Onomázein
ISSN: 0717-1285
onomazein@uc.cl
Pontificia Universidad Católica de Chile
Chile

Cid Uribe, Miriam E.; Ross Arias, Paula

LA CONSTRUCCIÓN DE UN CORPUS DE HABLA PÚBLICA DE CHILE: CRITERIOS Y PROCEDIMIENTOS PARA LA SELECCIÓN DE UNA MUESTRA REPRESENTATIVA

Onomázein, núm. 13, 2006, pp. 21-33
Pontificia Universidad Católica de Chile
Santiago, Chile

Disponible en: http://www.redalyc.org/articulo.oa?id=134516555002

ONOMÁZEIN 13 (2006/1): 21-33

LA CONSTRUCCIÓN DE UN CORPUS DE HABLA PÚBLICA DE CHILE: CRITERIOS Y PROCEDIMIENTOS PARA LA SELECCIÓN DE UNA MUESTRA REPRESENTATIVA[1]

Miriam E. Cid Uribe
Paula Ross Arias
Pontificia Universidad Católica de Chile
mcidu@uc.cl - pross@uc.cl

Resumen

Para la realización del proyecto de investigación "Patrones prosódicos recurrentes en los actos de habla pública de Chile: descripción fonofonológica" se ha hecho necesario construir un corpus de habla que contenga manifestaciones de dichas instancias de ocurrencia en distintas situaciones comunicativas que favorezcan el uso de la variedad definida como habla pública. En esta comunicación se describen, en primer lugar, los procedimientos llevados a cabo con el fin de construir nuestro corpus de estudio; posteriormente, se definen los criterios que subyacen a la selección de la muestra y, finalmente, se presentan los procedimientos de etiquetación del corpus en su totalidad. El objetivo de esta comunicación es, pues, describir los procesos de elaboración de un corpus de habla pública, dar cuenta de los criterios de selección de la muestra utilizados y proponer una etiquetación holística que registre todos los componentes que identifican distintivamente cada una de las instancias comunicativas que componen el corpus base de nuestra investigación.

Palabras clave: corpus; etiquetado; representación holística; actos de habla.

[1] Proyecto FONDECYT Nº 1030953, en realización en el Departamento de Ciencias del Lenguaje de la Facultad de Letras de la Pontificia Universidad Católica de Chile.

Fecha de recepción: marzo de 2006
Fecha de aceptación: abril de 2006

Abstract

For carrying out the research "Recurrent prosodic patterns in Chile the speech acts of Chile public speech: phonophonological description", it was necessary to build a corpus showing manifestations of such acts in different communicative situations. In this paper we, firstly, describe the procedures undertaken to build the corpus; secondly, those criteria underlying the sample selection are defined and, finally the labelling procedures are described. The objective of this paper is, then, to account for the process of corpus collection and to propose a holistic labelling that duly contains for all the components that distinctly identify each of the communicative instances that make up our corpus.

Key words*: corpus; labelling; holistic representation; speech acts.*

1. INTRODUCCIÓN

El proyecto "Patrones prosódicos recurrentes en los actos de habla pública de Chile: descripción fonofonológica", en actual realización en la Facultad de Letras de la Pontificia Universidad Católica de Chile, tiene como objetivo principal describir la relación entre los subsistemas prosódicos recurrentes en el habla pública de Chile y los actos de habla que los determinan. Para lograr este objetivo se hace necesario recurrir a un corpus para cuya recopilación nos hemos basado en la caracterización del concepto operacional de habla pública que, como equipo de investigadores, postulamos y la cual describimos como sigue: "El habla pública es aquella actividad oral que realizan determinadas personas desde su papel institucional, cuya expresión idiomática tiende a un estilo más bien formal, sin que por ello se excluyan otros registros; que está dirigida, además, a auditorios colectivos reales o virtuales; que es producida en un espacio semiótico público o escenario público, y cuyos temas son de naturaleza también pública y de interés general". (Cid, M.E. *et al.*, *Onomázein* 10 [2004/2]: 179-184). El concepto operacional así descrito pareciera, entonces, alcanzar su máxima frecuencia de ocurrencia en instancias comunicativas cuyo canal de transmisión es preferentemente medial. Por lo anterior, el proceso de construcción del corpus sobre el cual basamos nuestra investigación ha hecho necesaria, en gran medida, la recopilación de muestras según ellas se producen y transmiten a través de los medios, sobre todo en la actualidad, de forma tal de alcanzar el auditorio colectivo meta.

