BFUCh XXXIV (1993-1994): 489-509

El léxico del español de Chile[1]
El léxico periodístico

Leopoldo Sáez Godoy
Universidad de Santiago de Chile
Universidad Católica Blas Cañas

¿Cómo conocer el acervo léxico del chileno? En promedio, ¿cuántas unidades (activas y pasivas) tiene según las edades, actividades y estudios? ¿Cómo se interrelacionan? ¿Cómo están sistematizadas en el cerebro? ¿Hay que imaginárselo como una especie de diccionario alfabético, uno conceptual o tal vez como una base de datos interactiva?

La importancia de abordar problemas de esta índole queda en evidencia si, por ejemplo, concebimos el lenguaje como el resultado de un proceso intelectual de una comunidad que además actúa como una suerte de red que nos permite orientarnos en el mundo e interpretarlo.

La identidad cultural de un pueblo se confunde en gran parte con su acervo léxico que es el depósito de la memoria colectiva. Allí se encuentran los productos de su proceso de interpretación del mundo, que han sido aceptados por los hablantes. Uno de los rasgos más importantes que nos diferencian de nuestros vecinos es nuestro repertorio léxico, su composición y frecuencia de uso relativa.

Sería necesario acceder a él, por ejemplo, para planificar la enseñanza de la lengua materna de modo tal que el niño aprendiera gradualmente desde las unidades léxicas más usuales y generales –las que son imprescindibles para entender cabalmente los textos corrientes: diarios, revistas, cartas, instrucciones, manuales, conversaciones familiares, películas– y luego avanzar a las menos frecuentes y más específicas que aparecen en textos de mayores exigencias como los literarios, los históricos, los filosóficos, suplementos culturales, los científicos.

[1] Este es un proyecto financiado por FONDECYT.

En la escuela se dispone de un tiempo limitado para introducir y ejercitar nuevos términos. Habría que imaginar un sistema diferencial que permitiera salvar el abismo que existe entre el léxico de un niño desfavorecido socialmente y el de otro que vive en un ambiente estimulante con conversaciones variadas e interesantes y tiene un contacto habitual con revistas, periódicos, libros, videos.

El método más directo sería poder introducirse de algún modo en los cerebros de hablantes representativos de distintas edades, sexo, formación cultural, procedencia geográfica, actividades, y extraer de allí las unidades léxicas de que disponen, y algunos datos más como el sistema de almacenamiento y de incorporación de nuevas unidades, el método de acceso al acervo léxico, la importancia relativa de unidades léxicas equivalentes, el sistema de relaciones entre unidades.

Por el momento, no parece ser posible este abordamiento directo, pero podría decirse, con algún grado de optimismo, que ya se está trabajando en esa dirección: el grupo Fujitsu y el Instituto de Investigaciones de la Universidad Hokkaida en Sapporo están experimentando en la lectura del lenguaje mental (*silent speech*) mediante computadoras. Mientras tanto tenemos que recurrir a otros métodos:

1. La dialectología, en especial los atlas lingüísticos, obtiene su material basándose en cuestionarios sobre un número de temas, entre otros, sobre el entorno y las actividades características de la región donde se habla el dialecto en estudio. Tradicionalmente estos estudios han estado muy ligados a lo etnográfico, como en la línea *Wörter und Sachen*.

2. La distinción sistemática entre léxico activo y léxico disponible ha traído consigo la elaboración de metodologías para extraer el acervo léxico de grupos de hablantes en torno a determinados campos de interés. La frecuencia y el rango de las unidades léxicas permite asignarles índices numéricos para discriminar sobre su importancia relativa.

3. Otro método es el de los corpora. La introducción de la computación en las ciencias humanas ha permitido disponer del vocabulario empleado por un autor en una obra, un tipo de obras, un período, su producción completa, además de índices rectos o inversos y concordancias completas.

Sin lugar a dudas, todos estos métodos son provechosos. Sus frutos son suficientemente conocidos, por lo que puedo ahorrarme una frondosa y erudita nota a pie de página.

Se trata de escoger la herramienta más adecuada para el objetivo específico. Me parece que 3 puede ser más útil que los otros para conocer objeti-

vamente el léxico global de un dialecto. 1 y 2 necesariamente son más restringidos, ya que el léxico que se recibe es el que corresponde a los campos, a menudo muy reducidos, que le interesan al investigador[2].

3 ofrece además posibilidades más amplias de caracterizar cada unidad léxica de acuerdo con su frecuencia y distribución en múltiples subconjuntos de tipos de textos y de hablantes. Estas informaciones son cada vez más importantes para la descripción de un dialecto.

Si no podemos abordar directamente el cerebro del hablante, por lo menos, sí está relativamente a nuestro alcance intentar analizar el resultado de su actividad: los textos. Esta decisión significa que nuestro objeto queda circunscrito al léxico activo del dialecto.

Los textos transmitidos pueden ser escritos u orales. Aquí nos encontramos con un nuevo obstáculo: no tenemos acceso a todos los textos escritos por los hablantes de español en Chile; queda fuera de nuestro alcance toda la producción no pública: cartas, diarios de vida, memorandos, informes. Mucho menos podemos disponer de los textos orales que producen los chilenos, por pequeña que sea la unidad de tiempo que nos fijemos.

En todo caso, si no existieran estas restricciones y pudiéramos disponer realmente de todos los textos orales o escritos producidos en un período de tiempo, prácticamente no podríamos procesarlos, ya que no disponemos de la infraestructura computacional, ni humana, ni mucho menos de los fondos para mantener esa empresa gigantesca, que tal vez no tendría sentido.

CIECh

Tamaño

Como en otras disciplinas, hay que acudir a muestras representativas. De acuerdo con experiencias de otros países, para el proyecto *Corpus Integral del Español de Chile (CIECh)*[3] nos pareció adecuada una muestra de dos millones de palabras-textuales (p-t) (*running words*): un millón de lengua escrita y un millón de lengua oral[4].

[2] Sin lugar a dudas podría ampliarse el número de campos, por ejemplo, siguiendo una línea como la de Hallig-Wartburg. Pero las posibilidades de trabajo práctico se complicarían extraordinariamente.

