1992) sostiene que un corpus computacional se constituye en un fenómeno nada excitante, pues resulta ser sólo una gran cantidad de textos almacenados en un computador. En este sentido, de modo algo simplista, Leech enfatiza la idea de que este tipo de corpus podría ser sólo una gran cantidad de textos con cierto formato.

> ... un corpus computacional es un fenómeno nada excitante: un helluya enorme de textos, almacenados en un computador (Leech, 1992: 106).

A pesar de ello, este mismo investigador reconoce que son las máquinas y este tipo de corpus digitales los que permiten realizar operaciones computacionales sobre cantidades masivas de textos, cosa impensable años atrás. En palabras de Leech (1991: 13):

> [...] la amplia disponibilidad de recursos de corpus computarizados ha permitido a los fenómenos sintácticos y léxicos de una lengua abrirse a la investigación empírica en una escala inimaginable.

Por su parte, Sinclair (1991: 171) sostiene que un corpus es:

> [...] una colección de textos de ocurrencias de lenguaje natural, escogidos para caracterizar un estado o una variedad de lengua.

Esta anterior definición se aprecia enriquecida en algunos aspectos en la propuesta de Crystal (1991: 32):

> Una colección de datos lingüísticos, ya sea de textos escritos o de transcripciones de habla grabada, los que pueden ser utilizados como punto de partida para descripciones lingüísticas o como un medio de verificación de hipótesis acerca de una lengua.

En particular, las alusiones directas a la escritura y a la oralidad, en especial a esta última modalidad de la lengua, enfrentan complejos desafíos para alcanzar un nivel sofisticado de transcripción y etiquetaje enriquecido a través del cual se dé cuenta de aspectos vitales para las interacciones orales, por ejemplo, los suprasegmentales. Dentro de este panorama, una definición posiblemente más rica y afinada es la que aporta, en el marco de un proyecto de la Unión Europea, el *Expert Advisory Group on Language Ingineering Standards* (EAGLES). El grupo EAGLES realiza recomendaciones o propuestas de estandarización con el fin de coordinar los trabajos que se realizan en las diferentes lenguas de Europa. Para ello, evalúa métodos y sistemas existentes y a partir de estos análisis realiza sus propuestas. El

proyecto a cargo del EAGLES busca la armonización de los recursos lingüísticos en diferentes lenguas europeas. EAGLES no pretende, por lo tanto, producir un etiquetario morfosintáctico, sino más bien entregar directrices que ayuden en el desarrollo de uno. Se ha propuesto, por ejemplo, tres criterios orientadores: a) flexibilidad, b) apertura teórica y c) búsqueda de consensos.

En esta línea de acciones, para EAGLES, un corpus es:

> una colección de partes de una lengua que son seleccionados y ordenados de acuerdo a explícitos criterios lingüísticos, con el fin de ser empleados como ejemplos de esa lengua [……] Un corpus el cual es codificado de un modo estandarizado y homogéneo para responder a tareas específicas de recuperación (EAGLES, 1996)

Un breve análisis de esta propuesta permite detectar, al menos, tres aspectos relevantes: 1) un corpus debe estar compuesto por textos producidos en situaciones reales, 2) la recolección de estas instancias de lengua en uso debe estar guiada por parámetros explícitos que permitan tener claridad de la constitución de las mismas, de modo que se apoyen en el análisis y se posibilite la replicabilidad en estudios posteriores, y 3) un corpus (aunque dicho de modo implícito) debe estar disponible en formato electrónico con el fin de ser analizado por medio de programas computacionales.

Buscando apoyar la construcción de corpus, EAGLES (1996) propone algunas recomendaciones para que un corpus pueda considerarse como tal:

1. El corpus debe ser lo más extenso posible de acuerdo con las tecnologías disponibles en cada época.
2. Debe incluir ejemplos de amplia gama de materiales en función de ser lo más representativo posible.
3. Debe existir una clasificación intermedia en los géneros entre el corpus en total y las muestras individuales.
4. Las muestras deben ser de tamaños similares.
5. El corpus, como un todo, debe tener una procedencia clara.

Del mismo modo, Biber, Reppen, Clark & Walter (2001) proponen cuatro ventajas para adoptar una aproximación basada en corpus:

1. Adecuada representación del discurso en su forma de ocurrencia natural en muestras amplias y representativas a partir de textos originales.
2. Procesamiento lingüístico (semi)automático de los textos mediante el uso de computadores. Ello permite análisis más amplios y profundos de los textos mediante conjuntos de rasgos lingüísticos caracterizadores.

3. Mayor confiabilidad y certeza en los análisis cuantitativos de los rasgos lingüísticos en grandes muestras de textos.
4. Posibilidad de resultados acumulativos y replicables. Posteriores investigaciones pueden utilizar los mismos corpus u otros pueden ser analizados con las mismas herramientas computacionales.

Como se desprende, existe cierta coincidencia entre lo propuesto por EAGLES (1996) y Biber *et al.* (2001). Aunque Biber *et al.* (2001) también apuntan claramente hacia rasgos de la constitución de un corpus, se detecta que ellos buscan afianzar una perspectiva metodológica más particular, cual es la de los estudios multidimensionales y multirregistros (Biber & Tracy-Ventura, 2007).

Considerando lo hasta aquí discutido, es factible detectar tensiones en cuanto al concepto de corpus. Ya sea si éste debe ser necesariamente uno de tipo digital o si aun es factible pensar en un conjunto de textos en papel. También se hace evidente que el asunto de la extensión cobra importancia. Seguramente se dirá que ello depende en gran medida de los objetivos de la investigación. Sin duda, ello es altamente relevante; no obstante, si se busca un proceso de investigación sinérgico con resultados de índole acumulativa y posibilidad de replicación, resulta indudable que se debe adherir a la mayoría de las indicaciones propuestas.

En mi opinión, al menos, se pueden identificar ocho características relevantes, llegado el momento de construir y comprender los alcances de un corpus. Ellos se listan a continuación sin mediar ningún sesgo jerárquico. Como es obvio, este conjunto no está cerrado ni pretende estarlo:

1. Extensión
2. Formato
3. Representatividad
4. Diversificación
5. Marcado o etiquetado
6. Procedencia
7. Tamaño de las muestras
8. Clasificación y adscripciones de tipos disciplinar, temático, etc.

