LINGÜÍSTICA DE CORPUS Y DISCURSOS ESPECIALIZADOS: PUNTOS DE MIRA

Giovanni Parodi
EDITOR

Ediciones Universitarias de Valparaíso
Pontificia Universidad Católica de Valparaíso

© Giovanni Parodi, Editor. 2007
Inscripción Nº 165.318

ISBN 978-956-17-0406-0

Derechos Reservados
Tirada: 400 ejemplares

Ediciones Universitarias de Valparaíso
Pontificia Universidad Católica de Valparaíso
Calle 12 de Febrero 187, Valparaíso
Fono (32) 227 3087 - Fax (32) 227 3429
E.mail: euvsa@ucv.cl
www.euv.cl

Diseño Gráfico: Guido Olivares S.
Asistente de Diseño: Mauricio Guerra P.
Asistente de Diagramación: Alejandra Larrain R.
Corrección de Pruebas: Osvaldo Oliva P.
Imagen de portada: Altavoz S.A.

Impresión Litogarín, Valparaíso

HECHO EN CHILE

Índice

Introducción

La verdad es que con gran alegría presento este libro, alegría que espero se entenderá a lo largo de los párrafos de esta *Introducción*. Es un lugar común decir -en este apartado de un libro- que estamos satisfechos por concluir la tarea y que entregamos este producto a los ojos de múltiples lectores para que se aproximen desde diversos puntos de mira. En efecto, es un lugar común muy propicio: mirando hacia atrás y al revisar todos los capítulos, resulta impresionante comprobar todo el trabajo científico que se congrega en estas páginas y que ha concluido satisfactoriamente. Otro lugar común es decir que he disfrutado cada paso y proceso de su producción y de los que aún están en ejecución. Pero, quizás, los que más añoranzas me producen son los encuentros gestadores, las charlas soñadoras y las discusiones acaloradas, eternas y fructíferas en entornos académicos con la maestra, los amigos, los pares, los colegas, los alumnos y los discípulos. En definitiva, más que lugares comunes que responden al género, todas estas son expresiones obligadas del fruto de un sentido de completitud y satisfacción colaborativa.

También es motivo de gran orgullo entregar este libro, el cual continúa y profundiza algunas líneas de investigación emergentes dentro de las más tradicionales de la **Escuela Lingüística de Valparaíso** (ELV) de la Pontificia Universidad Católica de Valparaíso (PUCV) (www.linguistica.cl), a saber, la de la lingüística de corpus y del discurso especializado. Estas últimas ya apuntadas en publicaciones como *Discurso Especializado e Instituciones Formadoras* (2005) y *Lingüística de Corpus* (2007). A la vez, también se mantiene un foco en temáticas más tradicionales de este grupo de investigadores, cual es la comprensión de textos escritos, pero con renovadas miradas y desde nuevos escenarios.

En este contexto, el libro que presento constituye una singular sinergia entre la innovación, a través de la producción de nuevas áreas de conocimiento; y la tradición a través de la consolidación de las ya más clásicas. Así, unas se benefician de las otras, por ejemplo, la investigación acerca de los procesos de comprensión y producción de textos de variada índole y que circulan en diversos contextos se ve potenciada por la tecnologización y el apoyo con soportes computacionales.

De modo más específico, el volumen contiene un conjunto de dieciocho trabajos, aparentemente disímiles en algunos casos, pero mayoritariamente desde el análisis de colecciones de textos tanto escritos como orales. Esta aproximación hoy en día es conocida como Lingüística de Corpus. Una de sus fortalezas como libro es la de ofrecer en un solo texto un conjunto de investigaciones que aborda la indagación de corpus multigéneros y multiregistros, vale decir, se estudian fenómenos lingüísticos a partir de muestras originales y actuales de textos naturales y, al mismo tiempo, se describe el uso variado y diversificado a lo largo de géneros y registros orales, escritos, especializados, no-especializados, escolares, periodísticos, didácticos, etc.

No debe sorprender a nadie la posible heterogeneidad de líneas de acción o de marcos teóricos referenciales a lo largo de los capítulos del libro. La adscripción a ciertas metodologías o a nuestro grupo disciplinar no introduce sesgos teóricos ni metodológicos y esa es tal vez una de sus fortalezas y su riqueza constitutiva. Sin embargo, sí existe un sustrato de principios e intereses comunes que cohesionan y dan forma a nuestra Escuela. Ellos son fácilmente rastreables a lo largo de estos dieciocho capítulos.