[3] Este proyecto fue financiado por FONDECYT en el período 1989-1991. Hay dos informes: Sáez, 1990b y Sáez, 1993b, seguramente en prensa.

[4] Fijemos algunos puntos de referencia: el *Corpus del Español de México* comprende cerca de dos millones de p-t; el *Freiburger Korpus* (alemán oral), 600.000; el *Limas Korpus* (alemán escrito), un millón; el clásico *Brown Corpus* (inglés norteamericano escrito), un millón; el *London-Lund Corpus* (inglés británico escrito y oral), un millón; el diccionario de frecuencias de Juilland y Chang-Rodríguez (español peninsular literario), medio millón; el PE77 de Gotemburgo (diario *El País* de Madrid), dos millones.

Límites temporales

Se han establecido como límites temporales del corpus los años 1970 y 1992. Este período es uno de los más importantes de la historia de Chile: comprende los tres años de gobierno del Presidente Salvador Allende, los dieciséis del "régimen autoritario" del General Pinochet y el retorno paulatino a la democracia con el Presidente Patricio Aylwin.

Lengua escrita

Disponemos de periódicos, revistas, libros y folletos (todos estos son textos que por ley deben ser enviados a la Biblioteca Nacional); boletines o publicaciones de circulación restringida y efímera; programas de radio y televisión, cartas privadas, canciones, diarios de vida.

En el caso de la prensa oficial (periódicos, diarios, revistas) y de los *libros y folletos* recibidos por la Biblioteca Nacional hemos procedido aleatoriamente. Mediante tablas seleccionamos once textos por año y tres páginas no seguidas en cada uno de ellos (previamente descartamos autores no chilenos, reimpresiones, traducciones, lenguas extranjeras, autores repetidos). Luego corregimos levemente la selección para que no quedara sin representación ninguna de las doce líneas temáticas por las que están clasificadas las obras en la Biblioteca Nacional. Tenemos 250 textos de esta clase (250.207 p-t).

En lo que respecta a *periódicos y revistas oficiales*, seleccionamos aleatoriamente, año por año, primero un título, luego mes, día, página. En esa página se escogió un artículo completo. Se ha procurado que estuvieran representadas las diferentes secciones de un periódico: editorial, reportajes, política, vida social, economía, deportes, hechos policiales, cultura. Hasta el momento tenemos 566 textos con un total de 278.137 p-t. Incluiremos también una muestra del lenguaje de los *comics*.

Para la *prensa no oficial* trabajamos con el fondo de boletines de la ONG "Educación y Comunicación" que cuenta con unos mil boletines de los más variados grupos de intereses: estudiantes, asociaciones femeninas, sindicatos, centros laborales, pobladores, agrupaciones culturales, defensa de derechos humanos, pensionados. Escogimos 400 textos. Son textos breves (promedio 511 p-t) que en conjunto aportaron al corpus 204.494 p-t.

También trabajamos sistemáticamente en lo que respecta a las *canciones populares*. Se aprovechó la información existente en las radioemisoras acerca de las canciones chilenas más populares año por año, desde 1970 hasta 1990. A estos "rankings" se agregó la producción de los compositores más connotados (Mann, Peralta, Grondona, Arenas, Plaza, Jara, Parra,...). De esta fuente obtuvimos 117 textos que aportaron 22.729 p-t.

Menos sistemática ha debido ser la selección de textos privados y de poesía popular. Coleccionamos alrededor de dos mil *cartas* de las más diversas temáticas y de ellas, aleatoriamente, sacamos 225 (84.784 p-t), procurando que estuvieran representados diversos tipos de hablantes. Más difícil ha sido disponer de *diarios de vida* (1.760 p-t).

En lo que respecta a *poesía poblacional* incluimos una muestra (40 poemas) de una recolección del Prof. Manuel Alcides Jofré efectuada en los barrios marginales de Santiago[5] (40 textos con 8.230 p-t).

Además recogimos textos muy variados y difícilmente clasificables, como: *volantes*, *avisos económicos*, *propaganda*, *cartas de restorán*, etc.: 92 textos con 33.965 p-t.

Nos propusimos trabajar con un millón de p-t. Hasta el momento llevamos digitadas 948.878, de las que dejaremos fuera topónimos, antropónimos, palabras de otras lenguas no incorporadas al español.

Lengua oral

Los textos escritos están disponibles. El problema es juntarlos y seleccionarlos. Muy distinto es el caso de la lengua oral. *Verba volant, scripta manent.* Es excepcional el caso de Ambrosio Rabanales y Lidia Contreras que han editado las entrevistas que les sirvieron de base para sus estudios sobre la norma culta de Santiago de Chile. Lo habitual es que no quede nada de estas recolecciones.

Para juntar una muestra oral representativa hemos tenido, por ello, que grabar y transcribir entrevistas, conversaciones cara a cara o telefónicas, discursos[6], charlas, disertaciones, homilías, prédicas, programas de radio y televisión, videos y películas, textos artísticos elaborados oralmente. Este es un trabajo lento y complejo.

Hemos realizado *entrevistas* en las que el informante reacciona ante una pauta que le presenta el entrevistador y habla sobre asuntos generales o sobre su actividad profesional. Hay entrevistas *directas* dirigidas, grabadas y transcritas por nuestro equipo. Hemos aprovechado también entrevistas difundidas en *radio* o *televisión* y las publicadas en *revistas* o *periódicos*. Estas últimas siempre tienen el riesgo de algún retoque ("edición") de parte del periodista (de este tipo tenemos 164 entrevistas con un total de 227.627 p-t).

[5] Es conveniente aclarar que poesía "poblacional" no es idéntico con "poesía popular", porque la "población" es un universo heterogéneo.

[6] Naturalmente, en todos estos casos, para esta sección nos sirven sólo los textos que no son una simple lectura de un manuscrito.