Hemos concebido nuestro corpus siguiendo, en gran medida, los postulados de J. Llisterri (1997) para quien un corpus es "un conjunto estructurado de materiales lingüísticos en el que se distinguen diversos modelos de representación correspondientes a diferentes grados de elaboración de los datos que lo constituyen". En el caso de nuestra investigación, el corpus que recolectamos es oral; por lo tanto, en

el tratamiento del mismo delimitamos los niveles de representación necesarios para su utilización como base de los análisis lingüísticos planteados en nuestro proyecto de investigación. Los niveles a que hacemos referencia son, a grandes rasgos, el discursivo y el fonético. El primero, porque es de ese nivel de análisis desde donde extraeremos los datos imprescindibles que nos permitan proponer una matriz discursiva de los actos de habla según ellos se manifiestan en el habla pública de Chile; este nivel requiere, evidentemente, una representación ortográfica. El segundo nivel que nos ha parecido del todo imprescindible es el fonético, porque de él extraeremos la información prosódica necesaria para relacionar los patrones suprasegmentales presentes en el habla pública de Chile con la matriz discursiva ya mencionada; este nivel requiere una representación prosódica que dé cuenta de la recurrencia y comportamiento de los patrones encontrados.

2. EL CORPUS

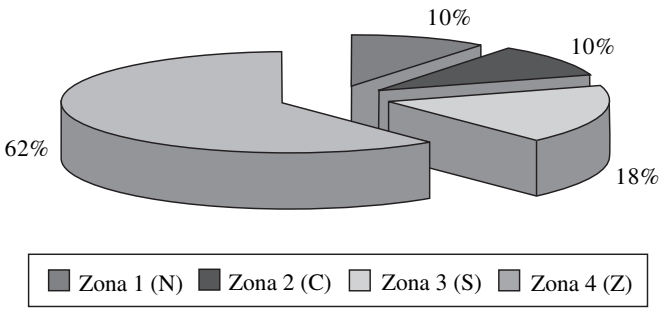
Comenzaremos por describir nuestro corpus el cual consiste en aproximadamente 25 horas de duración y fue recogido siguiendo una serie de criterios que a continuación se presentan.

2.1 Criterios de recopilación del corpus:

I. **Zona geográfica**: Siguiendo lo propuesto en el proyecto internacional SALA (Speech Across Latin America) 2000, Chile fue dividido en cuatro zonas geográficas tomando como base la densidad poblacional del país. Dichas zonas son: **Zona 1, Norte** y comprende las Regiones 1ª, 2ª, 3ª, y 4ª; **Zona 2, Centro** que incluye las Regiones 6ª y 7ª; **Zona 3, Sur** a la que pertenecen las Regiones 8ª, 9ª, 10ª, 11ª y 12ª; y, finalmente, la **Zona 4** formada por el **Área Metropolitana** y **Valparaíso**.
En términos de tiempo, el porcentaje de participación de cada zona en el corpus se detalla como sigue:

Zona 1	Norte	: 3.5 horas (14% del total del tiempo de grabación)
Zona 2	Centro	: 3.5 horas (14% del total del tiempo de grabación)
Zona 3	Sur	: 5.0 horas (21% del total del tiempo de grabación)
Zona 4	RM/Valpo	:13.0 horas. (50% del total del tiempo de grabación)

En términos de cantidad de informantes según zona de procedencia, la participación se grafica como sigue:



Como se puede apreciar, el porcentaje de participación de informantes según zona es directamente proporcional al porcentaje de tiempo de grabación por zona presente en el corpus, ya que –ordenadas de menor a mayor proporción– las zonas Norte y Centro reúnen el 10% del total de informantes presentes en el corpus, proporción que aumenta para la zona Sur con un 18% y que presenta su mayor participación en la Zona Región Metropolitana y Valparaíso, concentrando más de la mitad de los informantes de todo el país; lo anterior tomando en consideración la premisa de que parte de la población de estas regiones es originaria de una de las zonas restantes.

