Luis Fernando Lara
Roberto Ham Chande
Ma. Isabel García Hidalgo

INVESTIGACIONES LINGÜÍSTICAS EN LEXICOGRAFÍA

Jornadas
89

EL COLEGIO DE MÉXICO

JORNADAS 89

EL COLEGIO DE MEXICO

CENTRO DE ESTUDIOS LINGÜÍSTICOS Y LITERARIOS

Luis Fernando Lara
Roberto Ham Chande
Ma. Isabel García Hidalgo

INVESTIGACIONES LINGÜÍSTICAS EN LEXICOGRAFÍA

JORNADAS 89
EL COLEGIO DE MÉXICO

Primera edición (2 000 ejemplares) 1979

Camino al Ajusco 20, México 20, D.F.

Impreso y hecho en México
Printed and made in Mexico

ISBN 968-12-0032-2

Indice

PROLOGO

La lexicografía ha sido considerada siempre como "el arte de componer léxicos y diccionarios". Disciplina tan antigua como el comercio y la guerra, su definición ha sido siempre utilitaria: la lexicografía *sirve* para relacionar pueblos de distintas lenguas y permitir su inteligibilidad mutua. Por ese carácter de utilidad al que modernamente se suma la competencia editorial y las necesidades del mercado, la lexicografía se ha mantenido siempre alejada de la ciencia del lenguaje; el lexicógrafo, preocupado por la colección minuciosa y, pese a todo, dilatada de palabras; presionado por el tiempo y las estrecheces económicas; abrumado por el exceso de trabajo y la falta de personas que compartan su extraño oficio, suele desconfiar de los comportamientos "científicos" de los lingüistas, para quienes los claustros universitarios hacen de su trabajo lingüístico una ocupación algo más muelle y algo menos sometida a otros intereses.

La lingüística, a su vez, es una ciencia joven si descontamos el pensamiento filosófico sobre el lenguaje como parte de ella; es decir, si hablamos de lingüística solamente desde que se comenzó a ejercitar un trabajo sistemático y concienzudo sobre las lenguas más que sobre las cuestiones ontológicas, metafísicas, lógicas, éticas y estéticas que preocupan esencialmente al filósofo. Como ciencia joven y además como aquella que, de entre las ciencias humanas, ha cosechado

éxitos más interesantes, ha pasado sus últimos sesenta años elaborando su instrumental teórico y metodológico con la vista puesta en otras ciencias que le sirven de modelo. Así, con un "espíritu de geometría", ha visto las lenguas desde una imprescindible abstracción sistematizante que, en la época actual, se revela en los mismos nombres de sus disciplinas: lingüística formal (gramática transformacional), lingüística matemática, lingüística teórica, semántica intensional (gramática de Montague), etc. Al lado de ella y en constante contrapunto y debate, han venido creciendo interdisciplinas que, aunque impugnadoras de aquel núcleo, se caracterizan todavía como subsidiarias: sociolingüística, dialectología, psicolingüística, etc.

Para el lingüista inmerso en su época, la lexicografía todavía no llega a ser siquiera interdisciplina; es "arte" para él como para el lexicógrafo. Sólo que si para el uno es un mote casi peyorativo, para el otro es una defensiva soberbia.

Gracias al interés de un grupo de personas preocupadas por el estado de la lengua española en México y con la ayuda del gobierno mexicano, en 1973 se creó un fideicomiso para la elaboración del *Diccionario del español de México* bajo el cuidado de El Colegio de México. Con esa ayuda se formó un equipo de investigadores que desde entonces ha venido realizando ese trabajo. El apoyo económico del fideicomiso y el terreno intelectual de El Colegio nos han permitido conjuntar la práctica lexicográfica —el "arte"— con el enfoque teórico y metodológico de la lingüística. Nos encontramos así en la todavía rara —por desgracia— posición de reunir nuestros intereses, de buscar planteamientos y soluciones consecuentes, de hacer de la lingüística y sus diferen-

tes disciplinas un instrumento práctico de trabajo y, al mismo tiempo, de sembrar de dudas, de someter a reflexión la ciencia del lenguaje con la vieja sabiduría práctica de los lexicógrafos. De ahí el título de esta colección de trabajos. Uno por uno son contribuciones de órdenes diferentes a la lexicografía y a la factura del diccionario; pero cada uno es también una propuesta para la lingüística. No se trata en ninguno de ellos de resultados definitivos. Son estados de nuestra investigación. Concluyen algo y comienzan otra cosa. Por ahora presentamos dos trabajos sobre estadística lexicológica; el primero de ellos se publicó en el tomo XXIII del órgano del Centro de Estudios Lingüísticos y Literarios de El Colegio de México, la *Nueva Revista de Filología Hispánica.* Fue un artículo programático, que fijó las bases y los límites del uso de la estadística para la elaboración de nuestro diccionario. En él no hemos alterado nada, excepto tres errores no de imprenta sino de nuestra máquina de escribir. A partir de esa "Base estadística del Diccionario del español de México" se presenta un primer resultado cuantitativo en "Del 1 al 100 en lexicografía". Así es que este artículo requiere del primero, del que suponemos difícil que haya llegado a todos los lectores que se interesen por este libro.

El tercero está dedicado a nuestro trabajo en lingüística computacional y es una explicación completa de lo que hicimos para que una computadora electrónica se convirtiera en uno de los más necesarios instrumentos de trabajo en lexicografía. Reconocemos la dificultad que debe presentar su lectura a quien no tenga experiencia en la formalización matemática y en la programación de las computadoras, pero a la vez creímos más

importante exponerlo así y no con un relato quizá más sencillo pero por eso mismo más expuesto a la superficialidad. El último es un estudio sobre semántica en lexicografía. A partir de él se han desarrollado las guías de trabajo para la definición lexicográfica; al mismo tiempo es un esbozo de teoría para la semántica lingüística.

Cada trabajo tiene un autor, pero es resultado de la colaboración real y no, como a veces pasa, de una ficción interdisciplinaria. El equipo del DEM tiene más miembros y cada uno de ellos ha aportado pensamiento, brega con los materiales y horas de discusión. A todos ellos les agradecemos su participación.

Hay muchas otras personas a quienes, desde ahora hasta que aparezca el diccionario, hemos de estar agradecidos. Pero entre ellos deseamos destacar la confianza que depositaron en nosotros y la ayuda que nos han dado (sin la cual —lejos de toda retórica— efectivamente no habríamos podido hacer nada) Antonio Carrillo Flores, Víctor L. Urquidi y Antonio Alatorre. También hacemos patente nuestro agradecimiento a Enrique Calderón, cuya visión nos permitió hacer un uso exitoso de la computadora electrónica.

L. F. L.
Ciudad de México, octubre de 1978.

Base Estadística del Diccionario del Español de México

Luis Fernando Lara
y
Roberto Ham Chande

0. El *Diccionario del Español de México*[1] es una obra lexicográfica cuyo objetivo fundamental es reflejar el léxico del español utilizado actualmente en el país, en cuanto "lengua nacional" y en cuanto a sus modalidades escritas y orales, cultas y coloquiales, urbanas y rurales.

Tal pluralidad de necesidades –tanto más obligatorias cuanto que el trabajo tiene una finalidad práctica inmediata: servir a un hablante mexicano medio como obra de consulta y como punto de referencia en su apreciación pre-científica del idioma– se enfrenta con algunos problemas que derivan de la naturaleza misma de la lexicografía: la necesaria objetividad de la descripción lingüística, la adecuación entre los métodos utilizados y la realidad de los fenómenos léxicos, el volumen de datos que requiere una definición lexicográfica completa, y los resultados que ha alcanzado hasta hoy la lexicografía española (y, en particular, la hispanoamericana). Este último, en razón de la influencia que pudieran tener sobre nuestro trabajo los diccionarios existentes del español.

En este artículo nos referimos a los problemas de la objetividad en la descripción del léxico mexicano, y al de la cantidad de datos necesaria para la labor lexicográfica. Particularmente nos ocupamos de la aplicación de la estadística lexicológica

[1] Este trabajo se publicó por primera vez en la *Nueva Revista de Filología Hispánica*, 23 (1974), pp 245-267

en la investigación del español de México como el mejor instrumento de documentación y análisis del vocabulario.

1. Los problemas básicos que se plantean al iniciar las labores de documentación para cualquier diccionario de interés lingüístico se pueden definir respecto a: *a)* los documentos lexicológicos de que se dispone antes de iniciar el trabajo; *b)* el valor científico de tales documentos; *c)* el uso que se les pueda dar en el trabajo que se emprende; *d)* la recolección de nuevos datos que complementen o sustituyan a los anteriores.

Una caracterización muy general de la lexicografía española nos permite distinguir dos tipos de obras lexicográficas al alcance de los hablantes: los diccionarios *generales* del español y los diccionarios de *regionalismos* del español. Los primeros dependen, en su totalidad, de los registros que presenta el *Diccionario de la Lengua Española (DRAE)* como reflejo de la labor de las Academias de la Lengua en el mundo hispánico[2]. En todos ellos hay algunos vocablos no registrados por el *DRAE*, pero fundamentalmente se trata de trabajos individuales en los que la

[2] Cf. María Moliner, *Diccionario de uso del español* (Madrid, 1970): "En un principio se pensó tomar las definiciones para este 'diccionario de uso' del *Diccionario de la Lengua Española*, diccionario oficial de la 'Real Academia Española', como lo han hecho hasta ahora absolutamente todos los diccionarios españoles" (pp. xiii-xiv). Si bien en el aspecto de las definiciones Moliner sostiene haber reconstruido totalmente el diccionario *(loc. cit.)*, más adelante dice: "Están incluidas en el presente diccionario todas las voces contenidas en el D.R.A.E., con . . . excepciones" (p. xxiv); en la sección de "Obras utilizadas" dice que, "dejando aparte las obras de consulta empleadas esporádicamente, cuya relación completa sería difícil de hacer y carecería de interés, se basa fundamentalmente en el *Diccionario de la Lengua Española* de la Real Academia Española . . . seguido paso a paso en la redacción de los artículos, si bien refundiendo y reorganizando las acepciones". El *Diccionario general ilustrado de la*

inclusión de voces depende de las necesidades lexicológicas que el autor supone que existen. En ninguno de ellos hay una descripción del léxico como la lingüística moderna lo exigiría.

Los diccionarios de regionalismos eliminan, por definición, "el vocablo de estructura y significado estrictamente castizo, es decir, incluido como tal en el Diccionario vulgar de la Academia Española"[3]. No hay en ellos, por lo tanto, una descripción integral del español usado en cada región del mundo hispánico; su utilidad para el usuario se define solamente por las singularidades léxicas de la región en cuestión y no por la presentación de la común lengua española.

Esta clara bipartición entre diccionarios generales y de regionalismos permite suponer que, hasta hoy, la lexicografía hispánica ha intentado establecer una coordinación entre ambos tipos de diccionario; pero, como se puede ver, el resultado obtenido deja una amplia laguna respecto a la realidad del léxico español en cualquier parte de sus dominios y no permite que el hablante tenga ante sí la imagen real de una lengua extendida por millones de kilómetros, revitalizada por las diferentes comunidades que la utilizan y, sin embargo, unida de una manera sorprendente si se la compara con otras lenguas de cultura en condiciones semejantes.

La documentación lexicológica de que disponemos es fragmentaria, guiada por oscuros principios de recolección y de un valor relativo para los usuarios, ya que los diccionarios genera-

lengua española (Vox), de S. Gili y Gaya no parece haberse separado de esa tradición, aunque en su prólogo no se indica explícitamente (véase p. xxx).

[3] Francisco J. Santamaría, *Diccionario de Mejicanismos*, 1a. ed. Porrúa, México, 1959. Introducción, p. xii.

les no dan cabida suficiente al español de distintas regiones, y los de regionalismos no dan lugar al vocablo general o a diversas particularidades del significado de los vocablos.

Desde un punto de vista exclusivamente lingüístico el valor de la documentación de que se dispone actualmente es aún más discutible, puesto que los registros no son homogéneos, mezclan diferentes estados de lengua y se dejan llevar por criterios como el purismo, reales para una comunidad, pero totalmente ajenos al estudio descriptivo.

En cuanto al uso que se puede dar a los documentos lexicológicos anteriores a nuestro trabajo, se hace necesario definir previamente las características de nuestro diccionario para, de acuerdo con ellas, saber exactamente hasta dónde nos son útiles ya que todo registro lexicológico tiene valor en sí mismo. Por esta razón queremos exponer, aunque sea brevemente, las pautas que han guiado nuestro modelo de diccionario.

1.1. El *DEM* se define como un diccionario sincrónico[4], descriptivo y —por limitaciones de tiempo y dinero— selectivo[5]. Nos interesa mostrar en él el léxico del español que se utiliza entre

[4] La "sincronía práctica" (como propone llamarla J. Rey-Debove) para el *DEM* ha quedado definida respecto a todo texto (hablado o escrito) que se haya producido entre 1970 y 1973; en el caso de obras literarias (o científicas) cuya publicación sea anterior a esa fecha y aparezcan como textos fundamentales para el corpus, las publicadas después de 1921.

[5] La selectividad del *DEM* se refiere exclusivamente a la cantidad de vocablos que hemos calculado podrá contener y no a la existencia de criterios prescriptivistas. Evidentemente, para reducir el tamaño de la nomenclatura habrá que seleccionar los vocablos más usuales. Para evitar toda confusión con otras acepciones del término "selectivo" vale la pena recordar que, según la tipología que ofrece L. Zgusta, el *DEM* se inscribe entre los "standard-descriptive dictionaries"; cf. Ladislav Zgusta, *Manual of lexicography*, Praga y La Haya, 1971, p. 210.

las fronteras geográficas de México[6] y, a diferencia de los diccionarios de regionalismos, lo entendemos como un *diccionario regional* de la común lengua española.