Hasta el momento hemos procesado muestras de los siguientes oficios, profesiones o grupos (*entrevistas jergales*): abogado, actor de teatro, administrador público, albañil, arquitecto, artesano ceramista, artesana en calzado, artesano en carteras, artesano en cuero, artesano en metales, asesora del hogar[7], autobusero, auxiliar de párvulos, barman, bibliotecaria, campesino, carabinero, carnicero, carpintero, cartonera, cientista literario, cocinera, conductor de microbús, drogadicto, electricista, empleado de laboratorio, empresario chacarero, enfermera universitaria, escribiente de la Armada, estilista, estucador, feriante, ferretero, fotógrafo, futbolista, gásfiter, hortelano, ingeniero acústico, ingeniero agrónomo, ingeniero forestal, lustrabotas, manipuladora de alimentos, mantenedor de ascensores, marinero, matrona, mecánico de combustión interna, mecánico de automóviles, mecánico desabollador, médico laboratorista, mediero, minero, modista, mueblista, músico popular, panadero, paramédico, peluquera, pescador, prensista offset, profesor de castellano, profesor de idiomas, profesor de química, programador, refinador de cobre, relojero, reponedor de lácteos, sastre, secretaria, soldado, supervisor educacional, suplementero, taxista, tipógrafo, tornero, vendedor de artículos deportivos, vendedor de artículos fotográficos, vendedor de botillería, vendedor de flores, vendedor de maderas, vendedor de pescados y mariscos, vendedor de materiales de construcción, vendedora de la vega, veterinario, yerbatero.

La lista podría ampliarse hasta alcanzar varios cientos o miles. Pero no tiene sentido intentar la exhaustividad. Sólo agregaremos todavía algunas decenas hasta contar con muestras de los oficios que tienen mayor número de cultores en nuestra sociedad. Siempre está abierta la posibilidad de incluir materiales para reforzar el estudio de una jerga en particular. Hasta el momento hemos digitado 159 entrevistas jergales (182.646 p-t).

Se ha incorporado un número considerable de entrevistas que no están centradas en una actividad específica, sino que recogen la lengua común (*entrevistas generales*): 100 con un total de 136.391 p-t.

Hemos grabado *conversaciones informales* sin un conductor predeterminado. Los hablantes ignoran que se está grabando la conversación. Se han registrado conversaciones en buses, bares, reuniones: 32 textos con 77.017 p-t.

Los *discursos* o *disertaciones* constituyen una situación formal típica con emisor y receptores en papeles no intercambiables. En esta sección nos interesan sólo aquellos no leídos por el hablante, sino improvisados o producidos según una pauta. Tenemos 19 textos con 25.851 p-t.

[7] Eufemismo que intenta valorizar la profesión que se ha denominado *criada, niña de mano, china, empleada doméstica, doméstica* y, actualmente, *nana.*

Las *conversaciones telefónicas* tienen características singulares: el receptor no está presente, pero puede asumir de inmediato el papel de emisor. Los interlocutores pueden hablarse, pero no tienen un entorno común y no se están mirando, lo que suele dar gran libertad a la expresión. Como no disponemos de los recursos técnicos adecuados, nos hemos contentado con grabar a sólo uno de los interlocutores. Tenemos 41 conversaciones con 10.505 p-t.

La gran difusión de la *radio* y de la *televisión* hacen ineludible contar con muestras de *programas*, en los que los participantes no se ciñan a un libreto y hablen espontáneamente. Incluimos 53 textos radiales (55.301 p-t) y 21 de televisión (31.568 p-t).

Hemos recogido también *usos artísticos del lenguaje: cuentos* (19 con 19.010 p-t), *payas* (5, 3.123 p-t), *poemas* (50, 13.807 p-t). La condición es que hayan sido creados en la lengua oral.

Hasta noviembre de 1993 llevamos digitados 2.361 textos (669 orales y 1.744 escritos) con un total de 1.754.063 p-t (805.185 p-t del lenguaje oral y 948.878 del lenguaje escrito). En los cuadros puede verse un panorama del estado actual.

Informantes

Los informantes representan una amplia variedad de hablantes. Empleamos cinco criterios de clasificación:

a) *Origen geográfico: Norte* (Tarapacá, Antofagasta, Atacama y Coquimbo), *Centro* (Valparaíso, O'Higgins y Maule), *Santiago* y *Sur* (Bío-Bío, Araucanía, Los Lagos, Aysén y Magallanes) (N/C/S/U).

b) *Edad: jóvenes* (15-29 años), *adultos* (30-59), *mayores* (desde 60 años) (J/A/M).

c) *Sexo:* (H/M).

d) *Grado de instrucción:* hablantes analfabetos o con algún curso de enseñanza *básica*, hablantes que alcanzaron la enseñanza *media* y hablantes con enseñanza *superior*, esto es, postmedia, no necesariamente universitaria (B/M/S).

e) *Urbano-rural:* (U/R)

Combinando todos estos criterios tenemos representados 144 tipos distintos de informantes.

PROCESAMIENTO

Si ya es laborioso el proceso de recoger, seleccionar, digitar los textos, el paso siguiente no lo es menos. ¿Qué haremos con todo este material?

Hay diversas posibilidades de organización. Veamos algunas:

a) Confeccionar un *listado alfabético de palabras-textuales*[8] con sus *referencias*, que en nuestro caso serían el *código del texto* (tipo de texto y número correlativo), *código del hablante* (las variables de los cinco criterios de clasificación) y *número de la línea* donde aparece la palabra-textual. A esto podría agregársele el contexto (una o más líneas) (*concordancia*).

b) Un pequeño avance en el grado de abstracción sería, en lugar de preocuparse de cada palabra-textual, hacer un listado alfabético de las *formas-gráficas* con todas las palabras-textuales correspondientes y algunos datos estadísticos como la *frecuencia* total y el *rango* (esto es, el lugar que tiene en un listado de las formas más frecuentes).
Las frecuencias pueden ser totales, por texto, tipo de texto, período, etc.
En esta solución aparece, por ejemplo, la forma gráfica *bajo* con todas las realizaciones que tiene en el corpus, sean verbos, adjetivos, substantivos o preposiciones, sin ninguna discriminación.