No abordaremos puntualmente aquí cada uno de estos aspectos pues estimo que ellos han sido o serán comentados a través de este trabajo. Sólo los entrego a modo de resumen de los principios a tener en cuenta, en parte, como se dijo, dependiendo de los objetivos de cada investigador y de las posibilidades tecnológicas al alcance. No obstante ello, en lo revisado hasta aquí del concepto de corpus, una característica se hace recurrente y reviste ciertas complejidades: aquella denominada representatividad. Es bien sabido que incluso los grandes corpus no logran dar cuenta de la lengua como un todo ni tampoco se pretende que así sea. La

lengua en su dinamismo y heterogeneidad es mucho más rica de lo que se puede imaginar y no logra ser captada en un solo corpus, por gigantesco que sea su tamaño. Tal como apunta acertadamente Leech (2002), un corpus puede ofrecer información detallada acerca de una lengua particular, pero es imposible recolectar un corpus que abarque toda una lengua. Si ese fuera el caso, sería necesario recolectar todos los usos de dicha lengua. De este modo, se debe siempre tener presente que un corpus es sólo una colección finita de un universo infinito.

Por ello, el desafío de contar con un corpus representativo de una variedad determinada de lengua –incluso de un único registro específico de tal o cual lengua– es una cuestión compleja debido a la enorme diversidad y variedad inherente a cada lengua particular. En cuanto a la llamada representatividad estadística, Biber (2005) entrega lineamientos y alternativas en la construcción de un corpus con atención a este asunto, pero –en mi opinión– sólo aplicable desde ciertas perspectivas metodológicas. Muy posiblemente muchos de los investigadores en LC, y contrariamente a lo que sostiene Biber (2005), no buscan dotar a sus corpus de un carácter representativo, así entendido desde la metodología de la investigación científica y desde los principios estadísticos de representatividad (Hernández, Fernández & Baptista, 2003; Hair, Anderson, Tatham & Black, 1999). En este sentido, en lingüística, el universo de estudio (en el giro técnico) no es en muchas investigaciones fácilmente determinable ni calculable, por ende tampoco lo es la población o muestra estadísticamente representativa que de él se desprende. Por ejemplo, esto se aplica al trabajo con los corpus orales correspondientes, digamos, a una ciudad, cuyo universo no resulta del todo fácil de estimar. Es muy cierto que se podría determinar el tipo y cantidad de hablantes por estratos específicos, pero otra cosa es decidir el tamaño de cada entrevista, de cada grabación o de cada muestra textual. En otras palabras: ¿cuántas horas de entrevistas son necesarias para alcanzar la representatividad estadística del discurso oral en un registro específico de los hablantes de una ciudad cualquiera? Ciertamente es un asunto de complejidades. Algunos podrían decir que no existe límite. Otros pueden sostener que se deben hacer opciones y definir claramente los parámetros, variedades y estratos a abordar. Esto último es, sin duda, una salida posible.

Al respecto, cabe señalar lo que sucede en el caso de la investigación de que se da cuenta en este artículo. De cara al estudio del discurso especializado, se recolecta el total de textos escritos que circulan en una institución de educación durante un período formal de estudio sistemático. En otras palabras, el corpus está compuesto por el universo de los textos que reciben como lectura obligatoria y complementaria los alumnos de determinadas áreas técnico-profesionales como parte del currículo de formación. Este corpus constituye así el universo de indagación y en base a él, sí es factible determinar estadísticamente una muestra representativa. Por supuesto que éste no es siempre el caso en investigaciones lingüísticas.

Otra opción es que, más bien, se busque una proporcionalidad adecuada del corpus y que ello conduzca a solo ciertas proyecciones. Por supuesto que no será posible realizar generalizaciones, como desde otros modelos estadísticos inferenciales. Así, queda claro que las indicaciones de Biber (2005) son prudentes, pero sólo logran encontrar acogida en cierto tipo de investigaciones cuantitativas que logren, por ejemplo, determinar previamente, en base al universo estudiado, su corpus de análisis.

4.1. Mi definición de corpus

Propongo, en términos iniciales, que un corpus es una colección o conjunto de textos que está formado por al menos dos o más textos (dicho de otro modo, corpus aquí sería algo así como corpus textual). En este sentido, un corpus debe contener un número importante de textos que comparten ciertos rasgos definitorios, limitado sólo por características inherentes a la naturaleza de los mismos. Partiendo de estas ideas, se puede afirmar que el objetivo de la LC sería el analisis y descripción de la lengua en uso, tal como se realiza a través de texto(s). De este modo, una premisa fundamental es que los textos son el medio primario de creación y transmisión de significado. Esta amplia y algo vaga definición preliminar permite, en mi opinión que, al menos, un par de textos constituya así un corpus (acogiendo todas las posibilidades mono o multimodos o mono o multimedios). En este punto, es relevante señalar que un texto no es lo mismo que un corpus. Son diversas las comparaciones y contrastes que se puede ofrecer (Tognini-Bonelli, 2001). De modo breve, baste decir que un texto se constituye en una pieza comunicativa única y que se define por su cierre semántico y su coherencia. Un corpus, por su parte, reúne un conjunto de unidades textuales y no es una única instancia comunicativa, tampoco cuenta con cierre de ningún tipo. En este sentido, un corpus busca entregar datos acerca de la lengua en una proyección mayor que la que busca un texto como instancia de habla.