Esta breve descripción tipológica de nuestro trabajo nos permite evaluar los registros lexicológicos que existían antes del inicio de nuestra investigación. Para usar los diccionarios del español que se encuentran hasta hoy tendríamos que distinguir entre varias sincronías allí mezcladas, analizar severamente los criterios de inclusión de vocablos en los diccionarios y comprobar la existencia de cada uno de ellos en México con el objeto de mostrar una realidad. Resulta fácil ver que, por un lado, el trabajo de desglose y comprobación de los datos sería inmenso y, por otro, de ninguna manera conduciría a la descripción del español usual en México. Es necesario, por lo tanto, buscar otro método más adecuado a nuestras necesidades.

1.2. Son bien conocidas las aplicaciones que se han hecho de las cuantificaciones estadísticas a los estudios lingüísticos[7]; también se sabe hasta qué punto un análisis estadístico puede llegar a presentar situaciones de solución imposi-

[6] Las fronteras geográficas de la República Mexicana difícilmente tienen alguna realidad desde el punto de vista lingüístico. Al norte, el español se extiende por el sur de los Estados Unidos y al sur solamente podría existir alguna frontera dialectal en alguna parte de Centro América. Pero el hecho de que el *DEM* se preocupe por mostrar el español "nacional" de México nos obliga a respetar cuidadosamente nuestras fronteras políticas. Hay que señalar, sin embargo que, indudablemente, el *DEM* será útil para cualquier comunidad hispanohablante.

[7] Cf. Alphonse Juilland y Emilio Chang-Rodríguez, *Frequency dictionary of Spanish words*, La Haya, 1964, y U. Bortolini, C. Tagliavini y A. Zampolli, *Lessico di frequenza della lingua italiana contemporanea*, Milán, 1972 en donde aparecen reseñas críticas de varios trabajos dedicados a la estadística lingüística; entre otros citamos los de M. A. Buchanan, *A graded Spanish*

ble para la lexicografía[8]. No obstante, en vista de las dificultades aún mayores que representa el uso de procedimientos tradicionales de recolección lexicológica, hemos considerado que el método estadístico es el único capaz de darnos los registros necesarios y la cantidad de datos suficientes para nuestra tarea lexicográfica de un modo objetivo e imparcial. Describiremos ahora los puntos de vista que han quedado en la base de nuestro análisis, con el objeto de demostrar la utilidad del método seleccionado.

2. Del análisis estadístico de un corpus de datos deseamos obtener:

a) Un número elevado de *vocablos*[9] que puedan constituir la mayor parte de las entradas del diccionario.

b) Una base imparcial de selección de vocablos para la primera edición del *DEM.*

word book, Toronto, 1927; L. Rodríguez Bou, *Recuento de vocabulario español,* Puerto Rico, 1952; y V. García Hoz, *Vocabulario usual, vocabulario común y vocabulario fundamental,* Madrid, 1953.

[8] Análisis críticos del valor de la estadística en lexicografía se pueden encontrar, por ejemplo, en J. Rey-Debove, *Étude linguistique et sémiotique des dictionnaires françaises contemporains,* Mouton, La Haya, 1971, especialmente pp. 67-68; Charles Muller, "Un dictionnaire de fréquence de l'espagnol moderne", *ZRPh,* 81 (1965), 476-483 (en adelante, *Muller 65a); id.,* "Fréquence, dispersion et usage: à propòs d'un dictionnaire de fréquence", *CLex,* 7 (1965), 33-42 (en adelante, *Muller 65b).* Véase también K. H. Deutrich y G. Schoental, "Der Stellenwert der Statistik im Freiburger Analyse-Modell gesprochener Sprache", *Linguistique et statistique,* Colloque organisé par le Centre d'Analyse Syntaxique de l'Université de Metz, J. David y R. Martin (eds.), Klincksieck, Paris, 1974, pp. 95-104.

[9] Utilizamos la distinción propuesta por Ch. Muller ("Le mot, unité de texte et unité de lexique en statistique lexicologique", *TLL,* 1, 1963, 155-173) y K. Heger ("Die Semantik und die Dichotomie von langue und parole", *ZRPh,* 85, 1969, 144-215), según la cual la unidad de lengua que nos interesa es el *vocablo* al que corresponde la *ocurrencia* en el *habla* y el *tipo* en la Σ *hablas.*

c) Un punto de referencia que nos permita detectar los usos diferentes de los vocablos en la sociedad mexicana.

Estas tres necesidades nos colocan frente a frente, por una parte, con lo que significa un corpus para la lexicografía y la lingüística y, por la otra, con la concepción del corpus para la estadística.

2.1. Desde el punto de vista de la lingüística se hace necesario recordar, como lo ha señalado K. Heger[10], que todo corpus de datos produce tanto un número de documentaciones menor que el número de posibles ocurrencias que se pueden obtener del sistema, como un número mayor de ocurrencias que las que puede generar el sistema (como en el caso de las erratas, que son documentables y sin embargo no son realizaciones del sistema). Además, un corpus, por exhaustivo que sea, no será un documento completo del sistema y por ello el lingüista se verá obligado a trabajar continuamente extrapolando entidades del sistema de entre las realizaciones y, en consecuencia, un corpus de datos lingüísticos es una ayuda muy necesaria para el trabajo, pero no constituye una fuente exclusiva de materiales[11].

2.2. Para la lexicografía, en virtud de su carácter aplicado, la consideración del corpus puede tomar en cuenta las exigencias que impone la lingüística (y en nuestro caso debe hacerlo), pero también debe basarse en la riqueza del material y en el grado de objetividad de los datos

[10] Cf. Klaus Heger, "Belegbarkeit, Akzeptabilität und Häufigkeit", *Theorie und Empirie der Sprachforschung*, H. Pilch y H. Richter (eds.), Basilea, 1970, 22-33.

[11] K. Heger. *Monem, Wort und Satz*, Tübingen, 1971, § 1.2, p. 9.

respecto a una realidad léxica determinada. Estas dos condiciones son de una importancia extrema, puesto que son las que verdaderamente califican la utilidad del corpus; así, en cuanto a la riqueza de los materiales, J. Rey-Debove ha señalado que el tamaño reducido de todo corpus en comparación con el volumen real del léxico "entraîne des conséquences genântes: (1) l'absence de très nombreaux mots, (2) l'importance relative accrue des mots employés plusieurs fois par le même auteur (un idiolecte) par rapport à ceux employés une fois par plusieurs auteurs (plusieurs idiolectes), qui ont une valeur d'échange plus grande (d'où la nécessité de corriger la notion de fréquence par celle de répartition), (3) la faible fréquence des mots thématiques (liés à un thème conceptuel à l'exclusion d'un autre), même courantes, due au fait que tous les thèmes ne sont pas abordés dans le corpus. D'où la nécessité de corriger la notion de fréquence par celle de disponibilité"[12].

En cuanto al grado de objetividad del muestreo, conviene señalar la diferencia entre juzgar los resultados a partir de una consideración intuitiva de la "realidad" léxica (con lo que la evaluación se torna imposible al quedar sujeta a la experiencia de cada hablante), juzgarlos en comparación con trabajos realizados bajo muy diferentes enfoques (por ejemplo, con diccionarios anteriores elaborados de manera tradicional) y evaluarlos tras un análisis de sus diferencias con obras estadísticas previamente elaboradas. Este último criterio nos parece el único consecuente con el método seleccionado, pero aun aceptando otro tipo de evaluación subjetiva, la forma en que se realizó

[12] J. Rey-Debove, *op. cit.*, § 3.4.2.2, p. 68.

el muestreo para el *DEM* nos permite esperar un gran acercamiento a las "realidades léxicas" de los hablantes[13].

2.3. Para la estadística, el léxico de una lengua es el resultado de la unión de los léxicos individuales de los hablantes y constituye un conjunto finito. Pero, como Martinet observa[14], una característica esencial de todo léxico es su carácter "abierto", el constante aumento de vocablos dentro de una lengua, con lo que la identificación del conjunto resulta imposible; además, el léxico individual depende de una multitud de fenómenos que van desde la edad y el sexo del hablante, hasta las diferentes peculiaridades de su educación y de su actividad diaria, por lo que cada hablante conoce un léxico distinto. El resultado de esto es que, en estadística lexicológica, el conjunto del léxico no solamente no se puede identificar en su totalidad, sino que además puede variar de acuerdo con el tipo de hablantes cuyos léxicos particulares se han investigado. Dadas estas circunstancias, solamente se puede definir el léxico común del español mexicano como una intersección de léxicos individuales.

Al tomar en cuenta a los hablantes para seleccionar sus léxicos particulares, encontramos que, en nuestro caso, el número de mexicanos hablantes del español es elevadísimo y hace imposible conocer todos sus léxicos individuales; por ello nos vemos obligados a establecer, en primer término, una muestra de hablantes o, lo que es

[13] Cf. *Muller 65b*, p. 33: "En matière de lexique au contraire, la probabilité ne pourra jamais être confrontée qu'avec les fréquences constatées dans de nouveaux échantillons extérieurs au corpus, mais appartenant au même état de langue".

[14] Cf. André Martinet, *Eléments de linguistique générale*, 6a. ed., Paris, 1966, § 4.19.

lo mismo, una muestra de los textos producidos por los hablantes. En consecuencia, desde un punto de vista estadístico, el universo es única y exclusivamente el conjunto de los textos que forman la muestra, es decir, el conjunto de vocablos con sus frecuencias de uso que podemos encontrar en los textos seleccionados. Ahora bien, en vista de que los objetivos de la muestra no son hacer inferencias sobre este universo restringido, sino sobre uno muchísimas veces mayor y –para los hablantes– de mayor interés como lo es el "español de México", se hace necesario suponer que el conjunto de textos que hemos establecido *representa* al español de los mexicanos.

2.4. Esta última suposición se basa en presupuestos lingüísticos –como señala J. Rey-Debove[15]– puesto que no existe ningún medio estadístico de asegurar previamente las características de representatividad que debe tener la muestra en el momento de seleccionarla. Queda, por lo tanto, la interrogante de cómo seleccionar el corpus de tal manera que no solamente podamos confiar en su acercamiento a la "realidad" del léxico del español mexicano, sino que también resulte un corpus en el que las limitaciones a que se hizo referencia en el § 2.2 sean superadas al máximo.

3. Tras revisar las causas más evidentes de distorsión en una muestra lingüístico-estadística podemos confiar en que nuestro corpus podrá eliminarlas en la medida en que tome en cuenta:

a) Una cantidad de textos lo suficientemente grande como para obtener el número de vocablos que deseamos (aproximadamente 30 000) y

[15] J. Rey-Debove, *loc. cit.*

lo suficientemente reducida como para que sea económico y rápido su tratamiento con una computadora electrónica.

b) Una gran diversidad de textos que asegure la aparición del mayor número de vocablos "disponibles"[16], es decir, que permita la entrada de muy diferentes temas.

c) Una gran diversidad de autores, que elimine tanto como sea posible, los estilos individuales.

d) Una adecuada estratificación de los textos que permita obtener buenos resultados en el campo de la dispersión y el uso estadístico[17].

e) Una longitud suficiente de los textos, que permita la recuperación del significado global en que aparezcan los vocablos, para que la definición lexicográfica cuente con todos los elementos de juicio necesarios para el análisis semántico.

3.1 Respecto al tamaño y costo de la muestra nos hemos orientado por los objetivos centrales del *DEM* (presentación de una "lengua nacional" con todas sus variedades, cf. *supra* § 0) y no por

[16] El término *disponible* ha tenido su origen en los estudios de G. Gougenheim en torno al "francés fundamental"; se refiere a aquellos vocablos de baja frecuencia cuya utilidad es muy grande para cualquier usuario de un idioma; cf. Jean Dubois *et al.*, *Dictionnaire de linguistique*, Larousse, Paris, 1973, s. v. La disponibilidad de un vocablo puede verse directamente afectada por la mayor o menor variedad de textos que se exploren. Generalmente es más probable encontrar vocablos disponibles en una muestra muy diversificada que en una demasiado homogénea.

[17] Cf. *Muller, 65a*, p. 481: "C'est là surtout que la brièveté relative du corpus devrait être compensée par une stratification intensive à fin d'éliminer les mots propres à une discipline particulière, et de ne retenir que le vocabulaire commun du langage scientifique, et surtout à fin d'éviter des effets du sort comme celui que vient d'être cité" (a propósito del reducido número de "universos" empleados por Juilland y Chang Rodríguez). La frase de R. Moreau "estratifiquen a ultranza" se ha vuelto clásica en la formación de muestras estadísticas; véase su "Au sujet de l'utilisation de la notion de fréquence en linguistique", *CLex*, 3 (1962), 140-159.

cuestiones de precisión y confiabilidad sobre variables numéricas, como podría ser, por ejemplo, el cálculo de las frecuencias estimadas de uso de cada vocablo[18]. Esto significa que nos ha interesado menos el valor probabilístico de los vocablos en el corpus y que, en cambio, ha sido más importante calcular el tamaño y el costo de nuestro corpus respecto al número absoluto de vocablos que podamos encontrar en la muestra[19]. Para esto último las experiencias anteriores a la nuestra en cuanto al tamaño del corpus han sido muy aleccionadoras (ver el cuadro siguiente).

Como se puede deducir de los datos del cuadro anterior, no es necesario contar con una muestra muy grande para obtener el número de vocablos

[18] La muestra que constituye el corpus proporcionará una serie de estimaciones estadísticas sobre cada vocablo. Estas estimaciones son, por una parte, la frecuencia de uso del vocablo tanto dentro del total de la lengua, como dentro de cada género, y por otra las medidas específicas de la estadística lingüística. Si fijamos nuestra atención sobre alguna de las estimaciones –por ejemplo alguna de las frecuencias del uso del vocablo– estamos seguros de que, debido a las características del corpus como muestra y del léxico total como población teórica, no tendremos el valor exacto de esa frecuencia sino sólo una estimación que esperamos se acerque a ella de manera suficiente. Desde un punto de vista exclusivamente teórico estadístico, se podría diseñar un muestreo que nos garantizara, con una alta probabilidad (que nunca puede ser del 100%), que la estimación obtenida no se aleje en más de cierta cantidad (que nunca puede ser 0) del valor que tratamos de estimar. Sin embargo, para lograr una muestra tal para todos los vocablos identificados, o para la mayoría de ellos, con una precisión y un grado de confiabilidad aceptables, llegaríamos a tal complejidad en el diseño de la muestra y a un tamaño tan enorme, que automáticamente se invalida esta forma de analizar el problema.