Hasta aquí está trabajando el computador con programas muy conocidos y la intervención de los lingüistas es reducida. Comienza a acentuarse en las posibilidades siguientes:

c) Puede hacerse una pequeña elaboración del material si a cada palabra textual se la provee de un *especificador funcional* que indique la categoría gramatical que tiene en el contexto y luego se reúnen aquellas que tienen la misma categoría. De este modo, se separan los *vino* verbo de los *vino* substantivo. A continuación pueden ponerse las concordancias completas de cada una de estas *formas* que llamaremos *funcionales*.

d) Otro avance sería juntar todas las *formas funcionales* que pertenecen a un mismo *conjunto flexional* y agruparlas bajo una de las formas posibles, que suele llamarse *lema*. Así separaríamos los *fue* verbo *ir* de los *fue* verbo *ser* y reuniríamos todas las formas de cada uno de estos verbos documentadas en el corpus.
El lema representa a todo el conjunto flexional; puede ser una forma existente en el corpus.
Los conjuntos en este nivel son polisémicos, es decir, las diferencias semánticas que se aprecien se entienden como diferentes *acepciones* de una misma unidad.

[8] Para la terminología, véase SÁEZ, 1988.

Además, resulta operativo emplear la convención de que los conjuntos contienen sólo elementos monoverbales. Esto permite encontrar con facilidad cualquier elemento léxico. Naturalmente las unidades léxicas pluriverbales existen, pero hay un amplio campo de indecisión entre las combinaciones libres y las fijas, que suele resolver cada lingüista de acuerdo con las condiciones que él determine. La lengua es dinámica y hay una serie de construcciones que se están moviendo entre los extremos. A menudo no coinciden las decisiones de diferentes investigadores. No hay dudas con respecto a *haber + participio*, pero hay una decena de otras perífrases verbales cuyo status no es tan claro.

Igual sucede con las lexías complejas que, como es bien sabido, tienen muy variadas estructuras: fijas, semifijas, con o sin elementos intercalados, con posibilidades de substitución de un elemento, con un significado unitario, con elementos ligados, etc.

Los lingüistas tendemos a caer en la desesperación y la frustración cuando las unidades que resultan del procesamiento computacional de textos no coinciden exactamente con las unidades con las que estamos acostumbrados a trabajar. Creo que hay que tener conciencia de que con grandes masas de datos en lugar de involucrarse en trabajos manuales interminables puede ser más conveniente operar con entidades convencionales, cuyas características estén claramente definidas, y estudiar el modo de establecer relaciones con las unidades lingüísticas habituales.

Naturalmente se debe intentar acercar lo más posible unas y otras, por ejemplo, mejorando los procedimientos de párser.

e) En esta etapa los conjuntos flexionales se convierten en *conjuntos léxicos*, mediante la consideración del factor semántico léxico.

Se separan homónimos, lo que se materializa en un *especificador semántico* (v.g. un número que diferencie cada homónimo) que puede agregarse al *especificador funcional.*

Se reúnen los elementos de las lexías complejas, que son tratadas unitariamente y se separan los de las contracciones Se consideran las lexías textuales (discurso repetido) como unidades.

Esta etapa constituye el trabajo propiamente lingüístico que aprovecha el material puesto a disposición del especialista en cualquiera de las etapas previas.

Me parece que en la elaboración del corpus una meta deseable sería llegar a d) y dejar e) para el trabajo de los lingüistas que utilicen nuestro material.

De este modo, podríamos contar con una estructura que contuviera de algún modo los siguientes elementos:

a) *lema*
b) *especificador gramatical*
c) *conjunto funcional*

Cada *conjunto funcional* está compuesto de las

c_n) *formas funcionales* correspondientes que estarán acompañadas de
d) *concordancias completas* (tres líneas de texto) y sus
e) *referencias* (e_1 tipo de texto con su e_2 número correlativo)
f) *características del hablante*

indicaciones estadísticas:

g) *frecuencias*: totales de todo el conjunto flexional, porcentaje con respecto al total, individuales, totales de cada forma funcional, de cada una de las cinco variables que consideramos en los hablantes, de lengua oral/escrita, de cada tipo de texto, de cada texto individual, etc.

Naturalmente se contaría además con la colección de textos digitados.

La posibilidad de comparar resultados de trabajos similares depende de las decisiones que se tomen en este punto. No son comparables *formas gráficas* con *formas funcionales* o con *formas flexionales* o con *formas léxicas* (la frecuencia es decreciente). Hay puntos conflictivos para clasificar gramaticalmente todas las palabras-textuales de un corpus (¿adjetivos o participios?, ¿cuántos *ses*?) o para lematizar (¿*yo* o *yo*, *me*, *mí* como lemas?). Habrá que especificarlos. Pero la tarea se hace prácticamente incontrolable si pasamos a e) e introducimos el factor semántico.

Hay que tener presente que se está trabajando con millones de unidades y sería deseable que al investigador que inicia el proyecto le alcance su vida para preparar el material y aprovecharlo en una mínima parte.

LA PRENSA

Géneros periodísticos

Dado que nuestro primer interés recae en la lengua de la vida diaria, en los textos escritos hemos privilegiado el mundo de la prensa[9]. Hay además va-

[9] En esto desoímos la experiencia de Luis Fernando Lara que estima que el lenguaje de prensa le resultó poco productivo para su corpus del español de México. Esperamos no tener que lamentarnos de esta decisión. En todo caso siempre podremos rectificar rumbos, aunque estos cambios siempre tienen sus costos.

rias razones que le otorgan una situación especial: mayor alcance y difusión, influencia permanente, llegada a todo tipo de receptor. Televisión, radiodifusión y prensa escrita tienen un efecto incalculablemente superior al de la literatura[10], lo que implica una gran responsabilidad de la que no parecen estar conscientes muchos periodistas.