Así, unida a mi concepción de LC, mi definición de corpus corresponde a un conjunto amplio de textos digitales de naturaleza específica y que cuenta con una organización predeterminada en torno a categorías identificables para la descripción y análisis de una variedad de lengua. Este conjunto de textos debe mostrar, de preferencia, accesibilidad desde entornos computacionales y visibilidad de modo que se posibilite su uso en diversas investigaciones con el fin de asegurar acumulación de conocimientos e integración de la investigación de una lengua particular o en comparación con otra. También debe cumplir con aportar detalles relevantes acerca de su recolección y procedencia. De modo más específico, se espera se almacene en conjunto con otros corpus diversos con el fin que se permita su compara-

ción e, idealmente, su contraste. Debe quedar claro que esta definición no se aplica a casos de corpus especializados, pues se comprende que muchas veces a éstos sólo existe acceso restringido o su naturaleza misma los hace escasos y, por ende, su tamaño puede ser muy reducido.

En esta línea, entiendo que un corpus en la actualidad, de ser factible, debe cumplir algunas o todas estas características:

1. Recolección de textos en entornos naturales.
2. Explicitud de los rasgos definitorios y compartidos por los textos constitutivos.
3. Formato final de tipo digital plano (*txt.) para cada texto o documento.
4. Tamaño, preferentemente, extenso
5. Respeto a principios ecológicos.
6. Etiquetaje computacional semiautomático de naturaleza morfosintáctica u otra para cada texto.
7. Disponibilidad a través de medios computacionales.
8. Acceso a visualización completa de los textos que lo componen en formato plano.
9. Búsqueda de principios de proporcionalidad o representatividad (posiblemente estadística).
10. Sustento o procedencia inicial especificada.
11. Identificación de una organización en torno a temas, tipos de textos, registros, géneros, etc.
12. Registro de datos cuantitativos que permita la comparación y posible normalización de cifras.

Por su parte, respecto a los textos que componen un corpus, se espera que ellos preferentemente:

1. Sean unidades completas.
2. Sean de modalidad oral, escrita o de diversas variedades multimodales las cuales deberán ser identificadas en detalle.
3. Cuenten con registro del número de palabras y de oraciones que los componen.
4. Cuenten con datos de proveniencia tales como fecha, contexto de recolección, recolector, etc.

Enmarcado en estas ideas reguladoras, también estimo que un corpus debe mostrar más de alguna clasificación de la colección que recoge, ya sea de índole temática, de registro, de género o de disciplina. Idealmente un corpus debiera tender a una cierta representación, aceptando que esto encierra complejidades diversas. Adhiero a la idea de que debemos recolectar corpus muy amplios, tan extensos como sea

factible, y que la cuestión de la "saturación" no resulta muy clara ni ventajosa en este tipo de investigaciones de corte más bien cuantitativo. En mi opinión, la constitución de un corpus debería, preferentemente, contar con la posibilidad de disponer de otros tipos de corpus de naturaleza diversa en alguna dimensión. Ello permite la comparación y, de este modo, el contraste hace emerger características distintivas y prototípicas que, de otro modo, sería imposible llegar a descubrir. En este sentido, la recolección de un solo y muy focalizado corpus, por amplio que sea, no brindará una gran riqueza en su descripción, salvo que ya se cuente con otros corpus disponibles previamente y, así, la comparación emerja con mayor facilidad. O, por el contrario, que se encuadre en objetivos de investigación muy acotados por sus recolectores e investigadores; o que busque constituirse en un sentido de precorpus.

Desde esta óptica, la descripción de un corpus radica de modo importante en la búsqueda de una especificación de sus características prototípicas, las que –en mi opinión– resultan únicamente detectables de modo certero a través de la comparación y contraste con otros corpus diversos. Del mismo modo, este procedimiento también permite la determinación de similitudes y de rasgos idénticos y compartidos entre los corpus en estudio. Por ejemplo, en nuestras propias investigaciones esta cuestión emergió como un rasgo sorprendentemente clarificador, llegado el momento de caracterizar y describir un corpus de textos especializados escritos que circulaban en la educación técnica profesional chilena. Sólo logramos identificar la prototipicidad del discurso de los textos escritos especializados de esta variedad de lengua cuando los comparamos con otros diversos, tales como un corpus de literatura latinoamericana escrita (CLL) y otro de entrevistas orales semiestructuradas (CEO).

Siguiendo esta última idea, y a pesar de lo dicho más arriba, estoy cierto que existen propósitos investigativos y realidades de estudio que no necesariamente deben cumplir con todas estas exigencias. Por ejemplo, se pueden efectuar estudios de precorpus con el fin de proponer hipótesis de trabajo o con el objetivo de explorar ciertas características o categorías para una posterior recolección más amplia y robusta. Dado un corpus altamente especializado, puede que sea imposible conseguir una amplia y variada cantidad de textos de esa naturaleza, pues el universo de textos puede ser muy restringido y escaso; el estudio de textos institucionalizados o profesionales impone restricciones de índole legal y ética que complejiza una recolección amplia y ecológica y, muchas veces, sólo obliga a contar con muestras ejemplares o prototípicas (sus autores o usuarios deben respetar estrictas normas de confidencialidad con el fin de no difundir información reservada que pueda dañar a terceros). No obstante ello, es muy cierto que la tendencia actual impone unas ciertas normas o principios que nos llevan a pensar que "más es mejor" y también que "mayor diversidad asegura mayor confiabilidad en la comparación", en especial, de cara a una descripción profunda.

5. LA LC CONTEMPORANEA: ¿NUEVOS ORIGENES?

El (re)florecimiento de los estudios basados en corpus se puede fijar a comienzos de la década del sesenta, marcado, en parte, por los cambios paradigmáticos ya comentados y también afectado por la incorporación de los computadores en el ámbito lingüístico. Junto a esto se debe destacar el auge de grandes proyectos de investigación en Inglaterra y en los países escandinavos, a partir de la construcción de grandes corpus lingüísticos digitales para el inglés. Ellos constituyen el eje de avanzada de esta nueva reposición. Desde este escenario, es posible establecer, a lo menos, tres momentos relevantes.