[19] Cada una de las ocurrencias en el corpus tiene prácticamente el mismo costo de recolección y recuento, con la ligera excepción de aquellos textos hablados que deben ser transcritos previamente, pero que en realidad no representan costos adicionales notorios, por lo que la asignación de muestreo no lleva la complicación adicional de costos variables.

Obra	Extensión del Corpus	Vocablos Obtenidos
Frequency Dictionary of Spanish Words[20]	500 000	5 024
Computational Analysis of Present-day American English[21]	1 014 232	50 406
Trésor de la Langue Française (TLF)[22]	70 317 234	71 415

que nos proponemos incluir en el *DEM*. Consideramos que un corpus de 2 000 000 de ocurrencias nos proporciona un vocabulario lo suficientemente amplio para nuestras necesidades, y que darle mayor extensión elevaría el costo desproporcionadamente respecto al número de palabras que se podrían obtener, como demuestran los resultados que obtuvo la exploración del *TLF*.) A esto último hay que agregar que, dadas las proporciones que deseamos dar al *DEM*, los vocablos que se obtuvieran más allá de los dos millones de ocurrencias serían los menos usuales y que, más allá de la frontera de los 30 000 vocablos no solamente resultaría más sencillo sino también más aconsejable pasar a un procedimiento de documentación del tipo de diccionario-tesoro[23] en que todo registro es válido (como en

20 El objetivo del trabajo de A. Juilland y E. Chang-Rodríguez no era hacer un diccionario en el sentido estricto del término, sino un estudio estadístico con finalidades estructuralistas. Los 5 024 vocablos que registra son solamente los que obtuvieron una frecuencia lo suficientemente alta como para hacer proyecciones científicas.

21 Este corpus fue posteriormente parte básica de la nomenclatura del *American heritage dictionary of the English language*, William Morris (ed.), 1969.

22 La cifra corresponde exclusivamente al corpus literario del *TLF*. Se ha publicado un prolijo estudio estadístico en que se explican los resultados de la investigación. Cf. *Dictionnaire alphabétique de fréquences*, C.N.R.S., Nancy, 1973.

23 Esta idea, desde luego, corresponde a un pensamiento esencialmente práctico acerca de la factura de diccionarios. Cf. el

el caso de voces usadas por solamente un autor en cierto texto, de vocablos que dependen excesivamente de las modas lingüísticas, etc.). Por otra parte, si se busca una fuerte estratificación de los tipos de texto, como se recomienda en el § 3.*d* y en la nota 17, puede acrecentarse el rendimiento final de la muestra para los fines de la lexicografía.

3.2. La estratificación interna de la muestra se orienta también por los objetivos finales del *DEM* en cuanto implica una idea de lo que entendemos por léxico del español mexicano. Hay que señalar que, al definir nuestro diccionario como *regional,* pensamos que el léxico que presente pertenecerá por lo menos a tres conjuntos: al de los vocablos comunes a todo el mundo hispánico, al de las voces comunes a dos o más comunidades hispanohablantes (una de ellas México), y al de los vocablos usuales solamente en nuestro país. Esto significa que la limitación a las fronteras políticas de la República Mexicana es válida únicamente para *fijar las documentaciones de los vocablos* y no para dar la impresión de que el español mexicano es totalmente distinto del español general, ni para negar la realidad de que, desde el punto de vista de la dialectología, nuestras fronteras no marcan, seguramente, regiones dialectales distintas entre México, Guatemala y el sur de los Estados Unidos.

El punto de referencia en la formación interna del corpus necesitaba, en consecuencia, quedar determinado por un análisis de los tipos de texto que se producen en México. Para ello acudimos al concepto de *lengua culta,* de larga

prólogo de don Ramón Menéndez Pidal a la primera edición del diccionario Vox, citado en nota 2.

tradición lexicográfica, pero también objeto de estudio por parte de algunas corrientes de la lingüística contemporánea.

Entendemos por *lengua culta* el uso de un idioma en la comunicación intelectual de sus hablantes, uso lo suficientemente fijo como para permitir un amplio entendimiento entre los usuarios, pero también lo suficientemente flexible como para aceptar todas las innovaciones que impone la vida cultural de la comunidad[24]. La *lengua culta* es, en este sentido, el registro sociolingüístico de la lengua española en que *a)* predomina la función referencial sobre las otras funciones del lenguaje (según las definiciones clásicas de R. Jakobson), y *b)* se efectúan las comunicaciones lingüísticas de la mayor parte de los hispanohablantes educados.

La *lengua culta* entendida de esta manera, viene a ser más amplia que el español canonizado por las academias y, al mismo tiempo, consideramos que es la que constituye el punto de referencia en la apreciación de los hablantes a propósito de la "corrección idiomática"[25].

Con la *lengua culta* como punto de partida,

[24] Cf. Paul L. Garvin, "The standard language problem: concepts and methods", en D. H. Hymes (ed.), *Language in culture and society*, Nueva York, 1964, pp. 521-528: las propiedades de la lengua estándar (cf. *infra*, nota 26) son "flexible stability and intellectualization. Flexible stability here refers to the requirement that a standard language be stabilized by appropriate codification, and that the codification be flexible enough 'to allow for modification in line with culture change'. Intellectualization here refers to the requirement of increasing accuracy along an ascending scale of functional dialects from conversational to scientific" (p. 521).

[25] Garvin (*op. cit.*, p. 522) se refiere precisamente a la capacidad de la lengua culta de servir "as a frame of reference for correctness and for the perception and evaluation of poetic speech".

elaboramos un modelo de los usos sociales del español mexicano, es decir, un modelo diastrático, que albergara todos los posibles registros en que se realiza el español mexicano. Para ello se propuso una hipótesis a partir de nuestro conocimiento de la comunidad lingüística mexicana y de lo que en otros diccionarios del español y otras lenguas significan etiquetas como *popular, elevado, coloquial,* etc., connotativas de la función sintomática del signo lingüístico.

Una condición básica para elaborar el modelo fue siempre la de prever posibles diferencias en el uso del vocabulario y establecer las cotas suficientes que permitieran realizar, durante el estudio estadístico, agrupaciones que el modelo mismo no supone pero que gracias a él son fácilmente identificables. El modelo del uso del español mexicano, por lo tanto, constituye una hipótesis de valor únicamente operativo.

3.2.1. La *lengua culta* corresponde al registro más alto de los usos del idioma y forma el marco de referencia necesario para el sentido de la corrección lingüística del hablante. Se trata aquí del nivel a partir del cual los diccionarios establecen las calificaciones de uso del léxico y generalmente, en cuanto registro, no aparece marcado de ninguna manera.

Paul L. Garvin ha propuesto en varias ocasiones el concepto de *lengua estándar* como sinónimo de lo que la Escuela de Praga denominaba *lengua literaria o lengua escrita*[26] (que nosotros

[26] La escuela de Praga se refería indistintamente a la *langue littéraire* ("Thèses présentées au Premier Congrès des Philologues Slaves", p. 43) o a la *Schriftsprache* (B. Havránek, "Zum Problem der Norm in der heutigen Sprachwissenschaft und Sprachkultur"), en J. Vachek (ed.), *A Prague school reader in linguistics,* Indiana University Press, 1964, pp. 33-58 y 413-420 respectivamente.

hemos igualado con *lengua culta);* al aplicar estos conceptos a la práctica lexicográfica hemos creído necesario relacionarlos con los otros niveles de la lengua de manera tal que nuestras calificaciones de uso queden bien definidas con respecto a una visión global de los usos sociales del español mexicano.

En tales circunstancias preferimos distinguir entre *lengua estándar y lengua culta* haciendo más amplia a la primera y más restringida a la última. Y esto por una razón: creemos que en México hay un español uniforme en todo el país, resultado de la poderosa influencia no sólo de la educación, sino también de los medios masivos de información. Este *español estándar* se caracterizaría de la manera siguiente (véase cuadro A): es *general* en todas las regiones de México, es producto de lo que los antropólogos llaman "cultura urbana"[27] y se propaga continuamente a partir de los principales centros de irradiación del país (especialmente la ciudad de México). El nivel elevado del *español estándar* es la *lengua culta,* nivel de la literatura, de los textos científicos, de las conferencias, del periodismo, etc. Hay también un nivel del español mexicano estándar que se desvía de la lengua culta y es más familiar, más del dominio popular: lo llamamos *lengua sub-culta.* El español mexicano estándar cuenta, en consecuencia con dos niveles de uso por lo menos: el de la lengua culta y el de la lengua sub-culta.

Por contraposición con la lengua estándar, creemos que también existen usos del español poco extendidos, limitados a ciertas regiones

[27] Cf. Robert Redfield, *The folk culture of Yucatán,* Chicago, 1941, y Garvin, *op, cit.*

CUADRO A. NIVELES DE LA LENGUA EN LA MUESTRA DEL DEM; FUNCION PREDOMINANTE: REFERENCIAL

Lengua	*Nivel*		*Actualización*
ESTANDAR 1. general (geogr.) 2. urbana (sociol.) 3. irradiadora	culta	a. vocabulario intelectualizado y rico b. sintaxis rica c. modelo de corrección	escrita
	sub-culta	a. vocabulario no intelectualizado b. sintaxis limitada c. desviación del modelo de corrección	
NO ESTANDAR 1. limitada (geogr.) (sociol.) 2. rural (regional) urbana (grupos cerrados) 3. poco irradiadora	dialectal	a. vocabulario no intelectualizado, pero rico b. sintaxis regional c. modelos propios (?)	
	jergal	a. vocabulario limitado (terminologías) b. sintaxis pobre c. sujeta a modas	hablada

geográficas (dialectos del español mexicano) o a ciertos grupos sociales cerrados (jergas del hampa, de algunas profesiones, etc.); en el caso de los dialectos geográficos generalmente supondríamos su relación con "culturas rurales". Tanto los dialectos como las jergas resultan generalmente poco capaces de irradiar sus características a grandes zonas del país. Ambos forman lo que denominamos español mexicano *no-estándar.*

3.2.2. Para mostrar con más claridad cómo han surgido las oposiciones entre diferentes niveles correspondientes a la lengua estándar y a la no-estándar, nos pareció conveniente desglosar las características que Garvin y Havránek atribuyen a la lengua culta y, a base de una serie de rasgos opositivos, definir el resto de los niveles; así tendríamos que:

I. La lengua culta se caracteriza por un vocabulario muy vasto y, sobre todo, intelectualizado; por una explotación muy amplia de las posibilidades sintácticas del sistema y por su capacidad de servir como modelo de corrección para los hablantes.

2. La lengua sub-culta, en contraposición con la anterior, no presenta gran intelectualización del vocabulario, tiende hacia la repetición de un número menor de vocablos y de patrones sintácticos de la lengua y, algo muy importante, se concibe como "desvío del modelo de corrección"[28].

3. En la lengua no-estándar, los dialectos del español mexicano tienen vocabularios amplios y característicos de cada zona, pero menos intelectualizados; posiblemente muestren una explotación sintáctica singular en cada caso y, por lo

[28] Cf. Jean et Claude Dubois, *Introduction à la lexicographie: le dictionnaire,* Larousse, Paris, § 11.2 pp. 100-101.

menos teóricamente (en el caso de México), pueden formar marcos de referencia para el sentido de la corrección lingüística de sus hablantes[29].

4. Con las jergas se trata ante todo de vocabularios reducidos y algunos clichés sintácticos, sujetos totalmente a las modas y, por ello, capaces sólo fugazmente de formar modelos de prestigio.

Debemos señalar que todas estas oposiciones (a excepción de los dialectos, que se tratarán en seguida) no son resultado de una oposición previa entre tipos de hablantes (por ejemplo: hablantes considerados "cultos" frente a hablantes analfabetos), sino que solamente se refieren a niveles de uso de la lengua, a registros que todos los hablantes pueden utilizar en diferentes situaciones. Un profesor universitario, por ejemplo, puede utilizar todos los niveles, aunque posiblemente tenga más dominio del primero (lengua culta) que de los demás. Esto se vuelve evidente en el caso de las jergas puesto que conviven con la lengua culta: un médico, por ejemplo, alterna el vocabulario intelectualizado con vocablos o expresiones propias de su gremio.

Tratándose de los dialectos la situación es más complicada, puesto que generalmente el hablante de un dialecto domina también la lengua estándar, pero en cambio difícilmente es capaz de utilizar un dialecto diferente del suyo. De ser así, estos hablantes se tendrían que considerar como "bidialectales" o "multidialectales".

3.2.3. Al traducir el modelo en clases o "géne-

[29] Dado el carácter de los dialectos del español mexicano, que no son sino modalidades del castellano trasplantado a América, nos parece difícil que el sentido de corrección de los hablantes de un dialecto se manifieste tan nítidamente como en otras comunidades de la Península ibérica; podría ser así, sin embargo, en los casos de dialectos muy caracterizados como el veracruzano o el del noreste (Monterrey, N. L.).

ros" de textos para la muestra, tenemos la siguiente división[30]:

Lengua Estándar:

Lengua Culta. **Literatura**: novela, cuento, ensayo, teatro. **Periodismo**: reportajes de autor mexicano, editoriales, reseñas (políticas, sociales, culturales, deportivas, policiacas, taurinas).—**Ciencias**: humanas (bibliotecología, filosofía, historia, culturas indígenas, pedagogía y educación, psicología); *sociales* (antropología, arqueología, derecho, economía, geografía, politología, sociología); *físico-matemáticas* (astronomía, electricidad y electrónica, física, geofísica, matemáticas, computación); *químico-biológicas* (biología, química, farmacología); *administrativas* (administración, contabilidad, comercio); *medicina* (humana, veterinaria); *arte* (arquitectura, danza, artes plásticas, artes gráficas, arte dramático, música, cine).—**Técnicas**: *ingeniería* (civil, industrial, química, automotriz, aeronáutica, naval, de ferrocarriles, de minas); *comunicación* (correo y filatelia, periodismo, publicidad y mercadotecnia, radio y televisión); *oficios* (agricultura, ganadería, pesca, carpintería, electricidad, etc.); *labores domésticas* (costura, cocina, decoración, etc.); *deportes* (charrería, tauromaquia, futbol, etc.).—**Otros**: discursos políticos, religión, habla de la ciudad de México[31].