Es conveniente destacar que, pese a la fuerza de la televisión, la radiodifusión sigue siendo un medio importante de comunicación masiva en nuestro país (en la actualidad existen más de quinientas radioemisoras que tienen mucha influencia, especialmente en los sectores rurales). Las radioemisoras trabajan con textos escritos (libretos) –noticias, horóscopos, programas científicos y culturales de diverso tipo, consejos para el hogar, propaganda, radioteatro– y textos orales: entrevistas, programas de habla espontánea, como los deportivos o de proselitismo religioso, o los basados en conversaciones telefónicas de los locutores con sus auditores, en los que con igual ligereza satisfacen preferencias musicales o proponen fórmulas para solucionar tanto males físicos como problemas espirituales de la más variada índole.

En televisión se da también esta combinación de textos escritos y orales. Son usuales los programas de entrevistas, foros, concursos, junto con los de carácter periodístico, cultural o científico, por lo general escritos.

Para nuestro corpus utilizamos más de mil textos orales y escritos de medios de comunicación masiva, que se constituyen así en la principal fuente de nuestra documentación.

Dentro de un periódico conviven diversos tipos de textos[11], niveles de lenguaje y lenguajes especializados. Las secciones poseen estructuras características: editorial, deportes, crónica roja, política, necrología y defunciones, avisos económicos, crítica, artes plásticas, literatura, cine, música clásica, espectáculos, economía y negocios, moda, etc.

[10] Una encuesta hecha entre los habitantes de Santiago muestra esta situación en cifras (Céneca-Flacso, *Encuesta. Consumo Cultural,* Santiago, enero de 1988, 154 pp.):

En los días anteriores a la encuesta, el 91,3% de los informantes había visto televisión. En cambio, un 67% había leído algún diario en la semana, un 55,9%, alguna revista en los tres meses anteriores y un 45,3% no había hojeado ningún libro en todo el año (26,3% estaba leyendo alguno, 15,4% había leído alguno en los tres últimos meses y 2,9% en el último año).

La audición de las radios es diferente si se trata de FM (68,3% durante la semana y 68,9% los fines de semana) o AM (58% y 40,9%, respectivamente).

[11] En este último tiempo, en Chile se han publicado algunos trabajos que tienen como objeto el lenguaje de la prensa. BURDACH *et al.* (1992) analizan los editoriales del diario *El Mercurio,* el periódico chileno más influyente. MAYORGA y CIFUENTES (1992) hacen un enfoque textual de un artículo periodístico de EDUARDO GALEANO.

Es frecuente encontrar suplementos sobre deportes, agricultura, economía, cultura, arquitectura y construcción, en los que naturalmente se emplean conjuntamente el lenguaje especializado y el estándar.

El lenguaje de la prensa

El lector habitual va asimilando términos de la jerga política: *bajar candidatos* (retirar su candidatura), *sensibilidades* dentro de un partido (tendencias), *primarias* (elección de representantes para elegir al candidato final), *votos duros* (los electores militantes), *votos cruzados* (por candidatos de distintas tendencias políticas), *renovado* (el no ortodoxo); de la jerga jurídica: *constituirse* en el lugar de los hechos (visitarlo oficialmente el juez), *substanciar un proceso* (llevarlo adelante), quedar una causa *en acuerdo* (suspender el dictamen). Y lo mismo sucede con las jergas de la economía, la moda, los deportes, etc.

Estos términos pasan a engrosar el léxico pasivo del lector y de acuerdo con una serie de factores –como la frecuencia de uso, la mayor cercanía del quehacer especializado con su quehacer vital, la moda– pueden ir incorporándose lentamente al léxico activo; pero estamos lejos de ello todavía. Ni siquiera un abogado diría, por ejemplo: *"Dejemos este trabajo en acuerdo, constituyámonos en el bar 'La picá de On Andrés' y substanciémonos unos perniles"*.

El periodismo puede así cumplir el papel de introductor de términos de los lenguajes especializados en la lengua estándar. Es muy probable que cree neologismos, especialmente en algunas jergas muy cambiantes como las de los comentaristas deportivos, ávidos de novedad en la expresión.

Es frecuente encontrar disparidades entre la norma académica y la norma descriptiva objetiva en uso en la prensa. Pueden ser un reflejo de tendencias vivas en el dialecto (como en el *dequeísmo* o el uso personal de *haber*) o a veces simplemente de la ignorancia o premura del periodista[12].

En general, puede sostenerse que el lenguaje de la prensa es un buen representante de la lengua estándar de la comunidad, que prácticamente no experimenta variaciones diatópicas.

Sin embargo, hay que tener presente que tiene ciertas limitaciones, por lo que cabría esperar omisiones léxicas importantes:

- El lenguaje de la prensa escrita naturalmente deja fuera muchos elementos del lenguaje oral: conectores pragmáticos, deixis, recursos fáticos, muletillas.

[12] Pueden verse numerosos ejemplos en Sáez (1989) y en la Sección "Materiales Lingüísticos", de *Literatura y Lingüística*, Nos. 4 (1991) y 5 (1992).

- Hay temas que son públicos y otros que no lo son. Hay asuntos que no pueden contarse en un periódico, porque podrían tomarse como una ofensa a las buenas costumbres, al buen gusto o a la moral, por ejemplo. El léxico de lo no público no aparece en la prensa.
- El concepto de noticia suele restringirse a lo espectacular, a lo que rompe lo habitual. Hay determinados ámbitos en que ocurre lo que se entiende por noticias. El léxico de los otros ámbitos tiene mucho menores posibilidades de aparición.
- Los destinatarios normalmente son plurales, indefinidos (salvo en el caso de los boletines, que se dirigen específicamente al público homogéneo de un sindicato, de una población, de un centro cultural). Esto significa una predominancia del léxico de mayor extensión y menor comprensión.
- No cualquier persona es productora de textos periodísticos. El hablante es ser alfabeto, por lo general, tiene estudios secundarios o universitarios, y casi sin excepción pertenece a la cultura urbana. Luego comúnmente no aparece en la prensa el léxico popular ni el léxico rural.
- La función primordial es la referencial, informativa: contar sucesos. En segundo término es interpelativa, exhortativa, trata de impulsar a la acción. Muy rara vez es expresiva. Por ello, debieran tener mucho menor frecuencia las formas verbales y los pronombres personales de primera persona.