El primero surge, como se decía más arriba, a partir de la recolección de grandes corpus de textos auténticos, además de estar ahora debidamente digitalizados y operados a través de herramientas computacionales *ad hoc*. Estos corpus incluyen una diversidad de usos lingüísticos que permiten alcanzar observaciones generales acerca de la estructura y el uso de registros tanto orales como escritos por medio de una jerarquización y organización pertinente. Como es bien sabido, estos primeros avances se desarrollan básicamente para la lengua inglesa: el corpus Brown de inglés norteamericano escrito (constituido por reportes de prensa, documentos gubernamentales y narrativa de ficción) alcanzó un millón de palabras. Complementariamente, el corpus Lancaster-Oslo-Bergen (LOB), en su versión de inglés británico, compiló un millón de palabras. Como primer desarrollo que diera cuenta de la oralidad, el corpus London-Lund incluyó quinientas mil palabras de textos orales de inglés británico, incorporando una variedad importante de diversos géneros. Un dato importante de consignar es que, en su momento, estos corpus fueron considerados como construidos "a gran escala", ya que superaban largamente el estudio de textos ejemplares o de corpus muy reducidos tradicionalmente almacenados en formato papel y organizados, muchas veces, a través de fichas.

Desde esta óptica, los requerimientos de análisis semiautomáticos y exhaustivos de textos sobre la base de herramientas computacionales (tales como etiquetadores morfosintácticos) derivó en descripciones en términos probabilísticos y llevó al desarrollo de gramáticas independientes del contexto (*context-free-grammars*). Como se sabe, desde el enfoque probabilístico, la variación es tomada como parte integral del funcionamiento lingüístico en la formulación de los mecanismos de selección, ya que ellos emergen de distribuciones observables, frecuencias relativas y correlaciones estadísticas. La probabilidad de una secuencia de palabras se determina por la suma de las probabilidades individuales de todas las estructuras. En estos términos, una gramática probabilística es muy similar a algunas gramáticas convencionales, excepto que además de asignar un conjunto de estructuras para cada secuencia de palabras de una lengua, también entrega una probabilidad para cada una de ellas (Halliday, 1992; Aarts, 1991; Stubbs, 1996, 2006). Una característica impor-

tante de las gramáticas y de los etiquetadores probabilísticos es que se van construyendo a partir de la interacción entre unos resultados preliminares y la revisión de expertos que retroalimentan los posibles problemas del sistema, de modo que el etiquetador o la gramática en cuestión se vuelve cada vez más preciso y robusto.

Un segundo giro o momento en la LC, en lo relativo a textos de orientación general, se detecta a partir de la década del ochenta. Este dice relación con la recolección de los megacorpus, los que según su nombre indica pasan a constituir dimensiones gigantescas. Ello nos lleva a mirar ahora a la denominada "primera generación de corpus digitales" y juzgarlos, desde la privilegiada mirada actual, como "de escala menor". Algunos de los megacorpus son el caso del corpus Bank of English que contiene 450 millones de palabras; el corpus Internacional de Cambridge con 100 millones de palabras; el corpus Longman del inglés oral y escrito, formado por 40 millones de palabras y el corpus Nacional Británico que alcanza 100 millones de palabras. Recientemente se encuentran en construcción algunos corpus de más de un billón de palabras, muchos de ellos compilados a partir de herramientas computacionales automáticas que utilizan la red de Internet como fuente de información.

Un rasgo que vale la pena destacar y tener presente a partir de los corpus de lo que hemos denominado como segundo giro lo constituye el hecho de que la mayoría de estos megacorpus o de muchos de los corpus actualmente en construcción contienen, a diferencia de lo que sucedía con los primeros corpus digitales, textos completos más que secciones o trozos ejemplares de textos determinados (en algunos casos se extraían sólo 2.000 palabras por texto). Sin lugar a dudas, este hecho presenta implicancias considerables para cualquier análisis posterior, pues ya no se trabaja sobre textos mutilados o parcialmente representativos sino sobre unidades reales completas. Paralelamente, también se debe tener presente que estos nuevos grandes corpus se constituyen mucho más organizada y jerárquicamente, es decir, se establecen a partir de una conjugación de diversos tipos de variables diversificadas. Por ejemplo, acogen variedades orales y escritas, formales e informales, planificadas y espontáneas, monológicas y dialógicas y, en el caso de la lengua inglesa, incorporan, al menos, variantes del inglés británico y del americano.

Como se aprecia, sólo unas pocas décadas más tarde de su florecimiento, el perfil de la LC y de los corpus generales ha experimentado una tremenda transformación, ya no únicamente en cuanto a su tamaño sino también en términos de su composición interna, tornándose ésta cada vez más precisa, diversificada y de mayor impacto y envergadura. Estos desarrollos sólo han sido posibles gracias a un avance también vertiginoso que ha corrido en paralelo al de la LC como es el de la tecnología computacional, tanto en lo que dice relación con sistemas físicos (*hardware*) como de programas computacionales (*software*). Estos impresionantes avances tecnológicos, ejecutados en un periodo brevísimo de tiempo, han posibili-

tado la construcción y almacenamiento de estas bases de datos computarizadas así como el desarrollo de sistemas de interrogación y recuperación de la información contenida en dichos sistemas.

El impacto de estos avances se refleja en la investigación focalizada en la lengua inglesa, en donde se ha explorado una amplia gama de rasgos lingüísticos a través de enormes cantidades de textos pertenecientes a variados tipos textuales (Biber, 1988; Louwerse, McCarthy, McNamara & Graesser, 2004). Todo ello ha dado origen a, entre otros, varias gramáticas y diccionarios, construidas desde los principios de la LC, las cuales reúnen y distinguen variantes regionales y usos de la lengua oral y la escrita (Quirk, Greenbaum, Leech & Svartvik, 1985; Biber, Johansson, Conrad & Finegan, 1999). Estos avances para la lengua inglesa tienden a superar –de cierto modo– la clásica tendencia en la elaboración de gramáticas con una concentración preferente sino exclusiva en el modo escrito de la lengua, con base en un único registro y/o un único género y desde enfoques eminentemente normativos.