II. **Ámbito**: Identifica al agente emisor de habla pública según la dimensión del quehacer social al que pertenece. Los ámbitos considerados así como los agentes y escenarios que los componen, son:

- **Académico/cultural**: docentes y paradocentes (bibliotecarios, etc.), miembros académicos (Universidades, Academias de Lengua, de Ciencia, etc.), representantes del mundo artístico (actores, cantantes, escritores, etc.) en situaciones de:
  - Inauguraciones de años
  - Discursos de despedida y/o agradecimiento
  - Clases magistrales
  - Mesas redondas, paneles, debates
  - Entrevista medial
  - Presentación y análisis de obras literarias
  - Etc.

- **Castrense**: miembros de las Fuerzas Armadas, de Orden y Seguridad y Policía de Investigaciones en situaciones de entrevista medial.

- **Cívico/político**: directivo o representante de una institución pública y/u ONG; miembro de la clase política; representantes de organismos gubernamentales pertenecientes al Poder Ejecutivo, Judicial y Legislativo en situaciones de:

  – Discursos
  – Debates
  – Entrevistas
  – Comentarios
  – Mesas redondas, paneles

- **Empresarial**: representantes del mundo empresarial en situación de:

  – Panel
  – Entrevista
  – Comentario

- **Gremial/sindical**: un representante, elegido por su colectivo o asumiendo su representación, en situaciones de:

  – Entrevista medial (radio, TV)
  – Panel
  – Comentario
  – Discurso
  – Mesa Redonda
  – Debate

- **Medial**: trabajadores de los medios de comunicación –radio, televisión– en situaciones de:

  – Lectura y comentarios de noticias
  – Comentarios de la actualidad contingente
  – Debates
  – Paneles
  – Discursos

- **Religioso**: miembros del clero y/o representantes de diferentes credos y/u organizaciones ecuménicas en situaciones de:

  – Entrevistas mediales
  – Sermones, homilías, lectura de oraciones.

III. Género discursivo: Este componente tiene, a nuestro juicio, una injerencia fundamental a la hora de caracterizar un corpus de habla

pública, ya que en cada intercambio oral el género discursivo pareciera favorecer la ocurrencia y recurrencia de ciertos actos de habla por sobre otros. En el caso de nuestro proyecto y para la consecución de nuestros objetivos, hemos considerado siete géneros discursivos claramente identificables.

- **Clase Magistral**: Aquella instancia formal a través de la cual un informante, generalmente perteneciente al mundo académico y/o cultural, elabora en torno a un tema determinado en una instancia formal de comunicación. El discurso puede ser pauteado o no pauteado. En el primer caso, el hablante lee, aunque puede ocasionalmente expandir en forma espontánea, y en el segundo, el hablante elabora sin apoyo de texto escrito.

- **Comentario de actualidad y lectura de noticias**: En este género discursivo se han incluido tanto los comentarios en torno a un tema dado realizados a través de los medios como la lectura de noticias, las que, en la actualidad, tienden a contener comentarios y/o apreciaciones periodísticas relativas a un evento de actualidad. Como en el caso anterior, el habla producida es el resultado tanto de la lectura como de comentarios espontáneos.

- **Debate/panel**: También un género discursivo importante a la hora de caracterizar interacciones verbales en torno a un tema central, tiene una incidencia relevante en este corpus, por la riqueza lingüística esperada en intercambios orales en los que los aspectos de toma y entrega de turnos, interrupciones, cambios de ideas, pausas llenas y vacías, uso de rellenos para hacer tiempo y elaborar ideas parecieran presentar características particulares. Como es natural, este género se ve favorecido en intercambios de tipo polilógicos.

- **Discursos**: Instancias discursivas en formato monológico, se esperaría encontrar en ellos la utilización de un registro formal que favorezca actos de habla claramente identificables. Corresponden, invariablemente, a lecturas.

- **Entrevista**: Recopiladas desde los medios, las muestras de habla recogidas apuntan a elicitar una producción oral espontánea en situaciones comunicativas dialógicas. La espontaneidad esperada nos provee con un habla no pau-

teada y, por ende, caracterizada por el uso de rasgos de simplificación, partidas falsas, repeticiones, interrupciones, cambios de ideas, entre otros.

- **Mesa Redonda**: Muy semejante al género discursivo anterior, la diferencia estriba en la esperada, aunque no siempre realizada, calidad polilógica de la interacción.