Estas posibles deficiencias se anulan cuando se consideran otras fuentes escritas, como libros, folletos, discursos y especialmente textos orales variados.

SUBCORPUS: LA PRENSA ESCRITA

Para probar la factibilidad de los métodos en un subconjunto homogéneo, decidimos trabajar con los textos periodísticos escritos.

Son 1.001 textos formados por 525.079 p-t. En promedio tienen 525 p-t.

Hemos procesado:

566 textos de prensa escrita	278.187 p-t	(promedio 491 p-t)
25 programas radiales	35.149	(promedio 1.406 p-t)
10 programas de televisión	7.249	(promedio 725 p-t)
400 boletines	204.494	(promedio 511 p-t)

Estos 1.001 textos están digitados en 60.010 líneas.

Si no consideramos antropónimos, topónimos, diálogos en otros idiomas, las 525.079 palabras-textuales totales se reducen a 489.750, que constituyen nuestro primer objeto de estudio. Quedan 35.329 que no se procesan.

Método de trabajo

Para la clasificación gramatical empleamos el parser elaborado originalmente por Luis Fernando Lara y María Isabel García Hidalgo para la constitución del Corpus del Español de México y el *Diccionario del Español de México*[13]. La profesora García Hidalgo, con mucha generosidad, nos está haciendo una versión para computador personal adaptada a nuestras necesidades[14].

El programa pudo reconocer y clasificar 257.490 p-t (53%), lo que es un buen resultado, que de todas maneras nos deja con 232.260 (47%) que tenemos que clasificar y lematizar manualmente. Para tener una idea de lo que esto significa en tiempo podemos tomar la siguiente base: si tardáramos un minuto por p-t, nos tomaría 483 días de ocho horas de trabajo ininterrumpido, sin errores, ni pérdidas de archivos, situaciones que son habituales[15].

Para hacer el trabajo disponemos de dos computadores:

- un 386 DX, 25 mhz, 1 mega RAM, 80 megas de disco duro, que nos sirve para el ingreso y la corrección de los textos, y
- un 486 DX, 50 mhz, 8 megas RAM, dos discos duros con un total de 360 megas de memoria, uno para el procesamiento de los datos, y el otro como respaldo.

Las investigaciones lingüísticas exigen computadores de gran capacidad de memoria y muy rápidos, porque los procesos son repetitivos y, contrariamente a lo que sucede en otros campos, los resultados son enormemente más extensos que los datos iniciales: por ejemplo, transformar un texto de cien páginas en concordancias de tres líneas tiene un resultado de unas 2.400 páginas.

Con algo de exageración podría decirse que estamos trabajando artesanalmente un proceso que es industrial. Para aprovechar realmente la tecnología actual, tendríamos que haber partido por leer los textos escritos con

[13] Para los principios del parser, vid. García, 1979.

[14] El profesor Nelson Cartagena está preparando un corpus del español de España con otra variante del programa original de El Colegio de México.

[15] Este es sólo un ejemplo ilustrativo, no corresponde exactamente a la realidad.

un OCR, con lo que nos habríamos ahorrado toda la digitación y la corrección[16]. El ideal sería trabajar con dos o tres computadores en red, de modo de avanzar simultánea y no secuencialmente. Para solucionar el problema de que los computadores no están en el mismo lugar físico, hubiera sido conveniente disponer de un módem para el traspaso de archivos. El respaldo podría ser más seguro y mejor en cinta, para no pensar en procedimientos más sofisticados[17].

Estado actual[18]

- La jornada de trabajo con los computadores es de 20 horas diarias.
- En la lematización y clasificación manuales seguimos el orden de los textos digitados, línea por línea, de la línea 1 a la 60.010. Diariamente trabajamos unas 400 p-t nuevas.
- En la noche se revisan todas las líneas siguientes en busca de las mismas formas gráficas encontradas durante el día. Si aparecen, se les ponen provisoriamente el lema y la categoría gramatical más probables, como una manera de adelantar el trabajo siguiente. Cuando vuelvan a aparecer siguiendo el orden se rectificarán si fuera necesario.
  En promedio se necesitan unos 3 minutos para comparar una forma gráfica con los 232.260 registros de palabras-textuales, poner el lema y la categoría gramatical en cada una de las palabras-textuales que tienen la forma gráfica.
  En la noche se trabaja un promedio de 250 formas gráficas, que en este momento corresponden a unas 500 palabras-textuales. Ya han aparecido las formas gráficas más frecuentes y cada vez irá disminuyendo más la frecuencia hasta que quedemos con los hápax, que son muy abundantes.
  Actualmente ya están lematizadas y clasificadas en forma provisoria 210.000 p-t.

[16] Tuvimos una generosa oferta del colega Max S. Echeverría (Universidad de Concepción) para que úsaramos su OCR, pero la llegada del aparato tardó más de lo que podíamos esperar. En futuros trabajos podrá aprovecharse la experiencia de esta universidad.

[17] Justo es reconocer que no tengo memoria de haber tenido alguna vez en el país todo lo que se necesitaba para una investigación, lo que, mirado desde un punto de vista exageradamente optimista, es un acicate para la imaginación, la fantasía y la creatividad. En el caso actual, es muy satisfactorio el sólo hecho de tener apoyo para emprender una investigación de este tipo, que es a largo plazo.

[18] El detalle en la exposición tiene el fin didáctico auxiliar de ilustrar a quienes todavía piensan que el esfuerzo es mínimo si uno se ayuda con computadoras, porque ellas son las que lo hacen todo, y que basta con apretar un botón mágico para tener todos los resultados imaginables. Lo cierto es que se trabaja mucho más, las exigencias son mayores, pero uno puede proponerse metas de otro modo inalcanzables.