[30] La clasificación de las ciencias y las técnicas no tiene absolutamente ningún fundamento de orden epistemológico. Ha dependido, fundamentalmente, de nuestras necesidades prácticas de agrupar materias en grupos de tamaño relativamente homogéno. Véase en este mismo libro el artículo de R. Ham, "Del 1 al 100...", § 2 para la división definitiva del corpus.

[31] Estos textos pertenecen a las encuestas realizadas por el Centro de Lingüística Hispánica de la UNAM, publicadas bajo el

Lengua Sub-culta. **Literatura popular**: novela rosa, fotonovela, historietas.—**Otros**: conversaciones grabadas[32].

Lengua No-estándar: Textos regionales[33], documentos de estudios antropológicos, jergas, conversaciones grabadas.

3.2.4. Para que los autores de los textos fueran siempre distintos hemos vigilado que no se repitan más de dos veces en toda la muestra (en algunos casos excepcionales hay tres textos de un solo autor), y que en caso de repeticiones no sucedan dentro de un mismo "género".

3.2.5. Un problema especial se nos planteaba en el momento de asignar determinados porcentajes de importancia a cada género de la muestra. Inmediatamente nos dimos cuenta que no era posible tomar siempre el mismo número de textos de cada división, puesto que algunas estaban compuestas por textos cuyo léxico siempre es reducido (por ejemplo en la reseña deportiva o en algún texto científico). Optamos mejor por asignar diferentes "pesos" a cada género de acuerdo siempre con los objetivos finales del *DEM*. Nuestra ponderación resultó de la siguiente manera:

título de *El habla de la Ciudad de México. Materiales para su estudio,* UNAM, México, 1971. Corresponden al nivel que, dentro del trabajo dialectológico del Centro, denominan como "informantes cultos"

32 Se trata de encuestas realizadas por el Seminario de Dialectología de El Colegio de México sobre informantes de cultura "media". Las conversaciones que se citan en el nivel no-estándar corresponden, a su vez, a los informantes "analfabetas" de la Ciudad de México.

33 Aprovechamos los materiales que ha reunido el Seminario de Dialectología de El Colegio de México bajo la dirección del Prof. Lope Blanch para la delimitación de las zonas dialectales de México Cf. J.M. Lope Blanch, "Las zonas dialectales de México", *NRFH*, 19 (1970), 1-11.

Total de la muestra: 100%		*Porcentajes por géneros*	
Lengua Culta:	66.80%		100%
		Literatura	22.45
		Periodismo	26.34
		Ciencia	26.94
		Técnica	15.26
		Discurso político	2.69
		Religión	1.79
		Habla de la Ciudad de México	4.49
Lengua Sub-culta:	11.70%		100%
		Literatura popular	53.00
		Conversaciones grabadas	47.00
Lengua No-estándar	21.50%		100%
		Textos regionales	60.46
		Documentos de antropólogos	15.34
		Jergas	13.95
		Conversaciones grabadas	10.25

3.3. Los géneros de textos que han quedado establecidos a partir del modelo del § 3.2.2 se reparten entre el total de 2 000 000 de ocurrencias que fijamos previamente. Explicamos ahora cómo delimitamos cada texto. Como hemos dicho (ver § 2.2. y 2.4*c*), una de las causas principales de distorsión de la muestra estadística es la influencia del estilo individual de los autores; por lo que no es conveniente tomar para nuestro corpus un conjunto de libros completos (el hecho de que no intentemos hacer ni un diccionario literario ni un diccionario exhaustivo nos permite desechar esta tendencia generalizada por la tradición lexicográfica). La cuestión es entonces determinar el tamaño de los textos y la manera de seleccionarlos.

3.3.1. En cuanto al tamaño de los textos nos

ha parecido conveniente aplicar la fórmula utilizada en el estudio de H. Kučera y W. N. Francis sobre el inglés norteamericano contemporáneo[34], pues consideramos que, desde el punto de vista de las relaciones de frecuencia entre los vocablos, el español presentará características semejantes a las del inglés, ambas lenguas cultas contemporáneas. Un texto en nuestra muestra quedará formado, por lo tanto, por 2 000 ocurrencias extraídas de las obras que constituyen las fuentes para la muestra. Nuestro corpus queda como un conjunto de 1 000 textos con 2 000 palabras gráficas cada uno. Resumimos la lista presentada en § 3.2.3 indicando el número de textos en cada división:

Lengua Culta: 668; Literatura: 150; Periodismo: 176; Ciencias: 180; Técnicas: 102; Otros: 30.
Lengua Sub-culta: 117; Literatura popular: 62; Conversaciones grabadas: 55.
Lengua No-estándar: 215; Textos regionales: 130; Documentos de estudios antropológicos: 33; Jergas: 30: Conversaciones grabadas: 22.

3.3.2. En cuanto al modo de extraer los textos de las obras fuente, es necesario tomar en consideración tanto los resultados que se desea obtener como también la clase de datos que requiere el trabajo lexicográfico.

Como se ha dicho anteriormente no sólo nos interesa obtener del análisis de nuestro corpus la lista alfabética de los vocablos incluidos en él, sino que también deseamos conocer los contextos en que aparece utilizada cada palabra. Si el obje-

[34] Henry Kucera y W. Nelson Francis, *Computational Analysis of Present day American English*, Providence, R. I., 1970.

tivo fuera solamente el primero, lo más sencillo sería utilizar un muestreo aleatorio simple a todo lo largo de la obra fuente, es decir, se seleccionarían ocurrencias aisladas, y tendrían todas la misma probabilidad de aparecer en nuestro texto. Pero este procedimiento no conduce a la obtención de contextos y no permite obtener las selecciones tan rápidamente como lo desearíamos. Si, en cambio, tomamos las palabras con su contexto, la selección viene a ser continua y se aleja, en cierta medida, del ideal aleatorio. Dentro de un texto las palabras se reúnen en torno a una misma idea y de este modo se condicionan unas a otras. En muchos casos este mutuo condicionamiento es causa de repeticiones de vocablos, lo cual aumenta las frecuencias respecto del ideal estadístico. No obstante hemos preferido este último procedimiento tanto por las necesidades que se asientan antes, como por la facilidad que nos representa la selección de pasajes continuos.

Para definir la extensión de los pasajes que forman el texto convenimos en llamar *palabra* a la unidad tipográfica que aparece entre dos blancos y en llamar *párrafo* al conjunto de unidades tipográficas que aparecen entre dos puntos aparte. La identificación de estos elementos no ofrece ningún problema y, en cambio, nos permite obtener sentidos completos para las palabras que queramos analizar. Nuestra unidad de muestreo viene a ser el *párrafo* y un texto tendrá tantos párrafos como se necesite para alcanzar la extensión de 2 000 ocurrencias.

3.3.3. La selección de párrafos para cada texto se realiza con un esquema aleatorio: mediante una tabla de números al azar hecha especialmen-

te para nuestro trabajo, se efectúan las siguientes operaciones: *1)* se elige una página de libro o de revista (o un lugar dentro de una cinta magnética)[35]; *2)* se selecciona un párrafo de esa página (o cinta); *3)* se seleccionan tantos párrafos como sea necesario para lograr la cifra de 2 000 palabras gráficas; *4)* el proceso se repite cuantas veces sea necesario.

Hay casos en que no se encuentra un párrafo en la página escogida y se vuelve necesario tomar un nuevo número de página; hay otros en que el párrafo no comienza al principio de la página, por lo que el seleccionador toma el primero que cumpla con esa condición. También hay párrafos que terminan en la página siguiente y entonces el pasaje se tiene que extender hasta más allá de los límites de la página inicial.

4. Una vez realizadas las operaciones de selección de textos y de alimentación de datos a la computadora electrónica, iniciamos el análisis estadístico de los materiales aplicando las siguientes estimaciones y medidas estadísticas (véase cuadro B, en que se presentan en forma simbólica y con el formato numérico de salida de la computadora):

Gj caracteriza al j-ésimo género, en que j varía de 1 a m.

4.1. La clasificación de géneros se hará en dos versiones. En la primera se toma m = 3:

[35] En el caso de la transcripción de textos hablados se ha establecido previamente una tabla de convenciones de transcripción que no solamente permite aprovechar el material para el *DEM*, sino en general para cualquier trabajo de tipo dialectológico que no sea exclusivamente fonético.

G_1 = lengua culta
G_2 = lengua sub-culta
G_3 = lengua no-estándar

En la segunda clasificación tenemos m = 13:

G_1 = literatura
G_2 = periodismo
G_3 = ciencias
G_4 = técnicas
G_5 = discurso político
G_6 = religión
G_7 = habla Cd. México
G_8 = literatura popular
G_9 = conv. subculta
G_{10} = textos regionales
G_{11} = docs. antropológicos
G_{12} = jergas
G_{13} = conversacs. no-est.

4.2. Cada vocablo[36] identificado dentro del corpus aparecerá listado una sola vez junto con todas las frecuencias estimadas y las medidas estadísticas que le correspondan. Estas listas constituirán el núcleo de lo que podríamos llamar nuestro "diccionario estadístico del español de México".

Si se representa como V_i al i-ésimo vocablo de la muestra, se tendrá una lista de vocablos, esto es: $i = 1, 2, \ldots, \Omega$, donde Ω es el total de vocablos distintos en el corpus. La ordenación

[36] La obtención de *vocablos* según la definición que hemos seguido a lo largo de este trabajo (cf. *supra*, nota 9) supone que los problemas de lematización del corpus han sido resueltos en una etapa anterior.

de los V_i no será arbitraria, sino que se ajustará a los siguientes requisitos: *a)* una ordenación que tome en cuenta la frecuencia total de ocurrencias del vocablo, en sentido decreciente. Las palabras con frecuencia idéntica se agruparán alfabéticamente; *b)* una ordenación estrictamente alfabética.

Estas dos clasificaciones y las dos mencionadas por géneros dan lugar a cuatro tabulaciones básicas que producirá la computadora. La ordenación por frecuencias nos permitirá identificar el vocabulario más usual —estadísticamente hablando. La ordenación alfabética nos permitirá conocer las características estadísticas de cualquier vocablo (un vocablo que no aparezca dentro del corpus, es decir, $t = 0$ tendrá precisamente esa característica estadística)[37].

4.3. Las frecuencias absolutas cuentan el número de ocurrencias de cada vocablo dentro del corpus. De esta manera definimos como t_i al número de ocurrencias del vocablo V_i dentro del total; para contar la frecuencia de los vocablos en cada género tomamos como f_{ij} al número de ocurrencias del vocablo V_i dentro del j-ésimo género.

Las frecuencias relativas miden el porcentaje de ocurrencias de cada vocablo respecto al total dentro de cada género y entre los géneros, de acuerdo con las siguientes relaciones:

$$h_i = \frac{100\, t_i}{\sum_{i=1}^{\Omega} t_i} = \text{porcentaje de ocurrencia de } V_i$$

37 Una discusión a propósito de la frecuencia cero en estadística lingüística se puede encontrar en K. Heger, *Belegbarkeit. . .*, pp. 26-27.

$$c_{ij} = \frac{100\ f_{ij}}{t_i} =$$ porcentaje de ocurrencia entre géneros del vocablo i en el género j.

$$d_{ij} = \frac{100\ f_{ij}}{\sum_{i=1}^{\Omega} f_{ij}} =$$ porcentaje de ocurrencia dentro de géneros del vocablo i en el género j.

$$r_j = \frac{100 \sum_{i=1}^{\Omega} f_{ij}}{\sum_{i=1}^{\Omega} t_i} =$$ tamaño relativo del género j

4.4. La mera frecuencia total, ya sea absoluta o relativa, es en realidad una medida lingüística poco útil. Dos vocablos con la misma frecuencia, pero con distinta distribución entre géneros, tienen distinto comportamiento dentro del idioma: una distribución irregular entre los géneros denota que el vocablo está sujeto a determinantes de estilo o de tema, mientras que una distribución regular —en todos los géneros— indica que la palabra es independiente de la clasificación por géneros y que por lo tanto es mayor su importancia y su utilidad dentro del idioma. Se busca entonces una medida que combine tanto la frecuencia del vocablo como su *dispersión* entre géneros, de modo que entre vocablos de parecida frecuencia total se favorezcan aquellos con dispersión más uniforme.