- **Prédica/Sermón/Oración**: Producida exclusivamente por informantes del ámbito religioso, se espera un registro altamente formal caracterizado por el uso de fórmulas lingüísticas establecidas y en situaciones sociales no interactivas. Corresponden en su totalidad a lecturas.

IV. **Género**: En la medida de lo posible, se trata de favorecer un número similar de informantes masculinos y femeninos con el fin de no distorsionar la descripción. Como es de esperar, hay ciertos géneros discursivos y ciertos ámbitos que favorecen la producción oral masculina o femenina, pero en general se ha propendido a un equilibrio entre los informantes masculinos y los femeninos.

V. **Grupo etario**: Para efectos de este corpus, nuestros informantes han sido seleccionados tomando en cuenta los siguientes criterios:

- **Grupo 1**: Los informantes pertenecientes a este grupo van desde los **18 a los 35** años. Se espera obtener información que refleje el grado de dinamismo de la lengua que debería estar exento de rasgos de fosilización lingüística.

- **Grupo 2**: A este grupo pertenecen aquellos informantes cuyas edades fluctúan entre los **36 y los 55 años**, etapa en la que la competencia lingüístico oral ha alcanzado un nivel de proficiencia correspondiente a la norma estándar y en la cual comenzarían a aparecer algunos rasgos de fosilización. De este grupo de informantes se esperaría una mayor y más amplia disponibilidad léxica.

- **Grupo 3**: Formado por representantes de habla pública que tienen **56 años y más,** en este grupo se realizan instancias de habla que se acercan a las descripciones lingüísticas tradicionales correspondientes a la norma estándar en sus registros formal e informal.

2.2. La muestra: Criterios de selección

Un corpus con las características que hemos descrito hasta aquí, que es oral, que tiene una longitud de alrededor de veinticinco horas de duración, que, además, recoge información de hombres y mujeres de distintas procedencias geográficas, entre muchos otros factores, es, a nuestro juicio, un muy buen referente de análisis a la hora de caracterizar el habla pública de Chile. Sin embargo, para llegar a esa caracterización se hace necesario establecer ciertos criterios que nos permiten llevar a cabo un proceso de selección de las muestras; la aplicación de dichos criterios nos asegura que el cuerpo de análisis reunido en la muestra es, por una parte, representativo del habla bajo estudio y, por otra, lo suficientemente amplio como para posibilitar los distintos estudios parciales que surgen, necesariamente, de su análisis. Nos referimos a aquellos estudios tanto de tipo discursivo como prosódico de los cuales ha ido emergiendo, en forma paulatina, la caracterización prosódico-discursiva del habla pública de Chile.

La selección de nuestra muestra de análisis fue realizada considerando, en primer lugar, la **representatividad**. Para ello nos remitimos a lo usual en estos casos; es decir, se seleccionó un 20% del total del corpus. En segundo lugar, se dio importancia a los actos de habla y para ello seguimos la taxonomía de Searle. Desde este punto de vista, a cada enunciado se le adscribió uno o varios microactos de habla, esto con el objeto de proponer una matriz discursiva a la cual se superpondrá el componente prosódico de forma tal de obtener la caracterización prosódico-discursiva del habla pública de Chile. Por último, cabe destacar que cada uno de los criterios de recopilación del corpus se ve representado en la selección de la muestra.

3. TRANSLITERACIÓN Y ETIQUETADO DEL CORPUS

3.1 Transliteración del corpus

Es de destacar que, por ser el nuestro un material consistente en una colección de muestras orales de habla, se evidencia, desde el comienzo, la necesidad de tomar decisiones fundamentadas respecto a, por lo menos, dos aspectos importantes que tienen, a nuestro juicio, una incidencia directa en la expedición de acceso a la información recolectada y que, por lo mismo, ameritan una reflexión profunda. Nos referimos a los procesos de **transliteración** y de **etiquetado** de los materiales recolectados. En este apartado nos referiremos en detalle al proceso de transliteración del corpus para cuya realización hubo

que recurrir a una serie de convenciones tipográficas para representar rasgos propios de la oralidad. A continuación listamos las convenciones de transliteración que utilizamos:

- El cambio de estructura gramatical que refleja un cambio de idea en medio del discurso, se marca con el signo **~**
- Después de ~ va minúscula.
- La interrupción del discurso del hablante se marca con el signo #. Si el hablante sigue desarrollando su idea luego de la interrupción, el nuevo turno comienza con minúscula.
- Después de punto suspensivo se parte con mayúscula.
- La transliteración de guarismos, siglas y abreviaturas se realiza íntegramente según su pronunciación. Así, por ejemplo: ONU, mil novecientos noventa y ocho, etcétera. Pero, De Cé, Pe Pe Dé, Erre Ene.
- Para demostrar elisión de sílabas en posición final, utilizamos el apóstrofo, el cual puede coincidir o no con una sílaba completa. Así, pues, para, nada y sus variantes alomórficas podrán ser: p', po', pu', pa', na'.
- Las vacilaciones al interior del discurso son transliteradas según su pronunciación, correspondan ellas a sílabas independientes o no; por ejemplo, eh, m, ah, etcétera.
- Los marcadores discursivos del tipo no es cierto, digamos, claro, bueno, etcétera van separados por comas. Si la señal acústica así lo refleja, estos marcadores van entre signos de interrogación o de exclamación.
- Las repeticiones se transliteran fielmente según han ocurrido sin separar cada una de ellas por comas. Las repeticiones pueden ser de segmentos, sílabas, palabras, frases e incluso oraciones.
- Las expresiones del voseo verbal propias del español de Chile son respetadas en términos de fidelidad fónica.
- Los títulos de libros, los nombres de obras de arte, los nombres de revistas, diarios y programas mediales van entre comillas simples y con mayúscula inicial.
- Los neologismos y extranjerismos van en cursiva.
- Las citas textuales y autocitas irán precedidas de dos puntos e iniciadas por comillas y mayúscula.
- Un fragmento indescifrable de la grabación se representa de la siguiente forma: (…)
- Aquella parte del discurso que no interesa para el objetivo de la investigación, se representa de la siguiente forma: […]

- Aquella parte del discurso que no es analizada, pero que es necesaria para la comprensión global del evento discursivo se translitera en su totalidad, pero entre paréntesis cuadrado.

3.2 Etiquetación del corpus

Una vez transliterado el corpus, se hace necesario identificar las emisiones de habla pública a través de una etiqueta que considere cada uno de los aspectos que las configuran como variedad además de los criterios utilizados para su recopilación.

Como se mencionó anteriormente el género discursivo es, a nuestro juicio, un componente esencial en la caracterización de cualquier emisión, ya que pareciera favorecer determinadas combinatorias de actos de habla y da cuenta del escenario en la que el evento comunicativo se enmarca. Es por esta razón que este componente, identificado con una letra mayúscula, ocupa el primer lugar en la etiqueta.

En segundo lugar, se identifica al agente emisor de la variedad en estudio según su pertenencia a determinado ámbito del quehacer social. Vale la pena mencionar aquí que la esfera social en la que el agente se desempeña de manera regular se complementa, a la hora de definir el ámbito, con el rol institucional que este asume al momento de producir la emisión. Por ejemplo, un actor que habla en nombre de su colectivo es etiquetado en el ámbito gremial/sindical y no en el ámbito académico/cultural; por su parte, si un informante pertenece al mundo político, pero en determinada circunstancia está hablando como empresario, el ámbito que prevalece es el empresarial por sobre el cívico/político. En el etiquetado, el ámbito se expresa mediante el uso de una letra minúscula en segunda posición.

El tercer componente del etiquetado identifica la zona geográfica a la que pertenece el hablante: la inicial de la zona escrita en letra mayúscula y entre paréntesis evidencia la zona de procedencia. La zona norte será (**N**), la zona centro (**C**) y la zona sur (**S**), en tanto que la zona 4, correspondiente a la Región Metropolitana y Valparaíso, la letra mayúscula que la identifica es (**Z**).

El cuarto componente de la etiqueta identifica al hablante según su pertenencia al género masculino o femenino: **f** para las mujeres y **m** para los hombres, ambas letras en minúscula.

El quinto componente de la etiqueta corresponde al grupo etario. Como ya se describió, se clasifica al hablante como perteneciente a uno de tres determinados rangos. Cada grupo se identifica con el número arábico 1, 2 ó 3, según corresponda.