- Como decíamos, cuando vuelven a aparecer las formas gráficas ya tienen un lema y una categoría provisorios. En un porcentaje de los casos, el investigador hace los cambios exigidos por el contexto específico entregado por la concordancia, por ejemplo, una categoría o un lema por otro. Esto es frecuente en los usos de substantivo (8) –adjetivo (2), adjetivo (2) –adverbio (1), participio (9) –adjetivo (2), formas verbales (9) que coinciden con substantivos (8).
- Se van anotando todos los errores de digitación que serán corregidos de una vez en una etapa posterior.
- En la lematización-clasificación, rectificación de lemas y categorías, anotación de errores se ha alcanzado la línea 20.000. En una segunda vuelta (revisión) hemos llegado a la línea 6.000.
- Paralelamente se continúa con los trabajos que tienen relación con el corpus total: selección, grabación, transcripción, digitación, corrección de los textos que serán incorporados para alcanzar la meta de dos millones de p-t; preparación de otros subconjuntos que serán procesados más adelante, como, por ejemplo, entrevistas generales y entrevistas jergales.

Trabajos futuros

Una vez que se alcance la línea 60.010, quedarán todavía varias tareas pendientes:

- Revisar unos quinientos casos como *que* (pronombre o conjunción) o *la* (artículo o pronombre) que resulta más eficiente trabajar en conjunto.
- Comparar los resultados de la lematización y clasificación automáticas y manuales, hacer los ajustes necesarios y fundir ambas tablas en una sola.
- Corregir de una vez todos los errores detectados en la digitación.
- Se ve como necesario optimizar el parser: podría incorporarse un pequeño diccionario con las voces no reconocidas y que han demostrado ser unívocamente substantivos o verbos; a las formas que tienen más de una función gramatical podría agregárseles el lema y la categoría estadísticamente más frecuentes, lo que eliminaría el trabajo nocturno actual.

Utilización del subcorpus

Completado totalmente el análisis del subcorpus habría que emprender de inmediato algunas tareas que permitan su mejor utilización, además de aquellas que preparan la elaboración del resto del corpus:

- Determinación de los conjuntos flexionales presentes en el subcorpus:
  a) *Lema*,
  b) *especificador gramatical*,
  c) cada una de las *formas funcionales* que integren el conjunto denominado por el lema, y
  d) indicaciones estadísticas: *frecuencias* (totales de todo el conjunto flexional, porcentaje con respecto al total, individuales totales de cada forma gráfica, de cada una de las cinco variables que consideramos en los hablantes, de lengua oral/escrita, de cada tipo de texto, de cada texto individual, etc.).
- Determinación de los *rangos* de los conjuntos funcionales, representados por los lemas, y de las formas textuales.
- Comparación con otros corpora en número de conjuntos, composición, frecuencia: la lengua periodística del Corpus del Español en México, PE77 de Gotemburgo, 205 entrevistas del lenguaje oral del español de México empleadas por Raúl Avila (El Colegio de México), etc.
- Elaboración del mismo subcorpus con otros sistemas de párser o de tratamiento de corpora como TACT o Exégesis a fin de establecer pros y contras.
- Preparación de los criterios para determinar el léxico básico o común del español de Chile: frecuencia, distribución en tipos de textos y en tipos de hablantes.
- El material reunido es suficiente como para aprovecharlo en trabajos puntuales, no sólo léxicos, como por ejemplo: análisis de los conjuntos flexionales más frecuentes o de algunos campos léxicos[19], rección verbal, lexicogenesia: derivación[20], composición, reducción, siglificación, elipsis,...[21], coocurrencias (colocaciones), análisis sintáctico[22].

CONCLUSIONES

El procesamiento del subcorpus *textos periodísticos escritos* representa aproximadamente un quinto del trabajo total del corpus. Habrá que agilizar los métodos. Llevamos cuatro años de trabajo. Otros seis serían un plazo

[19] Hemos hecho algunos avances aprovechando los materiales preliminares: el léxico de la vida sociopolítica y de los derechos humanos (SÁEZ, 1990); el léxico de la droga (SÁEZ, 1993c).

[20] Ya hemos utilizado los materiales preliminares en unos artículos sobre los sufijos *-ton* y *-teca* (SÁEZ, 1991) y el complejo sufijal *-iz-* + *-ar* (SÁEZ y WAGNER, 1992).

[21] Sobre este tema, véase SÁEZ (1992).

[22] Los trabajos sintácticos lo tendrán bastante avanzado porque cada p-t estará provista de su especificador funcional, de modo que cada texto podría reemplazarse por las secuencias de especificadores, lo que permitiría seguir determinadas configuraciones sintácticas.

aceptable. Para ello, como decíamos, habrá que optimizar el parser y trabajar simultáneamente cuatro subconjuntos: *libros y folletos*, *textos escritos (resto)*, *entrevistas*, *textos orales (resto)*[23].

El trabajo de cada subconjunto nos dejará una cantera de futuras investigaciones:

- una base de datos interactiva,
- una colección de unos tres mil textos que pueden ser procesados computacionalmente,
- concordancias completas de cada una de los dos millones de palabras textuales.

Además tendremos subconjuntos acumulativos y podremos ampliar los que nos parezcan más productivos, o crear nuevos.

Es una recogida significativa de material léxico que no tenemos hasta el momento, especialmente importante en la lengua oral. Es, para decirlo en términos pesqueros, una suerte de "pesca de arrastre" donde se recogen merluzas, corvinas, congrios, jaivas, lenguados, blanquillos, pejesapos, anguilas. No se opone a la pesca al espinel. Ambos métodos son complementarios. Pero para la descripción léxica de un dialecto, ya no puede prescindirse de esta pesca de arrastre.

El material recogido permitirá que los especialistas trabajen con lengua viva, no con ejemplos construidos artificialmente sobre la base exclusiva de su idiolecto. Permitirá respaldar o rechazar sus intuiciones, siempre insustituibles. Entregará una primera aproximación a lo que podamos entender como *léxico básico del español de Chile* y de allí proyectarse al pan-español mediante las intersecciones con corpora semejantes de otros dialectos. Introducirá en problemas lexicológicos, morfológicos o sintácticos indicaciones de frecuencia de que no disponemos. Posibilitará estudios, como los de coocurrencias, hasta el momento inéditos en el país, porque requieren grandes bases de datos.