Luego de analizar distintas medidas que se han propuesto y usado en varios diccionarios de frecuencias, para reflejar la situación planteada, adoptamos como la más adecuada la creada por

CUADRO B

Núm de orden	*Palabra*	*Frecuencias absolutas*					*Frecuencias relativas*									*Medidas estadísticas*		
							Respecto al total	*Entre géneros*				*Dentro de géneros*						
		Total	G_1	G_2	...	G_m		G_1	G_2	...	G_m	G_1	G_2	...	G_m	*KF*	*S*	*C*
1	P_1	t_1	f_{11}	f_{12}	...	f_{1m}	h_1	e_{11}	e_{12}	...	e_{1m}	d_{11}	d_{12}	...	d_{1m}	KF_1	S_1	C_1
2	P_2	t_2	f_{21}	f_{22}	...	f_{2m}	h_2	e_{21}	e_{22}	...	e_{2m}	d_{21}	d_{22}	...	d_{2m}	KF_2	S_2	C_2
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
i	P_i	t_i	f_{i1}	f_{i2}	...	f_{im}	h_i	e_{i1}	e_{i2}	...	e_{im}	d_{i1}	d_{i2}	...	d_{im}	KF_i	S_i	C_i
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Ω	$P\Omega$	t_Ω	$f_{\Omega 1}$	$f_{\Omega 2}$	...	$f_{\Omega m}$	h_Ω	$e_{\Omega 1}$	$e_{\Omega 2}$	...	$e_{\Omega m}$	$d_{\Omega 1}$	$d_{\Omega 2}$	...	$d_{\Omega m}$	KF_Ω	S_Ω	C_Ω

J. Lanke[38], quien, tras sopesar distintas cualidades y defectos de las medidas que hasta ahora se han utilizado, llega a preferir el índice de frecuencia y dispersión entre géneros llamado KF (por *Korrigierte Frequenz),* que guarda la siguiente definición para el *i*-ésimo vocablo (de acuerdo con la notación adoptada en nuestro trabajo):

$$KF_i = \frac{1}{100} \left(\sum_{i=1}^{m} \sqrt{r_j \, f_{ij}} \right)^2$$

Para propósitos de exploración lingüística, también calcularemos un índice que, independientemente de la frecuencia, nos dé indicaciones de cómo se dispersa un vocablo entre géneros. El mismo J. Lanke propone la medida:

$$S_i = \frac{KF_i}{t_i}$$

Pero este índice presenta la dificultad de que, si se aplica a una muestra en que los géneros que la componen son desiguales —como en el caso presente—, su variación depende directamente de las desigualdades de la muestra y va desde la menor frecuencia relativa observada en el caso de la distribución más desigual, hasta 1 cuando la distribución es uniforme. Para corregir esta dificultad hemos optado por la siguiente modificación, a la que llamamos índice normalizado de dispersión:

$$C_i = \frac{100 \, S_i - \min_j r_j}{100 - \min_j r_j}$$

[38] Según reporta I. Rosengren, "The quantitative concept of language and its relation to the structure of frequency dictionaries", *ELA,* 1 (1971), 103-127.

El índice normalizado de dispersión entre géneros queda ahora establecido en un rango que varía entre 0 y 1 donde 0 indica la distribución más desigual y 1 la más uniforme.

5. Esperamos haber mostrado a lo largo de este trabajo que el estudio estadístico del léxico del español mexicano representa la mejor posibilidad de obtener una imagen real de la lengua que se utiliza en México y que, a pesar de las limitaciones que nos hemos ido imponiendo en el transcurso de nuestro proyecto, es el procedimiento más útil y adecuado para el problema que enfrentamos. Creemos haber tomado todas las precauciones necesarias para que el rendimiento del corpus sea suficiente y verdaderamente constituya el fundamento para la selección de vocablos que compongan el *DEM.* Sobre esta selección final (posterior al estudio en sí) deseamos agregar que el tema esencial que se propone a todo el equipo lexicográfico es el de saber cuáles vocablos deben integrar el diccionario y cuáles pueden desecharse. Dado que nuestro diccionario no es autoritario, el criterio de "uso de los buenos escritores", o los más conocidos del purismo y del casticismo no pueden constituir el punto de partida para la formación de la macroestructura del *DEM.* Como hemos querido demostrar en los párrafos iniciales de este trabajo, los léxicos de mexicanismos que se pueden encontrar hoy en día no pueden tampoco servir como base para nuestra selección de voces. Mediante el análisis estadístico en frecuencias, dispersión y los dos índices corregidos (KF y C) tendremos una base necesaria para toda consideración del vocabulario por incluir en el *DEM.* No obstante, por las razones que hemos apuntado, será necesario tener un criterio complemen-

tario de inclusión para vocablos de baja frecuencia, cuyo análisis estadístico no produzca resultados definitivos, y para vocablos que no hayan aparecido en la muestra y sin embargo sean de importancia para las finalidades del *DEM*.

En estos dos casos, la decisión dependerá del juicio intersubjetivo del cuerpo de redacción del *DEM* formado por el equipo lexicográfico, el consejo de redacción y el consejo consultivo. Con objeto de no falsear los resultados obtenidos en el estudio, las voces incluidas por esta otra vía recibirán una marca especial, aunque no hemos decidido si esto se hará en el cuerpo mismo del *DEM* o por separado.

Del 1 al 100 en Lexicografía

Roberto Ham Chande

1. Introducción

Cuando utilizamos como escala el abecedario y decimos que una tarea la llevamos a cabo desde la A hasta la Z, lo que estamos diciendo es que la realizamos paso a paso y que la completamos totalmente. Cuando en lugar de esa referencia alfabética utilizamos la escala de los números naturales (enteros y positivos) y comenzando en 1 llegamos a 100, en realidad quiere decir que apenas comenzamos la tarea y queda demasiado por cumplir. El título de este trabajo[1] justamente quiere decir eso, además de que el uso de cifras también intenta señalar que se trata de la parte numérica del trabajo, particularmente de los primeros resultados estadísticos, sobre el Corpus del Español Mexicano Contemporáneo (CEMC).

Una de las herramientas de primera importancia, y en cierta medida innovadora, que se utilizan en la construcción del *Diccionario del Español de México* (DEM), es sin duda alguna la metodología estadística elaborada ex profeso para el proyecto. La estadística constituye una base inicial y fundamental en la tarea lexicográfica. La idea básica que motiva y justifica tal modalidad técnica es la de que, pretendiendo ser el DEM el reflejo y la validación del vocabulario español utilizado por los hablantes mexicanos promedio, es necesario que se cumplan dos con-

[1] El título de este capítulo fue sugerencia de L. F. Lara, coordinador general del proyecto DEM, quien también hizo la revisión del primer borrador.

diciones: en primer lugar que sean precisamente los usuarios del idioma quienes identifiquen cuáles son los vocablos que utilizan, y en segundo lugar que también sean los usuarios los que informen dentro de qué contextos se utilizan estos vocablos.

La obtención de esos vocablos y sus contextos es imposible de efectuarse de modo exhaustivo, recogiendo todos los textos de todos los usuarios; aunque existe la alternativa, totalmente viable y capaz de solucionar satisfactoriamente el problema planteado, de hacerlo por muestreo. Una muestra es el estudio de una parte de un todo con el afán de hacer inferencias científicamente válidas sobre ese todo. La idea del muestreo no es una concepción extraña sino hasta familiar de la vida diaria, y todo el tiempo nos afectan decisiones hechas por muestreo que van desde cosas triviales como decidirnos a comprar fruta después de una "probadita" que nos obsequian en el mercado, hasta que un gobernante nos imponga un nuevo sistema educativo porque las "pruebas" indican que es mejor.

Este principio es el que origina de modo muy natural la concepción de un muestreo de usuarios de la lengua y de los textos que producen. Sobre este concepto de una muestra del español contemporáneo de México podemos decir, haciendo una gran esquematización, que estadísticamente la población a sondear es el universo de los textos, tanto hablados como escritos que producen los usuarios modernos de la lengua, y que la muestra es simplemente una parte de tales textos. A ese conjunto parcial de textos en la forma de muestra es a lo que denominamos CEMC. La organización de este corpus y la selección que lo materializó fueron hechas conforme a los li-

neamientos de muestreo estadístico especialmente diseñados para este caso. Estos lineamientos han obedecido a la necesidad de llenar tres requerimientos indispensables para la correcta consecución de la meta final del trabajo y que en última instancia es precisamente la producción del DEM. Los requerimientos que marcaron las pautas a seguir por el muestreo fueron: *a)* la practicabilidad de la selección dentro de la capacidad limitada de los recursos materiales, humanos y de tiempo asignados al proyecto, *b)* la necesidad de garantizar que el corpus "representara" con cierta calidad al universo del cual es muestra y *c)* la obtención de contextos capaces de identificar y demostrar cuál es el uso y significado de las ocurrencias dentro del corpus. Detalles más completos acerca de la necesidad y justificación de un corpus en el trabajo de lexicografía, las características que debe llenar, su tamaño, estratificación y la forma de seleccionarlo para cumplir con los principios de aleatoriedad y practicabilidad indispensables, se encuentran pormenorizados en el artículo de Luis Fernando Lara y Roberto Ham Chande, "Base estadística del Diccionario del Español de México" que se reproduce en este mismo volumen[2].

Se han terminado varias fases del plan de trabajo estadístico. Inicialmente, se efectuó la selección de la muestra para su registro magnético, lo que creó el CEMC propiamente dicho. Posteriormente, se procesó numéricamente el corpus para calcular las frecuencias y medidas estadísticas básicas[3].

2 Inicialmente publicado en *Nueva Revista de Filología Hispánica*, 23 (1974), pp. 245-267

3 Este procesamiento numérico se debió al gran trabajo rea-

Por último, hemos obtenido resultados estadísticos como los que hay en este trabajo, que constituyen una infraestructura fundamental para el desempeño del trabajo lexicográfico, pues sirven para ponderar la importancia relativa de los vocablos del idioma y ayudan a caracterizar cuantitativamente la lengua española utilizada en México. El presente trabajo tiene como meta consignar una primera descripción numérico-estadística de la estructura del CEMC.

2. Estratificación y tamaño del Corpus

Para llevar a cabo convenientemente la tarea propuesta de obtener la información estadística necesaria y adecuada para los propósitos del DEM, es requisito que se consideren los diversos niveles y géneros de los que se compone la lengua que denominamos español de México. Volviendo al lenguaje estadístico, el universo que representa el idioma está dividido en estratos y subestratos y si la muestra, o sea el corpus, debe ser "representativa" de este universo, entonces debemos cerciorarnos de que los estratos y subestratos estén tomados en cuenta y suficientemente "representados" en el corpus. Otra vez utilizando la terminología estadística, esto quiere decir que se trata de una muestra estratificada. Esta estratificación se refiere a la clasificación de la lengua en tres niveles y dentro de cada uno de ellos a una subclasificación en géneros de textos propios de cada nivel. Esta última clasificación comprende catorce categorías. En la primera parte del cuadro No. 1 se señalan cuáles

lizado por Isabel García Hidalgo, quien con el auxilio de Jorge Serrano y Javier Becerra, realiza la tarea de computación.

Cuadro 1

Nivel	Género	No. de textos		No. de Ocurrencias		% de Ocurrencias	
Lengua culta	Literatura	150		269 788		14.2666	
	Periodismo	176		299 775		15.8523	
	Ciencias	180		246 313		18.3133	
	Técnicas	102		202 716		10.7198	
	Discursos políticos	18		31 971		1.6907	
	Religión	12		21 277		1.1251	
	Habla culta	30	*688*	69 473	*1 241 313*	3.6758	*65.6416*
Lengua sub-culta	Literatura popular	63		127 459		6.7401	
	Habla media	30		59 567		3.1500	
	Lírica popular	24	*117*	45 149	*232 175*	2.3875	*12.2776*
Lengua no estándar	Textos dialectales	130		259 881		13.7427	
	Documentos antropológicos	33		68 376		3.6158	
	Jergas	20		34 839		1.8423	
	Habla popular	28	211	54 461	417 557	2.8799	22.0808
Totales			*996*		*1 891 045*		*100.0000*

son estos niveles y cuáles son los géneros que se tomaron en cuenta como subclasificaciones[4].

Al considerar la situación de tomar una muestra estratificada de una lengua en circunstancias como las que se dan en el caso particular del español de México, nos topamos con una peculiaridad que establece una diferencia sustancial entre este muestreo y la generalidad de los muestreos estratificados de otra índole. En esos otros muestreos aplicados a poblaciones clasificadas en estratos se conoce o se estima el tamaño de cada una de esas clases[5] componentes de la población, o su tamaño relativo dentro del total. En nuestro caso, siendo la lengua algo que se produce todos los días, por todos sus usuarios, en forma tan masiva y dinámica, y en una mezcla tan decididamente imponderable, se hace absolutamente imposible determinar el tamaño absoluto o relativo de los distintos niveles y géneros.

Podemos recordar que no hay una sola persona o escrito que recoja el acervo total de vocabulario que una lengua moderna y dinámica posee, e incluso podemos agregar que la cobertura total no se lograría ni siquiera contando con todos los diccionarios distintos en existencia. El simple número de vocablos disponibles es desconocido y cambiante todos los días. A esta dificultad ya de por sí insuperable se agrega una complicación adicional, que surge de la ponderación de la importancia de un texto producido y los vocablos que éste contiene. Si tomamos el caso de la conversación cotidiana, la producción de tex-

[4] Véase "Base estadística del Diccionario del Español de México", § 3.2.3, para la estratificación de los textos inicialmente planeada.

[5] En la terminología estadística estas clases se denominan estratos y deben ser sin traslapes y exhaustivas de la población.

tos hablados es inmensa; sin embargo, para cada hablante hay en general un número bastante limitado de oyentes. En otras circunstancias, como pueden ser los textos de los programas de la televisión o de los reportajes periodísticos, su producción en sí es menor, pero su difusión e impacto entre los receptores de estos mensajes es sumamente grande. Entrando en otras consideraciones, también tenemos la apreciación anímica de la importancia dentro del idioma de una obra literaria o científica de cierto nivel cultural, en la cual se tiene la desfortuna de un tiraje limitado, en comparación con, por ejemplo, fotonovelas sin valor literario alguno, en las que también hay una desfortuna consistente en su gran tiraje y comercialización. De esta manera, todas estas situaciones combinadas no dejan lugar más que a una sola alternativa posible, representada por la ponderación subjetiva. Bajo estos puntos de vista, que juegan un papel importante en lexicografía, se asigna una importancia relativa a cada nivel de la lengua y a cada género respecto al idioma como universo total de discurso. Se adopta la responsabilidad de asignar un criterio totalmente cuantitativo a partir de apreciaciones de tipo completamente cualitativo.

En el cuadro No. 1 se muestra el resumen y resultados de la clasificación, ponderación y selección del corpus. Ahí se señalan la gran estratificación en tres niveles de la lengua, la subestratificación en catorce géneros, el número de textos recopilados en cada clasificación, el número de ocurrencias[6] y el porcentaje de estas respecto a la clasificación. Las diferencias en estos porcen-

[6] En las ocurrencias no se tomaron en cuenta nombres propios ni fechas.

tajes respecto a lo planeado, de acuerdo con esa importancia subjetiva asignada, se debe al ajuste propio del trabajo de selección y recopilación, y no son realmente significativas.