Como se anunció, también es factible detectar un tercer giro. Este emerge debido al interés por estudiar los denominados discursos especializados. Esta variedad de discursos constituye normalmente, ya sea por su naturaleza o por otras razones, muestras relativamente pequeñas en comparación a los corpus de índole más general. Debido a que en algunas situaciones son textos escasos o a que se complica su disponibilidad por cuestiones de producción, acceso, ética o moral, su constitución suele ser reducida. Por ello, se identifica esta alternativa como un tercer giro en el cual nos movemos de los megacorpus a corpus comparativamente más pequeños, pero altamente focalizados temática, estructural y/o funcionalmente. En todo caso, cabe puntualizar que este camino paralelo no necesariamente implica que todo corpus especializado deba ser de tamaño reducido, ya que es posible también contar con corpus de naturaleza no general y de tamaño considerable (ver Parodi 2005b, 2007a y b).

5.1. Investigaciones en lengua española desde la perspectiva de la LC

La investigación reciente en lengua española ha mostrado un mayor énfasis en el uso de corpus digitales progresivamente más amplios y diversos con el fin de avanzar en descripciones lingüísticas más profundas y robustas. También se ha asentado la idea de que los principios de la LC entregan directrices empíricas eficaces para comprobar las hipótesis de los investigadores. Las distinciones entre, por ejemplo, un tipo de discurso especializado y uno de índole más general o de un tipo de registro escrito y otro oral sólo últimamente han logrado ser descritas de manera más acuciosa, aunque aún de modo preliminar (ver Parodi 2005b, 2007a y b). Desafortunadamente, ello todavía no logra materializarse en la forma de una gra-

mática del español que dé cuenta de estructuras y usos diversos de esta lengua particular y que muestre la heterogeneidad de géneros, registros y modos actuales, incluso incorporando información, por ejemplo, fonológica, prosódica o de tipo "toma de turnos", en el caso de textos orales (Leech, 2000). Tampoco se ha impactado aún en el sistema educativo y en las metodologías de lenguas, aprovechando –por ejemplo– los hoy denominados "corpus de aprendientes o aprendices" (*learner corpora*).

Ahora bien, debo aclarar que en este apartado no pretendo de modo alguno cubrir un relevamiento de las investigaciones en curso ni de los grupos que actualmente llevan a cabo trabajos dentro de los amplios marcos de los estudios de o con corpus. Comentamos sucintamente líneas iniciales y bosquejamos, *grosso modo,* la situación actual.

La investigación pionera en torno a la lengua española registra tanto en Latinoamérica como en España proyectos señeros muy relevantes como el *Proyecto de Estudio coordinado de la norma lingüística culta de la principales ciudades de España e Ibero América,* más conocido como Proyecto de la Norma Culta. Esta iniciativa, sin lugar a dudas, abrió y consolidó una oportunidad de trabajo mancomunado con investigaciones enmarcadas en principios de la LC, aunque sin los apoyos tecnológicos actuales (entre otros, Lope Blanch, 1969, 1977, 1990, 1994; Rabanales & Contreras, 1979; Oyanedel & Samaniego, 1998; Matus, 2002). También cabe destacar obras como la de Paul Garvin, *Breve introducción a la computación lingüística,* inicialmente publicada en Perú por la Universidad Mayor de San Marcos en el año 1969. En este libro se entrega herramientas y fundamentos informáticos y de lo que hoy denominamos LC para realizar trabajos en lingüística descriptiva. La obra es un compendio realizado a partir de conferencias y seminarios organizados por el PILEI (Programa Interamericano de Lingüística y Enseñanza de Idiomas) y la ALFAL (Asociación de Lingüística y Filología de América Latina) y que Garvin dictó en Montevideo, Uruguay. El texto definitivo fue revisado y editado por tan destacados especialistas como J.P. Rona, W. Mesías y A. Escobar.

Dentro de esta panorámica, aunque comparativamente de modo tardío, los estudiosos del español se han venido incorporando al campo de la LC en los términos actuales y han empleado las técnicas de recolección y construcción en cuestión. Un ejemplo interesante de acceso en línea y de modo gratuito lo constituye el trabajo que, en esta perspectiva, la Real Academia Española de la Lengua ha venido desarrollando. Ello se ha materializado en un sitio web con una interfaz de consulta de concordancias con dos corpus disponibles en línea: el Corpus de Referencia del Español Actual (CREA), que alcanza cerca de 140 millones de formas y el Corpus Diacrónico del Español (CORDE), que consta de 180 millones de formas. También cabe destacar que la RAE a través de su Departamento de Lingüística Computacional se encuentra implementando herramientas de análisis lingüístico que se espera estén disponibles en línea en un futuro próximo.

Entre otros varios grupos, un eje de acciones es el desarrollado por el Grupo Val.Es.Co en España, particularmente en cuanto a la lengua oral y registro coloquial y variedad conversacional (Briz & Grupo Val.Es.Co., 2002). También se debe destacar, entre otros, los trabajos del equipo de la Universidad de Santiago de Compostela con la Base de Datos Sintácticos del español actual (www.bds.usc.es) y del grupo del Instituto de Lingüística Aplicada de la Universidad Pompeu Fabra (http//:bwananet.iula.upf.edu). No obstante ello, existen ya una serie de bancos de datos y de recursos para el español disponibles gratuitamente en Internet, creados ya sea como iniciativas académicas institucionales y/o personales, algunos quedan registrados en la publicación del Instituto Cervantes (1996), otros en De Kock (2001) y en Parodi (2007b). Por supuesto que también destacamos nuestros propios avances en esta línea tanto en investigaciones empíricas señeras para el español (ver Parodi, 2004, 2005a, 2007a y b) como en desarrollo de tecnologías *ad hoc* (ver Parodi & Venegas, 2004; Parodi, 2007b, Venegas & Silva, 2007). En particular, resaltamos la mirada multigéneros, multirregistros y multimodos que nuestro equipo ha privilegiado desde sus comienzos, lo mismo que el impacto que ello ha tenido en tesis de pregrado, maestría y doctorado (Sabaj, 2004; Venegas, 2005; González, 2005; Silva, 2006; Gutiérrez 2007; Ferrari, 2007).