Hay que tener conciencia de que todo esto, que es una gran ayuda, también puede convertirse en un entorpecimiento. No es lo mismo enfrentarse a algunas muestras tomadas al azar de unas cuantas novelas o periódicos que analizar, por ejemplo, 100.000 *que*. En los casos en que es humanamente imposible la elaboración del material, siempre queda la solución de que la computadora haga la selección aleatoria que el investigador estime conveniente. De todas maneras, tendrá la ventaja de que se deja afuera la posibilidad de una elección sesgada por los gustos del investigador.

[23] Siguiendo una larga tradición en los trabajos lexicológicos y lexicográficos y según lo ya realizado, este plazo es sumamente optimista. Además, depende de muchas variables que no dominamos: que podamos controlar la hipertensión; que sigamos teniendo apoyo económico, una infraestructura adecuada, un equipo estable; que haya rápidos progresos en lingüística computacional y que podamos utilizarlos. Por otra parte, lo normal entre nosotros es que estos trabajos desgraciadamente no sean la principal actividad del investigador.

CIECh - INFORME 19 / noviembre de 1993

1. TEXTOS ESCRITOS

Categoría	Código	Código	Textos Digitados	Textos Total		Palabras-textuales frecuencia	promedio
Prensa			1.001			525.079	525
	escrita	epoe	001-566	566	(160)	278.187	491
oficial	radio	epor	001-025	25	(30)	35.149	1.406
	T.V.	epot	001-010	10	(60)	7.249	725
no oficial		epno	001-400	400	(400)	204.494	511
			96			46.208	
Varios		epev	001-085 090-096	92		33.965	369
Discursos		endi	001-004	4		12.243	3.061
Libros y folletos		el	001-250	250	(207)	250.207	1.001
			237			94.665	
T. privados		etmi	001-207	207	(120)	86.122	416
		etmb	001-030	30		8.543	285
		etd	001	1		1.760	
Canciones			119			22.729	191
		eoca	001-113	113	(80)	20.865	185
		eicb	001-006	6	(40)	1.864	311
Poesía poblacional		elpp	001-040	40	(40)	8.230	206
		Total textos:		1.744	(1.137)	*Prom.* 544 p-t	
		Total palabras textuales:		948.878	(1.005.300)		
En digitación		ECOM 001-008					

CIECh 19 / noviembre de 1993 (Cont.)

2. TEXTOS ORALES

	Código	Textos digitados			p-t	Prom.
Entrevistas reproducidas	onei	001-123 129-170	164	(175)	227.627	1.388
Entrevistas directas gen.	oned	001-100	100	(175)	136.391	1.364
Entrevistas jergales	onej	001-159	159	(100)	182.646	1.149
Programas de radio	opor	001-053	53		55.301	1.043
Programas de televisión	opot	001-021	21		31.568	1.503
Conversaciones informales	onci	006-037	32		77.017	2.407
Discursos-disertaciones	ondi	001-019	19		25.851	1.361
Conversaciones telefónicas	onct	001-041	41		10.505	256
Videos	oncv	001-006	6		22.339	3.723
Narrativa	olpn	001-019	19	(20)	19.010	1.001
Payas	olvy	001-005	5	(6)	3.123	625
Poesía	olvp	001-050	50		13.807	277
	Total textos:		669	(556)	*Prom.:* 1.204 p-t	
	Total palabras textuales:		805.185	(999.000)		

Total general de textos digitados:	2.413	(1.553)	Promedio p-t: 728
Total general de palabras-textuales:	1.754.063	(2.004.500)	

* Los paréntesis indican el plan originario.
* Las frecuencias indicadas no tienen descartes.

Faltan:	onei 124-128
En digitación:	oned 101-120 onej 160 ondi 020-022

BIBLIOGRAFIA CITADA

BURDACH, Ana María; Ana María MILLÁN y Marisela TOSELLI. 1992. "El recurso de referencia según el modelo de cohesión textual de Halliday y Hasan: aplicación a una manifestación del discurso escrito en español", *RLA*, 30: 97-117.

GARCÍA HIDALGO, María Isabel. 1976. "La formalización del Analizador Gramatical del DEM" en Lara *et al.*, 1979, 85-115.

LARA, Luis Fernando; Roberto HAM CHANDE, María Isabel GARCÍA HIDALGO. 1979. *Investigaciones Lingüísticas en Lexicografía*, México, El Colegio de México, 266 pp.

MAYORGA, Dora y Hugo CIFUENTES. 1992. "Algunas claves para la interpretación de un texto periodístico", *Nueva Revista del Pacífico*, 37: 7-30.

SÁEZ GODOY, Leopoldo. 1988. "Los inventarios léxicos automatizados y el español: proposiciones terminológicas" *Literatura y Lingüística* 1: 67-82.

—— 1989. "Desvíos de la norma culta en la prensa escrita de Chile: barbarismos y solecismos 'mercuriales'", *Literatura y Lingüística* 2: 105-134.

—— 1990a. "Novedades en el español de Chile (1973-1989) (Neologismos en el léxico de la vida sociopolítica y de los derechos humanos", *Literatura y Lingüística* 3: 117-151.

—— 1990b. "Corpus Integral del español de Chile", *Actas del Octavo Seminario Nacional de Investigación y Enseñanza de la Lingüística*, Santiago, Universidad de Santiago de Chile, 94-107.

—— 1991. "*-ton* y *-teca* en el español de Chile", *Literatura y Lingüística* 4: 129-140.

—— 1992. "Economía en el español de Chile (1979-1992): Elipsis, Aglutinación, Siglificación, Reducción y Abreviación", *Literatura y Lingüística*, 5: 161-196.

—— 1993b. "El Corpus Integral del español de Chile (CIECh). Estado actual", *Actas del IV Congreso del español de América*, Santiago (en prensa).

—— 1993c. "El léxico juvenil de las drogas en Chile II", *Literatura y Lingüística* 6 (en prensa).

SÁEZ GODOY, Leopoldo y Claudio WAGNER. 1993. "Un complejo sufijal productivo: *-iz-* + *-ar* en el español de Chile", *Estudios Filológicos*, 27 (1992): 29-42; 28: 97-122.