El libro recoge una variedad de trabajos empíricos en los que se describen diversos aspectos del lenguaje humano, tales como gramaticales, cognitivos, textuales y discursivos. Sobre estas bases de descripciones a partir de corpus auténticos y completos, se emprenden otras investigaciones de las que aquí también se da cuenta. Por ejemplo, la indagación de la comprensión de textos altamente especializados por estudiantes de establecimientos técnico-profesionales.

A través de la división en seis grandes apartados, he tratado de aglutinar los trabajos que dan forma al volumen. Al mismo tiempo, estas macroproposiciones intentan dar cuenta de las diversas opciones de investigación científica que algunos miembros de la ELV cultivamos actualmente. Estas se enuncian a continuación:

Parte I: Lingüística de corpus y recursos computacionales

Parte II: Análisis multiregistros del español

Parte III: Comprensión del discurso

Parte IV: Discurso periodístico

Parte V: Discurso escolar

Parte VI: Análisis semántico latente: Teoría y empiria

Así, los dieciocho capítulos del libro, reunidos en torno a estos seis grandes focos, ofrecen una diversidad de temas, de datos empíricos y de herramientas tecnológicas. Un aspecto destacable lo constituye la clara decisión de los investigadores en esta línea de apoyarse en herramientas computacionales, todas ellas desarrolladas por miembros del equipo. El Capítulo 2 es una muestra de ello. La descripción pormenorizada del programa computacional **El Grial** revela de manera certera la dirección de las investigaciones. Del mismo modo, el Capítulo 3 también muestra otra herramienta desarrollada dentro de la ELV: **El Manchador de Textos**.

En mi opinión, este libro es un aporte a la lingüística de corte interdisciplinario en el mundo de habla hispana, dirigido a lingüistas, gramáticos, periodistas, profesores de lenguas, investigadores de variada índole y, por supuesto, a analistas del discurso. A alumnos de pregrado y postgrado. Muy útil para cientistas de la educación y miembros de entidades gubernamentales en la definición de políticas educativas. Incluso, algunos de sus capítulos, me atrevo a aventurar, podrían ser de interés para ingenieros o lingüistas computacionales y para estudiosos del procesamiento del lenguaje natural.

No obstante todo lo anterior, estoy convencido de que este libro también constituye una muestra de trabajado mancomunado y de lo que un grupo humano fuertemente cohesionado y con lazos afectivos poderosos puede llegar a realizar en el terreno científico. En este sentido, este libro (al igual que otros, pero mucho más aquí) representa los valores y la mística de nuestra **Escuela Lingüística de Valparaíso** al dar cuenta -con orgullo y sin distingo de ningún tipo- de investigaciones de alumnos de pregrado, postgrado, tesistas de magíster y de doctorado, junto a investigadores de trayectoria.

Pero, la verdad, el libro es incluso más que todo eso, ya que en mi calidad de editor, debo reconocer que mi labor como tal ha sido absolutamente co-participada, pues todos los autores en este libro han leído y criticado activamente varios de sus capítulos en más de una oportunidad. Ello es muestra ferviente del espíritu que nos anima y de que la escritura académica crece y se nutre colaborativamente y entre escritores y lectores miembros de una comunidad discursiva y de nuevos ingresantes a la misma.

Sin lugar a dudas, los trabajos aquí presentados son responsabilidad de sus autores, mas queda muy claro que estos productos han alcanzado su expresión máxima a través de un entramado de diversos y singulares aportes. Estas texturas se han co-construido en múltiples encuentros a partir de lecturas reiteradas y críticas generosas que no hacen más que revelar una mística comunitaria muy identitaria.

Giovanni Parodi
Editor

Valparaíso, Chile, agosto de 2007.

PARTE I

LINGÜÍSTICA DE CORPUS Y RECURSOS COMPUTACIONALES

Capítulo 1

Lingüística de Corpus: Puntos de mira

Giovanni Parodi

Introducción

En el primer capítulo de este libro me ha parecido oportuno abordar la cuestión de lo que se entiende por Lingüística de Corpus (LC) y de las opciones que se ofrecen a quienes se inician en este ámbito. También he estimado prudente incluir mi propia concepción y algunos comentarios y discusiones al respecto. De modo más conciso, pretendo entregar una definición operacional de la LC, en el marco de una discusión abierta y en franco desarrollo. Así, busco aportar una reflexión en que se explique por qué durante un tiempo se produjo un menor impacto y difusión de la LC y cómo se ha gestado su (re)surgimiento e indiscutible potencial para los estudios lingüísticos contemporáneos.