6. LINGÜISTICA DE CORPUS: ¿METODOLOGIA O TEORIA?

La pregunta que da origen a este apartado revela que, aunque pueda hasta aquí haber aportado a la discusión del debate acerca de la LC como una metodología lingüística, aún se sigue debatiendo acerca de si la LC puede alcanzar un grado de independencia tal que le permita constituirse en un nuevo paradigma. Así, si uno se posiciona exclusivamente desde el nivel de los principios metodológicos, innegablemente sus aportes son innovadores y brindan gran soporte para un número creciente de investigaciones cuyos resultados, entre otros, se capitalizan hacia la elaboración de gramáticas y materiales didácticos, la construcción de diccionarios, diversas aportaciones a la ingeniería lingüística, a las tecnologías del habla, a los sistemas de recuperación de información y también, por supuesto, para las investigaciones de interés lingüístico *per se*. Es oportuno hacer notar que la aceptación y adhesión a este enfoque metodológico, de enorme importancia, acarrea dificultades o (pseudo)problemas que conviene tener presentes pues su consideración hará más potente sus desarrollos (Rojo, 2002).

Desde una mirada más ambiciosa, si se busca posicionar a la lingüística de corpus como una teoría explicativa de, al menos, parte del funcionamiento de la mente, las exigencias son mayores. De hecho, si se concibe el lenguaje humano como una facultad probabilística (Charniak, 1996; Manning & Schütze, 1999;

Bud, 2003; Jurafsky, 2003) y se acepta el procesamiento estadístico del lenguaje natural como un modo de operar de la mente, nos encontramos frente a un paradigma emergente. Ello pues los argumentos buscan ir más allá que principios metodológicos, sino que tratan de sustentar bases epistemológicas de la forma de procesar información por el ser humano, de la naturaleza de los datos lingüísticos y de la facultad del lenguaje. Desde luego, se deberá decidir si su visión más radical, posiblemente anclada en concepciones conexionistas del cerebro, con la consecuente negación de la mente con capacidad de representación simbólica del lenguaje es una alternativa plausible. En una versión extrema de esta naturaleza, es factible que la mente podría no existir y el procesamiento lingüístico quedaría restringido a una compleja red neuronal amparada en la metáfora de múltiples sistemas vectoriales interrelacionados.

Posturas intermedias, llamadas híbridas (Kintsch, 1998), parecen encontrar por ahora mayor acogida. Aunque el modo en que relacionan representaciones proposicionales simbólicas con modelos conexionistas no está aún suficientemente explicitado (Parodi, 2003, 2005b, 2007b; Ibáñez, 2007).

Resulta entonces altamente necesario preguntarse por el concepto de lenguaje que subyace a esta postura. Desde este enfoque, la LC llevaría a comprender el lenguaje humano como un fenómeno estadístico de índole estocástico. Concordando con esta postura, Bud (2003) postula que existiría una facultad probabilística exclusiva al ser humano. Por su parte, Moreno (1998), coincidiendo en esta línea, postula que el lenguaje humano es un mecanismo computacional de carácter biológico propio al ser humano.

Ahora bien, desde otros puntos de mira, Chafe (1992) parece ser, junto a Stubbs (1996, 2006) y Tognini-Bonelli (2001), son algunos de los más entusiastas respecto a la LC en sus potencialidades como teoría; no obstante ello, Chafe aboga al igual que Fillmore (1992) por el trabajo mancomunado de técnicas de investigación diversas (tanto cuantitativas como cualitativas), argumentando que las cuantitativas por sí solas no logran revelar los aspectos más profundos del lenguaje y la mente. Esta propuesta de Chafe (1994) resulta posiblemente la más interesante y vanguardista en cuanto visualiza que la tarea del lingüista de corpus es tratar de estudiar el lenguaje y, a través de éste, llegar a la mente humana, es decir, indaga la naturaleza del lenguaje como una manifestación de la mente con especial atención a la conciencia humana. No obstante ello, es cauteloso en cuanto a las etiquetas para uno u otro tipo de lingüística y, en definitiva, se inclina por denominaciones más genéricas que no provoquen disputas clásicas: introspección/experimentación (Chafe, 1992, 1994).

Stubbs (1996, 2006), a pesar de ser uno de los fuertes defensores de la LC como teoría, también deja entrever algunas reservas. Este científico sostiene que el empleo de corpus digitales otorga una nueva manera de considerar la relación entre

los datos y la teoría, revelando cómo la teoría puede fundarse a partir de corpus accesibles de lenguaje natural. Para este investigador, la teoría puede emerger inductivamente de los datos, dando así fuerza a una lingüística sustentada en corpus.

7. PALABRAS DE CIERRE

En este artículo he buscado entregar algunas reflexiones acerca de la LC y argumentar a favor de mi visión particular acerca de ella. Queda claro que no existe aún una posición homogénea; tal vez nunca la habrá. El recorrido ha pretendido ser abierto y con bibliografía que permita al interesado consultar otras fuentes, lograr juzgar los aportes y –de ser necesario– encontrar su propio camino.

Así las cosas, el desarrollo de la LC continúa en un marco extraordinariamente interesante y en ebullición. Las implicancias, que la perspectiva teórica que (ya sea *profunda* o *superficial*) pueda traer consigo (Hunston & Thompson, 2006), anuncia –en alguna medida– que estamos en medio de un proceso de cambios y ajustes, y avanzando hacia una mirada cada vez más compleja y enriquecida de los objetos de estudio. Miradas que ciertamente potencian las indagaciones empíricas del lenguaje y de las lenguas particulares, desde múltiples puntos de mira y haciendo confluir aproximaciones antes impensadas.

A modo de cierre, ofrecemos esta cita de Stubbs (1996: 231):

> La lingüística de corpus presenta aún sólo lineamientos muy preliminares de una teoría que pueda relacionar textos individuales con corpus textuales, que pueda usar lo que es frecuente en los corpus para identificar lo que es típico del lenguaje, y que pueda usar los hallazgos acerca de los patrones frecuentemente recurrentes para construir una teoría que relacione el uso rutinario y creativo del uso lingüístico.

REFERENCIAS

Aarts, J. 1991. "Intuition-based and observation-based grammars". En K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics. Studies in honor of Jan Svartvik*. Londres: Longman, pp. 44-62.

Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. 2005. "Representativeness in corpus design". En S. Geoffrey & D. McCarthy (Eds.), *Corpus linguistics: Reading in a Widening Discipline*. Londres: Continuum, pp. 174-197.

Biber, D. & Tracy-Ventura, N. 2007. "Dimensions of register variation in Spanish".

En G. Parodi (Ed.), *Working with Spanish corpora*. Londres: Continuum, pp. 54-89.
Biber, D., Johansson, S., Conrad, S. & Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. Harlow, GB: Longman.
Biber, D., Reppen, R., Clark, V. & Walter, J. 2001. "Representing spoken language in university settings: The design and construction of the spoken component of the T2K-SWAL Corpus". En R. Simpson & J. Swales (Eds.), *Corpus Linguistics in North America*. Ann Arbor: University Michigan Press, pp. 48-57.
Briz, A. & Grupo Val.Es.Co. 2002. *Corpus de conversaciones coloquiales*. Madrid: Arco Libro.
Bud, R. 2003. "Introduction to elementary probabilistic theory and formal stochastic language theory". En Bod, R., J. Hay & S. Jannedy (Eds.), *Probabilistic Linguistics*. Londres: MIT Press, pp. 11-37.
Caravedo, R. 1999. *Gramática española: enseñanza e investigación. Apuntes metodológicos: Lingüística del corpus*. Salamanca: Ediciones Universidad de Salamanca.
Chafe, W. 1992. "The importance of corpus linguistics to understand the nature of language". En J. Svartvik (Ed.), *Directions in corpus linguistics*. Berlín: Mouton de Gruyter, pp. 79-97.
Chafe, W. 1994. *Discourse, consciousness and time*. Chicago: The University of Chicago Press.
Charniak, E. 1996. *Statistical language learning*. Cambridge: MIT Press.
Chomsky, N. 1969. "Quine's empirical assumptions". En D. Davidson & J. Hintikka (Eds.), *Words and objections. Essay on the Work of W.V Quine*. Dordrecth: D. Reidel, pp. 53-68.
Church, K. & Mercer, R. 1993. "Introduction to the special issue on computational linguistics. Using large corpora". En *Computational Linguistics* 9(1), 1-24.
Conrad, S. & Biber, D. (Eds.) 2001. *Variation in English: Multi-Dimensional Studies*. Londres: Longman.
Crystal, D. 1991. *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.
De Kock, J. 2001. *Lingüística con corpus: Catorce aplicaciones sobre el español. Apuntes Metodológicos*. Salamanca, España: Universidad de Salamanca.
EAGLES. 1996. *Synopsis and comparison of morphosyntactic phenomena encoded inlexicons and corpora. A common proposal and applications to european languages*. Pisa: ILC-CNR.
Ferrari, S. 2007. *La variación de los rasgos de la informatividad y de los tipos de nominalizaciones en los manuales de dos áreas de formación académica*. Tesis de Magíster en Lingüística. Valparaíso: Pontificia Universidad Católica de Valparaíso.
Fillmore, Ch. 1992. "Corpus linguistics and computer–aided armchair linguistics". En J. Svartvik (Ed.), *Directions in Corpus Linguistics*. Berlín: Mouton de Gruyter, pp. 35-60.
Francis, N. 1979. "A tagged corpus: problems and prospects". En S. Greenbaum, G. Leech & J. Svartvik (Eds.), *Studies in English linguistics for Randolph Quirk*. Londres: Longman, pp. 192-209.
González, C. 2005. *La constitución del destinatario discursivo en los editoriales de prensa*.

Tesis de Doctorado en Lingüística, Pontificia Universidad Católica de Valparaíso, Chile. [En línea]. Disponible en http://cybertesis.ucv.cl/tesis/production/pucv/2005/gonzález_cr/html/index-frames.html [Consulta: diciembre de 2007].

Gutiérrez, A. 2007. *Realización lexicogramatical del sistema semántico de la modulación: una aproximación a la descripción sistémico-funcional del español*. Tesis de Doctorado en Lingüística. Valparaíso: Pontificia Universidad Católica de Valparaíso.

Hair, J., Anderson, R., Tatham, R. & Black, W. 1999. *Análisis multivariante*. Madrid: Prentice Hall.

Halliday, M. 1992. "Language as system and language as instance: the corpus as a theoretical construct". En J. Svartvik (Ed.), *Directions in Corpus Linguistics*. New York: Mouton de Gruyter, pp. 61-77.

Hernández, R., Fernández, C. & Baptista, P. 2003. *Metodología de la investigación*. México: McGraw-Hill.

Hunston, S. & Thompson, G. (Eds.). 2006. *System and corpus: Exploring connections*. Londres: Equinox.

Ibáñez, R. 2007. *Comprensión de textos disciplinares escritos en inglés. Un estudio multivariante*. Tesis Doctoral en Lingüística. Valparaíso: Pontificia Universidad Católica de Valparaíso.

Instituto Cervantes. 1996. *Informe sobre recursos lingüísticos para el español (II). Corpus escritos y orales disponibles y en desarrollo en España*. Alcalá de Henares: Instituto Cervantes.

Jurafsky, D. 2003. "Probabilistic modelling in psycholinguistics: Linguistics comprehension and production". En Bod, R., Hay, J. & Jannedy, S. (Eds.), *Probabilistic Linguistics*. Londres: MIT Press, pp. 38-95.

Kennedy, G. 1998. *An introduction to corpus linguistics*. New York: Longman.

Kintsch, W. 1998. *Comprehension. A paradigm for cognition*. Cambridge: Cambridge University Press.

Leech, G. 1991. "The state of the art in corpus linguistics". En K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics. Studies in honor of Jan Svartvik*. Londres: Longman, pp. 8-29.

Leech, G. 1992. "Corpora theories of linguistic performance". En J. Svartvik (Ed.), *Directions in Corpus Linguistics*. New York: Mouton de Gruyter, pp. 105-122.

Leech, G. 2000. "Grammars of Spoken English: New Outcomes of Corpus-Oriented Research". En *Language Learning*, 50(4), 275-724.