3. Principales índices estadísticos

Una de las tareas primordiales encomendadas al CEMC y a su estadística, es la de la identificación objetiva de los vocablos de mayor importancia dentro de esta lengua que conocemos como español de México.

Esos vocablos constituirán el vocabulario básico, cuerpo principal del DEM. Al esbozarse este objetivo, surge de inmediato la pregunta de cómo definir la importancia de un vocablo de modo de poder medirla concretamente y estar en posibilidad de establecer rangos y dimensiones entre vocablos.

Ante esta cuestión, la respuesta que con mayor sencillez surge es la de utilizar la frecuencia total observada de un vocablo como medida precisamente de su importancia.

Esta frecuencia reflejaría la intensidad con la cual se utiliza el vocablo en estudio. Mientras más alta fuera la frecuencia de aparición total en el corpus, mayor sería la importancia del vocablo y más merecería ser tomado en consideración.

Esta solución clara y sencilla, en realidad tiene fallas que impiden su aplicación simple. El origen de las deficiencias se encuentra en el hecho de que el CEMC intenta ser un reflejo del español de México a través de una mezcla subjetiva de los distintos niveles de la lengua y de sus géneros, en porcentajes de participación diferentes. La identificación de los géneros y de su representación porcentual sigue las pautas de los propó-

sitos del DEM y va de acuerdo con la ponderación atribuida a cada clasificación de los textos de la lengua, que varía de una clase a la otra. Esto da lugar a dos hechos que son los causantes principales de la inadecuación de la frecuencia total como ordenadora de los vocablos, y los cuales se comentan a continuación.

En primer lugar, aunque estrictamente nada impide que cualquier vocablo se utilice dentro de cualquier género, sí es inmediatamente reconocible que la frecuencia de uso de los vocablos se ve afectada directamente por la clase de género que se maneja. En el caso extremo nos encontramos con las palabras especializadas, que pudieran aparecer en cualquiera de los otros géneros, pero que en su frecuencia consignada será definitivamente mayor dentro del género especializado de tal vocablo. Así, esa frecuencia de aparición en ese género particular también aparecerá dentro de la frecuencia total. Si se toma entonces la frecuencia total como ordenador se le estará atribuyendo a una palabra la misma importancia que a otra que tiene la misma frecuencia total aunque su aparición no fuera exclusiva de un género y se repartiera entre todos[7]. Es intuitivo que una palabra "mejor repartida" es de mayor importancia que otra concentrada en un sólo género aunque su frecuencia total coincida para ambos casos.

En segundo lugar tenemos la influencia del tamaño del CEMC dentro de cada género. Un género con un tamaño relativo mayor, también da mayor oportunidad a la aparición de sus vocablos propios y especializados, lo que de nueva

[7] Nótese por ejemplo el número de veces que las palabras "vocablo", "género" o "frecuencia" aparecen en este artículo. Cf. "Base estadística. . .", § 3.

cuenta hace que se afecte la frecuencia total. Por el simple hecho de que un género esté mayormente representado dentro del corpus que otro, por eso mismo la frecuencia de aparición de ciertos vocablos se inclina hacia ellos.

Todas las anteriores consideraciones nos han llevado a concluir que la frecuencia total F debe ser sustituida por una "frecuencia corregida", que denotamos por KF, como la principal herramienta para medir la importancia y el orden de los vocablos. Este parámetro KF fue obtenido por J. Lanke, de la Universidad de Lund y lo justifica plenamente I. Rosengren[8] tanto en su interpretación como en su deducción. Aquí solamente nos corresponde consignar que KF toma en cuenta esos factores que intervienen en la ponderación de un vocablo que aparece dentro de un corpus, y que son la frecuencia de aparición, su repartición entre los distintos géneros y el tamaño relativo de cada género.

Junto con la frecuencia total y la frecuencia corregida, también es importante tener un indicador que señale cuál es la dispersión del vocablo entre los géneros. Independientemente de las frecuencias, este indicador nos dirá cuándo un vocablo experimenta concentraciones en uno o varios géneros, cuándo es en realidad propio del vocabulario general al encontrarse bien distribuido entre todos los géneros, o cuándo hay una situación intermedia y en qué grado. Para tal objeto contamos con un índice que denominamos C, que toma en cuenta tanto la frecuencia de aparición en cada género como los tamaños rela-

[8] Rosegren, I. "The quantitative concept of language and its relations to the structure of frequency dictionaries". *Etudes de linguistique appliquée,* 1 (1975), pp. 103-127.

tivos de éstos, y que normaliza el rango de este número de 0 a 1[9].

La interpretación de este índice es la de que cuando C se acerca a 1 entonces denota una distribución más regular del vocablo entre los géneros, siendo 1 la total regularidad, y cuando C se aproxima a 0 es cuando se tiene una mayor irregularidad en la distribución entre géneros de las ocurrencias del vocablo. C igual a 0 indicaría todas las ocurrencias dentro del género más pequeño.

Tanto las expresiones algebraicas como mayores detalles del porqué de los índices KF y C se consignan en el artículo "Base estadística del Diccionario del Español de México", en este mismo volumen. Una ejemplificación del significado y uso de los índices estadísticos discutidos se lleva a cabo en la siguiente sección.

4. Las cien primeras palabras

Es cierto que cada pedazo de información contenida dentro del CEMC, por pequeño que sea, es ilustrativo de algún aspecto del español de México. Sin embargo, como en toda otra clase de información de tipo masivo, existen prioridades, ponderaciones y resúmenes que nos permiten sintetizar los datos para describir un fenómeno dentro de propósitos y capacidades delimitadas.

Esta última noción es el objetivo de la estadística descriptiva. En el caso del CEMC una de las principales y más inmediatas expresiones de información sintetizada es la identificación y pon-

[9] En el trabajo citado en la nota 7 se utiliza un índice S, el cual modificamos para convertirlo en el índice C que nosotros usamos, principalmente para normalizarlo entre 0 y 1, ventaja que no se observa en S. Cf. "Base estadística. . ", § 4.4.

Núm.	VOCABLO	CAT. GRA.	KF	F	C
1	LA	ART	85919.13	87827	0.9780
2	EL	ART	78942.85	80845	0.9806
3	DE	PREP	63088.75	63541	0.9941
4	Y	CONJ	59255.11	59476	0.9962
5	QUE	PRON	58364.00	58595	0.9960
6	EN	PREP	48228.00	49299	0.9780
7	A	PREP	44569.77	44809	0.9946
8	SE	PRON	33442.49	33703	0.9922
9	NO	ADV	28670.65	31468	0.9101
10	SER	V	24428.99	24918	0.9802
11	UN	ART	20561.20	20646	0.9958
12	POR	PREP	19694.88	19862	0.9915
13	CON	PREP	18660.50	18775	0.9938
14	SU	ADJ	17233.52	17823	0.9665
15	UNA	ART	15597.41	15692	0.9939
16	HABER	V	14373.69	14619	0.9830
17	PARA	PREP	13951.97	14164	0.9849
18	AL	CONTR	11028.71	11207	0.9839
19	ESTAR	V	10940.95	11673	0.9366
20	COMO	ADV	10753.78	10818	0.9940
21	TENER	V	9402.55	9818	0.9572
22	LE	PRON	9192.13	10539	0.8707
23	HACER	V	8607.80	8826	0.9750
24	YA	ADV	8238.52	9815	0.8376
25	O	CONJ	7989.34	8368	0.9542
26	PERO	CONJ	7706.24	8343	0.9228
27	DECIR	V	7612.18	8881	0.8555
28	QUE	CONJ	7565.24	7978	0.9477
29	LO	ART	7335.91	7378	0.9942
30	ME	PRON	6928.90	10427	0.6607
31	MAS	ADV	6877.20	6964	0.9874
32	PODER	V	6200.40	6350	0.9762
33	ESTE	ADJ	6180.97	7142	0.8639
34	IR	V	5621.48	7047	0.7954
35	LO	PRON	5549.14	6118	0.9060
36	SÍ	ADV	4976.58	8656	0.5701
37	VER	V	4891.83	5566	0.8775
38	DAR	V	4876.03	5045	0.9661
39	CUANDO	ADV	4751.19	4907	0.9679
40	MUY	ADV	4503.17	4913	0.9156
41	YO	PRON	4390.14	7021	0.6210
42	PORQUE	CONJ	4202.55	5079	0.8255
43	EL	PRON	4101.76	4510	0.9085
44	MI	ADJ	4046.84	5744	0.7012
45	PUES	CONJ	4005.46	5918	0.6731
46	LA	PRON	3927.33	4259	0.9212
47	ASI	ADV	3739.74	4157	0.8985
48	ESTA	ADJ	3530.77	3735	0.9447
49	TODO	ADJ	3428.97	3455	0.9924
50	TAMBIEN	ADV	3388.18	3498	0.9682

DRO 2

Núm.	VOCABLO	CAT. GRA.	KF	F	C
51	VEZ	S	3152.41	3239	0.9730
52	NOS	PRON	3146.58	3389	0.9277
53	A:ÑO	S	3102.49	3294	0.9412
54	SABER	V	3085.00	3441	0.8954
55	SIN	PREP	2999.75	3188	0.9403
56	HASTA	PREP	2979.90	3034	0.9820
57	QUERER	V	2916.40	3546	0.8204
58	DEBER	V	2893.21	3199	0.9033
59	TODO	PRON	2892.46	3137	0.9212
60	AQUI	ADV	2818.17	4033	0.6953
61	UNO	PRON	2655.45	3097	0.8558
62	DIA	S	2623.75	2785	0.9414
63	ESO	PRON	2526.39	3352	0.7509
64	QU:É	PRON	2522.64	3176	0.7919
65	ELLA	PRON	2410.23	2612	0.9219
66	SOBRE	PREP	2410.01	2692	0.8941
67	BIEN	ADV	2361.08	2499	0.9442
68	LLEGAR	V	2348.04	2442	0.9611
69	MAS	ADJ	2341.78	2465	0.9494
70	DONDE	ADV	2274.15	2320	0.9800
71	ENTRE	PREP	2268.93	2457	0.9226
72	NI	CONJ	2249.99	2391	0.9404
73	OTRA	ADJ	2242.32	2265	0.9899
74	ENTONCES	ADV	2229.34	2960	0.7503
75	ESA	ADJ	2227.29	2333	0.9542
76	LLEVAR	V	2185.06	2297	0.9507
77	PONER	V	2129.97	2319	0.9176
78	PARTE	S	2116.71	2241	0.9439
79	TE	PRON	2048.26	3392	0.5993
80	TIEMPO	S	2047.34	2068	0.9899
81	DOS	S	1995.51	2019	0.9882
82	DESPUES	ADV	1988.68	2035	0.9770
83	DEJAR	V	1982.57	2098	0.9444
84	DESDE	PREP	1886.32	1910	0.9875
85	HOMBRE	S	1877.77	1989	0.9434
86	ESE	ADJ	1869.75	1980	0.9437
87	CADA	ADJ	1826.01	1884	0.9689
88	VENIR	V	1786.61	2094	0.8515
89	QUEDAR	V	1786.50	1862	0.9590
90	AHORA	ADV	1785.13	1920	0.9290
91	ESTO	PRON	1758.29	1811	0.9706
92	PASAR	V	1756.64	1906	0.9207
93	NADA	PRON	1722.46	2165	0.7933
94	SIEMPRE	ADV	1610.75	1667	0.9659
95	VIDA	S	1560.79	1697	0.9188
96	CASA	S	1546.17	1813	0.8512
97	S:ÓLO	ADV	1531.92	1713	0.8931
98	TOMAR	V	1505.79	1536	0.9801
99	FORMA	S	1501.64	1707	0.8783
100	TRABAJO	S	1495.21	1542	0.9693

deración estadística de las palabras más importantes[10].

Como se discutió en el punto 3 anterior, es la medida KF la que reúne toda la información numérica necesaria para calificar la importancia de un vocablo; mediante este criterio se identifican y ordenan las 100 primeras palabras[11], tal como se ve en el cuadro 2. En esta tabulación tenemos, en la primera columna, el número de orden del vocablo. En la segunda columna, se encuentra el vocablo en sí, y en la tercera aparece su categoría gramatical. La cuarta columna contiene el valor numérico de KF y en la quinta y sexta se consignan respectivamente la frecuencia absoluta y el valor numérico de C.

De esta manera, el cuadro 2 indica que la palabra de mayor KF es el artículo *la,* que tiene el valor KF = 85 919.13, y que a su vez también obtuvo la máxima frecuencia de aparición en el CEMC con un total de ocurrencias de 87 827. Esta primera palabra tiene un coeficiente de dispersión de 0.9780, lo cual señala una distribución sumamente regular entre los géneros. Esto último puede constatarse cuando se consulta el cuadro 3 donde aparece, otra vez para las primeras 100 palabras, la distribución de frecuencias tanto absolutas como relativas. Ahí notamos cómo para esa palabra *la,* los porcentajes de aparición entre los géneros no distan mucho de los tamaños relativos de los géneros de acuerdo con las

[10] Fue notorio que dentro del equipo que trabaja para el DEM existía la mayor excitación por conocer, al momento de producirse los primeros resultados de computadora, cuáles eran las palabras de mayor frecuencia y con qué frecuencias aparecían.

[11] Pudieron identificarse los 100 primeros vocablos gracias a la labor de identificación y de recuento que sobre los listados de computadora hizo Carlos Villanueva, del equipo del DEM. Véase nota 12.