Desde este marco, en este primer capítulo, nos aproximamos a uno de los dos ejes de este libro. El del discurso especializado se irá precisando poco a poco en los capítulos venideros.

Una vez dicho esto, abordemos sin más preámbulos lo que tenemos en el punto de mira.

1. Para comenzar: algunas precisiones

En un primer momento, quiero dejar en claro que en este capítulo no se indaga en los antecedentes fundacionales de la LC. En efecto, no pretendo llevar a cabo un sondeo histórico desde los inicios de la LC ni menos de su prehistoria, sino focalizar su (re)surgimiento actual y precisar sus deslindes.

Ahora bien, de entrada, sostengo que la LC en su versión actual constituye un enfoque metodológico para el estudio de la lengua y que presenta 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 (tal como se puede comprobar en la multiplicidad de investigaciones de las que se da cuenta en varios de los capítulos de este libro). Sin embargo, y como veremos en el desarrollo de este capítulo, adscribir a la LC también involucra un cierto modo de aproximación específica a los datos lingüísticos, pues subyacen a este enfoque determinados principios fundamentales que lo tiñen de un grado de singularidad.

Así, tal como propongo, la LC puede definirse, *strictu sensu*, como una metodología de investigación de textos, 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.

Ahora bien, 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 visión.

Así, la opción que propongo en este capítulo le confiere a la LC un carácter original y se enfoca desde una mirada interdisciplinaria del lenguaje 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). En parte, a través de

ella se busca explícitamente desprenderse y deslindarse de algunas visiones excesivamente descriptivistas e inmanentistas (en especial de aquellas con sesgos conductistas) y también de otras más 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 solo 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 texto/discurso (Parodi, 2003, 2005a,b; ver Capítulo 10). Siguiendo las ideas de Teubert (2005), no parece posible –en nuestra 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).

Ahora bien, para otros científicos, tales 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. Por su parte, 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 multiregistros y/o multigéneros en los cuales es posible establecer comparaciones entre variedades de una lengua o incluso entre lenguas (ver Parodi 2005a, 2007a; ver, entre otros, los capítulos 6, 7, 10 y 15 en este mismo volumen).

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.

Por último, como se sabe, algunos de los principios que sustentan el enfoque de corpus fueron propuestos en la década del sesenta o setenta del siglo pasado, particularmente desde figuras señeras en el ámbito anglosajón, tales como Sinclair, Francis, Leech, entre otros. La cuestión central radica más bien en qué diferencia a ese modo de hacer lingüística y el actual 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 de este capítulo, brindo argumentos que revelan que esta opinión, desde mi mirada, no es correcta.

2. Desde la lingüística de la competencia hacia la lingüística del uso

Tal como la preocupación por el estudio de la lengua en contexto y su correspondiente variación emana de manera simultánea a partir de múltiples vertientes, 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 estudio del lenguaje ni de metodologías de índole hipotético deductivo (Francis, 1979; Conrad & Biber, 1998; Chafe, 1992; Sinclair, 1991; Leech, 1991; Kennedy, 1998; McEnery & Wilson, 1996; Moreno, 1998).

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 que:

> "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).

Ahora bien, la sucesión de estos 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 ciencia. Todo esto implica que la LC no está exclusivamente comprometida con una aproximación analítica cuantitativa, sino que una mirada cualitativa de los hechos lingüístico 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.

3. El concepto de corpus y algunos criterios metodológicos

Definir lo que hoy en día se entiende por *corpus* en el ámbito de la LC no resulta una tarea simple. Existen complejidades de diversa índole, muchas veces entrecruzando planos, que resultan difíciles de soslayar. Algunas 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 al concepto de corpus.

Leech (1991, 1992), por su parte, sostiene que un corpus computacional se constituye en un fenómeno nada excitante, pues resuelta ser solo 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 solo una gran cantidad de textos con cierto formato.

> "On the face of it, a computer corpus is an unexciting phenomenon: a helluya lot of text, stored on a computer." (Leech, 1992: 106).

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

> "[...] the availability of vastly computer corpus resources has enabled syntactic and lexical phenomena of a language to be open to empirical investigation on a scale previously unimagined."