Finalmente, se considera importante para nuestro etiquetado identificar el número del hablante en forma correlativa, de manera tal

de explicitar a través de la adjudicación de un dígito identificatorio el número de informantes que tenemos en el corpus dentro de cada una de las categorías. Para diferenciar este componente del grupo etario ambos números se separan por /. Es este elemento el último en la etiqueta y comienza desde el 1 y continúa hasta agotar el número en determinado ámbito y según determinado género, masculino o femenino.

El detalle de las siglas que representan cada uno de los criterios de recopilación del corpus hasta aquí descritos se presenta a continuación:

– **Género Discursivo**

A Clase Magistral
B Comentario de actualidad y lectura de noticias
C Debate / Panel
D Discurso
E Entrevista
G Mesa Redonda
H Prédica/Sermón/Oración

– **Ámbito**

a Académico/Cultural
b Castrense
c Cívico/político
d Empresarial
e Gremial/sindical
f Medial
g Religioso

– **Zona**

(N) Zona 1: Norte
(C) Zona 2: Centro
(S) Zona 3: Sur
(Z) Zona 4: Región Metropolitana y Valparaíso

– **Género**

f Femenino
m Masculino

– **Grupo Etario**

1 Grupo 1: de 18-35
2 Grupo 2: de 36-55
3 Grupo 3: de 56-

– **Nº de informante**

Desde número *1*, hasta lo que corresponda, según cada grupo etario precedido de **/** para diferenciarlo de 1, 2, 3 correspondientes al grupo etario.

En resumen, hemos considerado seis componentes que se ordenan de izquierda a derecha, representados por la siguiente combinación de caracteres:

– Género Discursivo: A, B, C, D, E, G, H
– Ámbito: a, b, c, d, e, f, g
– Zona: (N), (C), (S), (Z)
– Género: f, m
– Grupo etario: 1, 2, 3.
– Nº hablante: /1, /2, /3, /4, ...

Es así como, a modo de ejemplo, una muestra etiquetada como E.c.(N).m.2/6, se trata de una emisión enmarcada en el género discursivo entrevista, realizada por un agente perteneciente al ámbito cívico-político, proveniente de la zona norte, de género masculino y del segundo grupo etario. Dentro del ámbito cívico político, este informante es el número 6 con las características de zona, género y grupo etario:

E.c.(N).m.2/6: Nosotro+s estamos súper complicados, porque aparece en el diario de que+ habría de todo, y no es así. Eh, yo tengo una lista de, e+m, de todos los medicamentos que ~ falta comprimido inyectable; inclusive tela para hacer las gasas.

Por otra parte, la etiqueta D.a.(S).m.3/2 representa a un informante masculino del ámbito académico-cultural, que produce una emisión con características del género discurso, proveniente del sur de Chile y que pertenece al grupo etario tres. Dentro de su ámbito, el hablante ocupa el segundo lugar de la lista de informantes de esa zona, género y grupo etario.

D.a.(S).m.3/2: Una vez soñé el cielo como un lugar donde disponía de toda la placidez y de todo el tiempo del mundo para escribir. Hoy recibo esta honrosa distinción y escucho las palabras de Don Hernán respecto de algo que yo escribí. No en el cielo, en esta tierra.

El corpus que hemos descrito y que es la fuente de información desde la cual se realiza el proyecto de investigación FONDECYT Nº 1030953 consiste, entonces, en una gran base de datos compuesta por informantes identificados individualmente a través de una etiqueta que resumirá las características ya descritas y que

variará, principalmente, si cambia el género discursivo o el rol institucional asumido por el hablante en una situación comunicativa determinada.

REFERENCIAS BIBLIOGRÁFICAS

CID, Miriam, Paula Ross y José Luis Samaniego (2004). "Habla pública: hacia un nuevo concepto", en *Onomázein* 10 (2004/2): 179-184.

LLISTERRI, Joaquín (1997). "Transcripción, etiquetado y codificación de corpus orales", en J. GÓMEZ GUINOVART, A. LORENZO SUÁREZ, J. PÉREZ GUERRA y A. ÁLVAREZ LUGRÍS (eds.). *Panorama de la investigación en lingüística informática. RESLA, Revista Española de Lingüística Aplicada*, Volumen monográfico, 1999, págs. 53-82.