CUADRO 3

Núm.	Vocablo	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
1	LA	12835	16893	21183	11703	1932	1366	2104	4054	1945	1562	7973	1837	992	1448
		14.61	19.23	24.12	13.33	2.20	1.56	2.40	4.62	2.21	1.78	9.08	2.09	1.13	1.65
2	EL	11562	16833	17920	10438	1639	1151	2197	3473	1716	1411	8291	1493	988	1373
		14.37	20.91	22.27	12.97	2.04	1.43	2.73	4.32	2.13	1.75	10.30	1.86	1.23	1.71
3	DE	10858	9453	9310	7288	818	635	2089	5313	2478	1588	8971	1836	1231	1593
		17.11	14.90	14.67	11.48	1.29	1.00	3.29	8.37	3.90	2.50	14.14	2.89	1.94	2.51
4	Y	8845	10392	10211	6270	1247	745	1343	3272	1933	1246	8781	2265	989	1937
		14.87	17.47	17.17	10.54	2.10	1.25	2.26	5.50	3.25	2.09	14.76	3.81	1.66	3.26
5	QUE	7407	10264	9977	5167	1101	629	2880	4340	2300	1676	7815	2074	1085	1880
		12.64	17.52	17.03	8.82	1.88	1.07	4.92	7.41	3.93	2.86	13.34	3.54	1.85	3.21
6	EN	6577	10566	11892	5637	1049	670	1647	2386	1133	664	4841	876	595	766
		13.34	21.43	24.12	11.43	2.13	1.36	3.34	4.84	2.30	1.35	9.82	1.78	1.21	1.55
7	A	6589	7305	6987	3701	813	467	1426	3482	1550	1131	6565	2312	875	1606
		14.70	16.30	15.59	8.26	1.81	1.04	3.18	7.77	3.46	2.52	14.65	5.16	1.95	3.58

Núm.	Vocablo	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
8	SE	4110	4931	6785	4746	412	295	1014	1952	890	622	5621	947	561	817
		12.19	14.63	20.13	14.08	1.22	0.88	3.01	5.79	2.64	1.85	16.68	2.81	1.66	2.42
9	NO	3982	2201	2865	1355	225	242	1917	2814	1876	969	8104	2011	951	1956
		12.65	6.99	9.10	4.31	0.72	0.77	6.09	8.94	5.96	3.08	25.75	6.39	3.02	6.22
10	SER	2958	2523	4588	2327	293	336	1375	1562	1251	592	4991	685	601	836
		11.87	10.13	18.41	9.34	1.18	1.35	5.52	6.27	5.02	2.38	20.03	2.75	2.41	2.36
11	UN	3450	2973	3680	2289	302	196	862	1493	703	505	2639	574	559	421
		16.71	14.40	17.82	11.09	1.46	0.95	4.18	7.23	3.41	2.45	12.78	2.78	2.71	2.04
12	POR	2672	4055	4140	2128	367	264	629	1342	465	435	1976	620	298	471
		13.45	20.42	20.84	10.71	1.85	1.33	3.17	6.76	2.34	2.19	9.95	3.12	1.50	2.37
13	CON	2735	3308	3681	2423	324	193	540	1377	452	459	1958	665	286	374
		14.57	17.62	19.61	12.91	1.73	1.03	2.88	7.33	2.41	2.44	10.43	3.54	1.52	1.99
14	SU	3709	3554	3288	1583	391	302	417	1802	351	233	1361	458	170	204
		20.81	19.94	18.45	8.88	2.19	1.69	2.34	10.11	1.97	1.31	7.64	2.57	0.95	1.14
15	UNA	2462	2245	3176	1897	201	197	794	1096	507	334	1799	398	279	307
		15.69	14.31	20.24	12.09	1.28	1.26	5.06	6.98	3.23	2.13	11.46	2.54	1.78	1.96

16	HABER	2228	2181	1957	993	246	140	672	1273	499	280	2925	477	314	434
		15.24	14.92	13.39	6.79	1.68	0.96	4.60	8.71	3.41	1.92	20.01	3.26	2.15	2.97
17	PARA	1500	2812	2554	2308	357	141	415	798	394	341	1658	388	202	296
		10.59	19.85	18.03	16.29	2.52	1.00	2.93	5.63	2.78	2.41	11.71	2.74	1.43	2.09
18	AL	1810	2387	2231	1279	227	130	291	740	188	253	1028	276	184	183
		16.15	21.30	19.91	11.41	2.03	1.16	2.60	6.60	1.68	2.26	9.17	2.46	1.64	1.63
19	ESTAR	1234	1259	1070	735	124	70	744	984	630	227	2727	770	445	654
		10.57	10.79	9.17	6.30	1.06	0.60	6.37	8.43	5.40	1.94	23.36	6.60	3.81	5.60
20	COMO	1734	1477	2117	944	137	162	348	668	315	181	1832	363	187	353
		16.03	13.65	19.57	8.73	1.27	1.50	3.22	6.17	2.91	1.67	16.93	3.36	1.73	3.26
21	TENER	1016	1056	1226	648	110	77	705	846	534	285	2077	572	200	466
		10.35	10.76	12.49	6.60	1.12	0.78	7.18	8.62	5.44	2.90	21.16	5.83	2.04	4.75
22	LE	1352	844	518	408	33	64	550	933	498	387	2982	909	363	698
		12.83	8.01	4.92	3.87	0.31	0.61	5.22	8.85	4.73	3.67	28.29	8.63	3.44	6.62
23	HACER	1125	1172	1016	804	101	91	578	782	374	147	1752	468	157	259
		12.75	13.28	11.51	9.11	1.14	1.03	6.55	8.86	4.24	1.67	19.85	5.30	1.78	2.93
24	YA	822	671	564	350	48	33	595	507	603	332	3116	840	452	882
		8.37	6.84	5.75	3.57	0.49	0.34	6.06	5.17	6.14	3.38	31.75	8.56	4.61	8.99

Núm.	Vocablo	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
25	O	816	759	2327	1359	130	78	409	331	329	28	1204	183	185	230
		9.75	9.07	27.81	16.24	1.55	0.93	4.89	3.96	3.93	0.33	14.39	2.19	2.21	2.75
26	PERO	1184	731	805	333	71	61	664	796	479	130	1814	622	215	438
		14.19	8.76	9.65	3.99	0.85	0.73	7.96	9.54	5.74	1.56	21.74	7.46	2.58	5.25
27	DECIR	1124	722	489	178	36	66	658	721	492	370	2158	831	258	778
		12.66	8.13	5.51	2.00	0.41	0.74	7.41	8.12	5.54	4.17	24.30	9.36	2.91	8.76
28	QUE	1715	797	952	548	65	39	156	528	204	101	2038	374	208	253
		21.50	9.99	11.93	6.87	0.81	0.49	1.96	6.62	2.56	1.27	25.55	4.69	2.61	3.17
29	LO	1128	1076	1357	603	106	82	332	670	269	113	1039	258	152	193
		15.29	14.58	18.39	8.17	1.44	1.11	4.50	9.08	3.65	1.53	14.08	3.50	2.06	2.62
30	ME	1526	146	107	75	26	2	690	1408	611	1087	1909	1601	287	952
		14.64	1.40	1.03	0.72	0.25	0.02	6.62	13.50	5.86	10.42	18.31	15.35	2.75	9.13
31	MAS	1026	755	1292	734	107	79	380	392	302	109	1258	201	106	223
		14.73	10.84	18.55	10.54	1.54	1.13	5.46	5.63	4.34	1.57	18.06	2.89	1.52	3.20
32	PODER	754	652	1724	856	86	60	266	538	237	122	614	199	109	133
		11.87	10.27	27.15	13.48	1.35	0.94	4.19	8.47	3.73	1.92	9.67	3.13	1.72	2.09

33	ESTE	376	631	922	511	74	52	309	183	433	58	2467	99	362	665
		5.26	8.84	12.91	7.15	1.04	0.73	4.33	2.56	6.06	0.81	34.54	1.39	5.07	9.31
34	IR	810	273	243	185	26	24	488	693	383	458	1851	837	257	519
		11.49	3.87	3.45	2.63	0.37	0.34	6.92	9.83	5.43	6.50	26.27	11.68	3.65	7.36
35	LO	1276	465	432	222	53	53	327	889	252	183	1257	357	121	231
		20.86	7.60	7.06	3.63	0.87	0.87	5.34	14.53	4.12	2.99	20.55	5.84	1.98	3.78
36	SI/	291	75	104	84	16	13	668	340	683	137	4844	300	255	846
		3.36	0.87	1.20	0.97	0.18	0.15	7.72	3.93	7.89	1.58	55.96	3.47	2.95	9.77
37	VER	845	360	348	215	16	20	396	499	334	312	1181	496	206	338
		15.18	6.47	6.25	3.86	0.29	0.36	7.11	8.97	6.00	5.61	21.22	8.91	3.70	6.07
38	DAR	687	589	637	320	57	71	235	455	193	243	807	403	103	245
		13.62	11.67	12.63	6.34	1.13	1.41	4.66	9.02	3.83	4.82	16.00	7.99	2.04	4.86
39	CUANDO	680	436	645	587	38	31	193	359	156	192	1001	356	87	146
		13.86	8.89	13.14	11.96	0.77	0.63	3.93	7.32	3.18	3.91	20.40	7.25	1.77	2.98
40	MUY	415	425	494	335	41	22	455	432	379	90	1299	206	152	168
		8.45	8.65	10.05	6.82	0.83	0.45	9.26	8.79	7.71	1.83	26.44	4.19	3.09	3.42
41	YO	819	51	36	45	10	15	737	597	439	479	1659	1080	317	737
		11.67	0.73	0.51	0.64	0.14	0.21	10.50	8.50	6.25	6.82	23.63	15.38	4.52	10.50

Núm.	Vocablo	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
42	PORQUE	494	314	256	120	48	34	410	242	421	260	1528	373	176	403
		9.73	6.18	5.04	2.36	0.95	0.67	8.07	4.76	8.29	5.12	30.08	7.34	3.47	7.93
43	EL	769	368	359	174	30	37	294	435	232	45	1063	355	102	247
		17.5	8.16	7.96	3.86	0.67	0.82	6.52	9.65	5.14	1.00	23.57	7.87	2.26	5.48
44	MI	1073	126	107	42	28	9	350	796	182	733	974	877	105	247
		18.68	2.19	1.86	0.73	0.49	0.16	6.09	13.86	3.17	12.76	16.96	15.27	1.83	5.95
45	PUES	262	235	208	80	16	25	513	157	559	32	3116	181	166	368
		4.43	3.97	3.51	1.35	0.27	0.42	8.67	2.65	9.45	0.54	52.65	3.06	2.81	6.22
46	LA	774	313	329	236	39	22	220	503	161	224	838	305	117	178
		18.17	7.35	7.72	5.54	0.92	0.52	5.17	11.81	3.78	5.26	19.68	7.16	2.75	4.18
47	ASI	357	384	406	221	37	36	195	225	190	41	1414	259	125	267
		8.59	9.24	9.77	5.32	0.89	0.87	4.69	5.41	4.57	0.99	34.01	6.23	3.01	6.42
48	ESTA	451	794	1017	554	114	46	80	203	72	67	161	76	40	60
		12.07	21.26	27.23	14.83	3.05	1.23	2.14	5.44	1.93	1.79	4.31	2.03	1.07	1.61
49	TODO	431	579	521	337	92	48	132	216	147	58	589	113	71	121
		12.47	16.76	15.08	9.75	2.66	1.39	3.82	6.25	4.25	1.68	17.05	3.27	2.05	3.50

50	TAMBIEN	332	527	547	303	44	52	191	151	174	44	856	118	73	86
		9.49	15.07	15.64	8.66	1.26	1.49	5.46	4.32	4.97	1.26	24.47	3.37	2.09	2.46
51	VEZ	538	338	478	287	40	27	105	265	130	20	686	174	51	100
		16.61	10.44	14.76	8.86	1.23	0.83	3.24	8.18	4.01	0.62	21.18	5.37	1.57	3.09
52	NOS	576	341	318	141	62	117	176	212	141	59	595	220	188	243
		17.00	10.06	9.38	4.16	1.83	3.45	5.19	6.26	4.16	1.74	17.56	6.49	5.55	7.17
53	A~~/~~O	282	703	477	280	81	7	226	95	127	7	758	73	47	131
		8.56	21.34	14.48	8.50	2.46	0.21	6.86	2.88	3.86	0.21	23.01	2.22	1.43	3.98
54	SABER	629	264	216	119	24	15	225	453	196	105	690	236	106	163
		18.28	7.67	6.28	3.46	0.70	0.44	6.54	13.16	5.70	3.05	20.05	6.86	3.08	4.74
55	SIN	776	440	703	395	58	30	66	338	43	53	149	79	31	27
		24.34	13.80	22.05	12.39	1.82	0.94	2.07	10.60	1.35	1.66	4.67	2.48	0.97	0.85
56	HASTA	450	350	418	377	28	24	96	210	104	73	481	211	64	148
		14.83	11.54	13.78	12.43	0.92	0.79	3.16	6.92	3.43	2.41	15.85	6.95	2.11	4.88
57	QUERER	575	210	135	79	40	28	174	583	127	420	606	325	89	155
		16.22	5.92	3.81	2.23	1.13	0.79	4.91	16.44	3.58	11.84	17.09	9.17	2.51	4.37
58	DEBER	307	477	998	710	65	67	81	198	52	18	122	49	33	22
		9.60	14.91	31.20	22.19	2.03	2.09	2.53	6.19	1.63	0.56	3.81	1.53	1.03	0.69