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

> "[...] a collection of naturally-occurring language texts, chosen to characterize a state or variety of a language."

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

> "A collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language."

En particular, las alusiones directas a la escritura y a la oralidad, en especial a esta ultima modalidad de la lengua enfrenta 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 rea-

liza 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:

> "a collection of pieces of language that are selected and ordered according explicit linguistic criteria, in order to be used as a sample of language [……] A corpus which is encoded in a standardised and homogeneous way for open-ended retrieval tasks." (EAGLES, 1996a).

En mi opinión, 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 tanto 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 (1996b) 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 de ser tamaños similares
5. El corpus, como un todo, debe tener una procedencia clara

Del mismo modo, Biber, Reppen, Clark y 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 (1996b) y Biber et al. (2001). Aunque Biber y colaboradores (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 multiregistros (Biber, 1988).

Considerando lo hasta aquí discutido, es factible detectar tensiones en cuanto al concepto de corpus. Ya sea si este 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.

Ocho aspectos, al menos, se identifican como cuestiones 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 en el capítulo. Solo 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– solo 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 algunas de las investigaciones de que se da cuenta en este libro. 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 carreras 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 este 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 solo logran encontrar acogida en cierto tipo de investigaciones cuantitativas que logren, por ejemplo, determinar previamente en base al universo estudiado, su corpus de análisis.

3.1. 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 solo por características inherentes a la naturaleza de los mismos. Esta amplia y algo vaga definición preliminar permite, en mi opinión, que al menos, un par de textos constituya así un corpus (acogiendo todas posibilidades mono o multimodos o mono o multimedios, sin entrar en las complejidades de lo que se entiende por texto).

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 comparación 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 estos solo existe acceso restringido o su naturaleza misma los hace escasos y, por ende, su tamaño puede ser 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 semi-automático 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 pre-corpus

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. Solo 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 semi-estructuradas (CEO) (Parodi, 2005b).

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 pre-corpus 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 (Tognini-Bonelli, 2001). 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, solo 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.

4. Nuevos orígenes de la LC

El (re)florecimiento de los estudios basados en corpus se puede fijar a comienzos de la década del sesenta, marcado –en parte- por la fuerte irrupción de los computadores en el ámbito lingüístico y el desarrollo 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 importante 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 mega-corpus 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 mega-corpus 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 posibilitado 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, Leech, Conrad & Finegan, 1999; Carter & McCarthy, 2006). 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 mega-corpus a corpus comparativamente más pequeños, pero altamente focalizados temática, estructural 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 el resto de capítulos de este libro).

4.1. La LC y la lengua española

Desde la lengua española, la investigación reciente ha revelado la necesidad de enfatizar el uso de corpus digitales progresivamente más amplios y diversos con el fin de avanzar en descripciones lingüísticas más profundas y robustas, y también como un medio empírico eficaz de 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 solo últimamente han logrado ser descritas de manera más acuciosa, aunque aún de modo preliminar (al respecto, ver Capítulos 4, 7, 8, 11, entre otros. También ver, Parodi 2005b, 2007a). Desafortunadamente, ello todavía no logra materializarse en la forma de una gramática 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 suscintamente 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; Pons & Ruiz, 2005). 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 (2007a),

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 el resto de capítulos de este libro) como en desarrollo de tecnologías *ad hoc* (ver Parodi & Venegas, 2004, y los Capítulos 2 y 3 en este mismo volumen). En particular, resaltamos la mirada multigéneros, multiregistros 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, 2004a; Venegas, 2005; González, 2005; King, 2006; Silva, 2006; Fierro & Céspedes, 2006; Gutiérrez 2007; Ferrari, 2007).

5. ¿Es la LC solo una metodología o una teoría de las facultades probabilísticas?

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; Bod, 2003; Juraksky, 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 (ver Capítulo 9). 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 (para mayores detalles de estas líneas de acción, véase los Capítulos 16, 17 y 18, en cuanto al *Análisis Semántico Latente*).

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; también véase en este libro, los Capítulos 9 y 10, de Ibáñez y de Parodi respectivamente).

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.

Desde puntos de mira similares, Chafe (1992) parece ser, junto a Stubbs (1996, 2006) y Tognini-Bonelli (2001), uno 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 muy 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. En palabras 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".

Así las cosas, la disputa 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 la indagaciones del lenguaje y de las lenguas particulares, desde múltiples puntos de mira.