Leech, G. 2002. "Sobre la importancia de los corpus de referencia". En *Donosit,* 24-25, 1-3.

Lope Blanch, J. 1969. Proyecto de estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica. En Actas del Simposio de México, 1969. México: PILEI.

Lope Blanch, J. 1977. *Estudios sobre el español hablado en las principales ciudades de América*. México: UNAM.

Lope Blanch, J. 1990. *Atlas Lingüístico*. México: Colegio de México.

Lope Blanch, J. 1994. *Estudios de historia lingüística hispánica*. Madrid: Arco/Libros.

Louwerse, M., McCarthy, P., McNamara, D. & Graesser, A. 2004. "Variation in language and cohesion across written and spoken registers". En K.D. Forbus, D.

Gentner & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum, pp. 235-242.

Malinowski, J. 1935. *An Ethnographic Theory of the Magical Word. Coral Gardens and Their Magic*, vol. II. Londres: Allen & Urwin.

Manning, C. & Schütze, H. (Eds.). 1999. *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Matus, A. 2002. "Corrección académica: ideal panhispánico y norma culta". En G. Parodi (Ed.), *Lingüística e interdisciplinariedad: desafíos del nuevo milenio. Ensayos en honor a Marianne Peronard*. Valparaíso: Ediciones Universitarias de Valparaíso, pp. 389-401.

McCarthy, M. (Ed.) 1988. *Vocabulary and language teaching*. Harlow: Longman.

McEnery, T. & Wilson, A. 1996. *Corpus linguistics*. Edinburg: Edinburg University Press.

Moreno, A. 1998. *Lingüística computacional: Introducción a los modelos simbólicos, estadísticos y biológicos*. Madrid: Síntesis.

Oyanedel, M., & Samaniego, L. 1998. "Notas para un nuevo perfil lingüístico de Santiago de Chile". En *Boletín de Filología de la Universidad de Chile*, 37, 899-913.

Parodi, G. 2003. *Relaciones entre lectura y escritura: una perspectiva cognitiva discursiva*. Valparaíso: Ediciones Universitarias de Valparaíso.

Parodi, G. 2004. "Textos de especialidad y comunidades discursivas técnico/profesionales: Una aproximación basada en corpus computarizado". En *Estudios Filológicos*, *39*(39), 7-36.

Parodi, G. (Ed.) 2005a. *Discurso especializado e instituciones formadoras*. Valparaíso: Ediciones Universitarias de Valparaíso.

Parodi, G. 2005b. *Comprensión de textos escritos*. Buenos Aires: EUDEBA.

Parodi, G. 2007a. "El discurso especializado escrito en el ámbito universitario y profesional: Constitución de un corpus de estudio". En *Revista Signos*, 40(63), 147-178.

Parodi, G. (Ed.) 2007b. *Working with Spanish corpora*. Londres: Continuum.

Parodi, G. & Venegas, R. 2004. "BUCÓLICO: Aplicación computacional para el análisis de textos. Hacia un análisis de rasgos de la informatividad". En *Revista Lingüística y Literatura* 15, 223-251.

Peronard, M. & Gómez, L. 1985. "Reflexiones acerca de la comprensión lingüística: hacia un modelo". En *Revista de Lingüística Teórica y Aplicada*, 23, 19-32.

Peronard, M., Gómez, L., Parodi, G., & Núñez, P. 1998. *Comprensión de textos escritos: de la teoría a la sala de clases*. Santiago de Chile: Editorial Andrés Bello.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1985. *A grammar of contemporary English*. Londres: Longman.

Rabanales, A. & Contreras, L. (Ed.). 1979. *El habla culta de Santiago de Chile. Materiales para su estudio*. Santiago: EUS.

Rojo, G. 2002. "Sobre la lingüística basada en análisis de corpus". [En línea]. Disponible en: http://www.uzei.com/Modulos/UsuariosFtp/Conexion/archivos54A.pdf [Consulta: diciembre de 2007].

Sabaj, O. 2004. *El comportamiento de los verbos abstractos en el corpus PUCV-2003*.

Tesis Doctoral en Lingüística. Valparaíso: Pontificia Universidad Católica de Valparaíso.

Silva, J. 2006. *Hacia un índice de lecturabilidad: El Manchador de Textos*. Tesina de grado. Valparaíso: Pontificia Universidad Católica de Valparaíso.

Simpson, R. & Swales, J. 2001. "Introduction to North American perspective on corpus linguistics at the millennium". En R. Simpson & J. Swales (Eds.), *Corpus linguistics in North America. Selections from the 1999 Symposium*. Ann Arbor: The University of Michigan Press, pp. 1-14.

Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stubbs. M. 1996. *Text and corpus analysis. Computer-assisted studies of language and culture*. Massachusetts: Blackwell.

Stubbs, M. 2001. *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell.

Stubbs, M. 2006. "Corpus analysis: the state of the art and three types of unanswered question". En Hunston, S. & Thompson, G. (Eds.), *System and corpus: Exploring connections*. Londres: Equinox, pp. 15-36.

Svartvik, J. (Ed.). 1992. *Directions in corpus linguistics: proceeding of Nobel symposium*. Berlín: Mouton de Gruyter.

Teubert, W. 2005. "My version of corpus linguistics". En *International Journal of Corpus Linguistics,* 10(1): 1-13.

Tognini- Bonelli, E. 2001. *Corpus linguistics at work*. Amsterdam: Benjamins.

Venegas, R. 2005. *Las Relaciones Léxico-semánticas en Artículos de Investigación Científica: Una Aproximación desde el Análisis Semántico Latente*. Tesis Doctoral en Lingüística. Valparaíso: Pontificia Universidad Católica de Valparaíso.

Venegas, R. & Silva, J. 2007. "El Manchador de Textos: Una herramienta computacional para el análisis de textos". En G. Parodi (Ed.), *Lingüística de Corpus y Discursos Especializados: Puntos de Mira*. Valparaíso: Ediciones Universitarias de Valparaíso, pp. 53-76.