Núm.	Vocablo	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
59	TODO	577	265	241	162	33	24	197	258	192	42	681	156	165	144
		18.39	8.45	7.68	5.16	1.05	0.77	6.28	8.22	6.12	1.34	21.71	4.97	5.26	4.59
60	AQUI	264	132	132	49	18	22	127	190	160	122	2224	189	108	296
		6.55	3.27	3.27	1.21	0.45	0.55	3.15	4.71	3.97	3.03	55.15	4.69	2.68	7.34
61	UNO	312	266	236	176	17	17	170	81	129	19	1183	104	120	267
		10.07	8.59	7.62	5.68	0.55	0.55	5.49	2.62	4.17	0.61	38.20	3.36	3.87	8.62
62	DIA	361	386	205	189	34	29	89	269	168	96	601	210	22	126
		12.96	13.86	7.36	6.79	1.22	1.04	3.20	9.66	6.03	3.45	21.58	7.54	0.79	4.52
63	ESO	392	150	56	58	9	20	257	303	308	38	1085	242	146	288
		11.69	4.47	1.67	1.73	0.27	0.60	7.67	9.04	9.19	1.13	32.37	7.22	4.36	8.59
64	QUE/	568	118	119	67	8	80	285	548	145	94	403	354	127	260
		17.88	3.72	3.75	2.11	0.25	2.52	8.97	17.25	4.57	2.96	12.69	11.15	4.00	8.19
65	ELLA	448	197	271	133	26	31	128	429	133	41	407	201	22	145
		17.15	7.54	10.38	5.09	1.00	1.19	4.90	16.42	5.09	1.57	15.58	7.70	0.84	5.55
66	SOBRE	480	528	761	454	44	63	69	111	35	4	78	18	33	14
		17.83	19.61	28.27	16.86	1.63	2.34	2.56	4.12	1.30	0.15	2.90	0.67	1.23	0.52

67	BIEN	253	210	326	194	18	10	134	282	150	46	507	111	109	149
		10.12	8.40	13.05	7.76	0.72	0.40	5.36	11.28	6.00	1.84	20.29	4.44	4.36	5.96
68	LLEGAR	343	345	269	123	28	12	149	209	130	72	423	161	78	100
		14.05	14.13	11.02	5.04	1.15	0.49	6.10	8.56	5.32	2.95	17.32	6.59	3.19	4.10
69	MAS	244	446	308	185	29	14	58	144	89	15	709	80	48	96
		9.90	18.09	12.49	7.51	1.18	0.57	2.35	5.84	3.61	0.61	28.76	3.25	1.95	3.89
70	DONDE	358	394	272	203	13	15	85	147	84	90	434	74	57	94
		15.43	16.98	11.72	8.75	0.56	0.65	3.66	6.34	3.62	3.88	18.71	3.19	2.46	4.05
71	ENTRE	426	574	691	314	40	25	33	119	37	16	110	25	32	15
		17.34	23.36	28.12	12.78	1.63	1.02	1.34	4.84	1.51	0.65	4.48	1.02	1.30	0.61
72	NI	459	246	238	104	49	28	91	221	105	77	340	224	77	132
		19.20	10.29	9.95	4.35	2.05	1.17	3.81	9.24	4.39	3.22	14.22	9.37	3.22	5.52
73	OTRA	281	385	455	219	29	27	98	110	86	21	382	86	24	62
		12.41	17.00	20.09	9.67	1.28	1.19	4.33	4.86	3.80	0.93	16.87	3.80	1.06	2.74
74	ENTONCES	220	97	213	55	10	21	396	104	293	1	882	273	139	256
		7.43	3.28	7.20	1.86	0.34	0.71	13.38	3.51	9.90	0.03	29.80	9.22	4.70	8.65
75	ESA	357	320	236	107	32	29	133	181	140	39	516	84	71	88
		15.30	13.72	10.12	4.59	1.37	1.24	5.70	7.76	6.00	1.67	22.12	3.60	3.04	3.77

Núm.	Vocablo	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
76	LLEVAR	251	334	211	156	30	15	96	165	94	139	447	137	92	130
		10.93	14.54	9.19	6.79	1.31	0.65	4.18	7.18	4.09	6.05	19.46	6.96	4.01	5.66
77	PONER	262	196	140	257	22	17	102	180	114	81	664	111	51	122
		11.30	8.45	6.04	11.08	0.95	0.73	4.40	7.76	4.92	3.49	28.63	7.79	2.20	5.26
78	PARTE	234	478	632	342	51	29	66	85	43	5	195	26	17	38
		10.44	21.33	28.20	15.26	2.28	1.29	2.95	3.79	1.92	0.22	8.70	1.16	0.76	1.70
79	TE	581	11	23	83	0	9	266	781	167	581	256	315	199	120
		17.13	0.32	0.68	2.45	0.00	0.27	7.84	23.02	4.92	17.13	7.55	9.29	5.87	3.54
80	TIEMPO	377	285	429	225	21	17	86	140	51	20	264	55	30	68
		18.23	13.78	20.74	10.88	1.02	0.82	4.16	6.77	2.47	0.97	12.77	2.66	1.45	3.29
81	DOS	267	324	358	232	4	14	76	112	62	38	333	78	36	85
		13.22	16.05	17.73	11.49	0.20	0.69	3.76	5.55	3.07	1.88	16.49	3.86	1.78	4.21
82	DESPUES	330	269	310	204	11	16	90	161	117	13	299	128	18	69
		16.22	13.22	15.23	10.02	0.54	0.79	4.42	7.91	5.75	0.64	14.69	6.29	0.88	3.39

83	DEJAR	410	198	199	198	12	7	81	262	66	91	272	179	37	86
		19.54	9.44	9.49	9.44	0.57	0.33	3.86	12.49	3.15	4.34	12.96	8.53	1.76	4.10
84	DESDE	303	337	368	222	34	15	99	122	68	23	171	51	55	42
		15.86	17.64	19.27	11.62	1.78	0.79	5.18	6.39	3.56	1.20	8.95	2.67	2.88	2.20
85	HOMBRE	490	255	260	138	40	71	125	189	48	64	178	101	3	27
		24.64	12.82	13.07	6.94	2.01	3.57	6.28	9.50	2.41	3.22	8.95	5.08	0.15	1.36
86	ESE	287	248	193	80	16	14	107	198	106	41	446	90	56	98
		14.49	12.53	9.75	4.04	0.81	0.71	5.40	10.00	5.35	2.07	22.53	4.55	2.83	4.95
87	CADA	238	259	429	333	61	12	46	86	68	9	247	40	15	41
		12.63	13.75	22.77	17.68	3.24	0.64	2.44	4.56	3.61	0.48	13.11	2.12	0.80	2.18
88	VENIR	266	173	89	54	18	17	64	185	96	124	628	179	79	122
		12.70	8.26	4.25	2.58	0.86	0.81	3.06	8.83	4.58	5.92	29.99	8.55	3.77	5.83
89	QUEDAR	299	206	202	122	11	12	68	164	68	60	396	103	42	109
		16.06	11.06	10.85	6.55	0.59	0.64	3.65	8.81	3.65	3.22	21.27	5.53	2.26	5.85
90	AHORA	298	285	160	66	16	11	143	252	124	77	326	46	38	78
		15.52	14.84	8.33	3.44	0.83	0.57	7.45	13.12	6.46	4.01	16.98	2.40	1.98	4.06

Núm.	Vocablo	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
91	ESTO	234	262	405	226	32	44	104	112	81	5	145	92	30	39
		12.92	14.47	22.36	12.48	1.77	2.43	5.74	6.18	4.47	0.28	8.01	5.08	1.66	2.15
92	PASAR	313	149	145	115	8	7	126	216	74	90	411	111	60	81
		16.42	7.82	7.61	6.03	0.42	0.37	6.61	11.33	3.88	4.72	21.56	5.82	3.15	4.25
93	NADA	282	125	60	38	10	8	94	212	123	38	715	208	98	154
		13.03	5.77	2.77	1.76	0.46	0.37	4.34	9.79	5.68	1.76	33.03	9.61	4.53	7.11
94	SIEMPRE	283	138	224	151	17	17	98	183	79	32	250	113	24	58
		16.98	8.28	13.44	9.06	1.02	1.02	5.88	10.98	4.74	1.92	15.00	6.78	1.44	3.48
95	VIDA	346	288	245	107	50	56	75	208	35	170	93	46	17	21
		20.39	13.44	14.44	6.31	2.95	3.30	4.42	12.26	2.06	10.02	5.48	2.71	1.00	1.24
96	CASA	272	137	60	80	8	3	100	189	102	50	500	203	39	70
		15.00	7.56	3.31	4.41	0.44	0.17	5.52	10.42	5.63	2.76	27.58	11.20	2.15	3.86
97	SO/LO	369	306	463	146	58	37	9	170	10	48	45	29	14	9
		21.54	17.86	27.03	8.52	3.39	2.16	0.53	9.92	0.58	2.80	2.63	1.69	0.82	0.53

98	TOMAR	247	193	259	148	7	17	59	195	63	46	173	69	18	42
		16.08	12.57	16.86	9.64	0.46	1.11	3.84	12.70	4.10	2.99	11.26	4.49	1.17	2.73
99	FORMA	183	279	615	332	28	31	35	55	21	0	84	11	18	15
		10.72	16.34	36.03	19.45	1.64	1.82	2.05	3.22	1.23	0.00	4.92	0.64	1.05	0.88
100	TRABAJO	98	302	332	177	38	20	66	63	62	10	236	59	39	40
		6.36	19.58	21.53	11.48	2.46	1.30	4.28	4.09	4.02	0.65	15.30	3.83	2.53	2.59

cifras que muestra el cuadro 1. La comparación se muestra en el cuadro 4, entre la primera y la segunda columna numéricas. El alto valor de C para el vocablo *la* sugiere un uso uniforme en todos los géneros, cosa que se corrobora en la comparación entre los porcentajes citados.

Cuadro 4

Género	Tamaño relativo del género	% entre géneros del vocablo *la* (art.)	% entre géneros del vocablo *me* (pron.)
Literatura	14.27	14.61	14.64
Periodismo	15.85	19.23	1.40
Ciencias	18.31	24.12	1.03
Técnicas	10.72	13.33	0.72
Discursos políticos	1.69	2.20	0.25
Religión	1.13	1.56	0.02
Habla culta	3.67	2.40	6.62
Literatura popular	6.74	4.62	13.50
Habla media	3.15	2.21	5.86
Lírica popular	2.39	1.78	10.42
Textos dialectales	13.74	9.08	18.31
Docts. antropológicos	3.62	2.09	15.36
Jergas	1.84	1.13	2.75
Habla popular	2.88	1.65	0.13

Con una interpretación semejante podemos seguir la lectura de los cuadros 2 y 3. La segunda palabra en importancia numérica del español de México es el artículo *el* (otro artículo, parece ser que denominamos mucho, y más frecuentemente en el femenino). Este vocablo tiene un KF de 78 942.85 y una frecuencia total de 80 485. Su dispersión es de 0.9806, lo cual indica que también tiene una alta uniformidad de uso entre todos los géneros. Siguiendo el análisis de esas medidas estadísticas nos topamos por ejemplo con

el vocablo *me,* utilizado como pronombre. El lugar que ocupa en la lista es el número 30 y su frecuencia de aparición, de 10 427, difiere notablemente de la KF con un valor de 6 928.90. Del lugar 22 que le hubiera correspondido debido a tan alta frecuencia absoluta, su sitio asignado descendió hasta el lugar 30. La explicación la proporciona el relativamente bajo valor del índice C que le corresponde y que es de 0.6607, el cual indica una alta irregularidad en su uso cuando se toman en cuenta los géneros. En el cuadro 4 se lee en la última columna los porcentajes de aparición de este vocablo entre los géneros. Al comparar las cifras con la de los tamaños relativos de los mismos géneros notamos la no uniformidad que señala el valor de 0.6607 para C. A pesar de que entre Periodismo, Ciencias y Técnicas hay un 44.88% del CEMC, el pronombre *me* solo ocurrió en un 3.15% en esos mismos géneros. Por el contrario, en géneros como Lírica Popular, Documentos Antropológicos y Habla Popular, que ocupan tan sólo un 8.89% del tamaño total del CEMC, el vocablo *me* ocurrió un 34.90% de las veces. Otras diferencias, aunque menos tajantes, también existen en los otros géneros. Otro ejemplo, también ilustrativo de este tipo de vocablos de uso no uniforme es el caso del adverbio *sí,* que ocupa el lugar 36. Por frecuencia total 8 656, debería ocupar el lugar 27, sin embargo su irregularidad en la distribución entre géneros lo lleva a un valor de KF de 4 976.58 y a un índice C de dispersión de 0.5701, el más bajo entre las 100 primeras palabras aquí estudiadas. El cuadro 3 nos da su distribución entre géneros y el aspecto que más salta a la vista es el inmenso uso que se hace de ese término en el género 11, de los textos dialectales, donde con un

tamaño relativo de 13 75% se observa un 55.96% de las ocurrencias de *si*[12].

Otro ejemplo interesante lo proporcionan los pronombres personales *nos* y *te*. *Nos* ocupa el lugar 52 y *te* el lugar 79. La frecuencia total observada es de 3 389 para el primer caso y de 3 392 para el segundo. Aunque de frecuencia parecida, e incluso ligeramente mayor para el caso de *te*, KF clasifica bastante prioritariamente a *nos* respecto a *te*. La explicación aparece en el cuadro 3, donde se especifican las distribuciones de frecuencia de observación entre los géneros. Mientras que para *nos* los porcentajes de aparición se parecen bastante a los tamaños relativos entre géneros, para *te* hay una franca preferencia por su aparición en géneros tales como la Literatura, la Literatura Popular y la Lírica. El índice C vuelve a reflejar la situación descrita, pues mientras que es de 0.9277 para *nos* es apenas de 0.5993 para *te*.

En esta misma forma se pueden analizar no sólo los 100 vocablos de este trabajo, sino las cifras y estadísticas de todos los vocablos y tipos a lo largo del CEMC. Para cada uno se tiene el recuento de las frecuencias absolutas y relativas, el cálculo de los índices estadísticos y las referencias para analizar los contextos dentro de los cuales se producen las palabras. Será el equipo del DEM y sus redactores quienes realicen la tarea exhaustiva de revisión e investigación para propósitos lexicográficos.

5. Indicadores de la riqueza del léxico

Una de las cuestiones más interesantes a las que se puede aspirar cuando se dispone de un

[12] Algo en las encuestas dialectales hace que sus informantes afirmen demasiadas cosas.

corpus de gran extensión, como es el caso del CEMC, es la de tener la oportunidad de explorar la idea de "qué tan rico es un léxico". Este último concepto en realidad merecería el esfuerzo de que se precisara y se buscara una definición concreta del significado que encierra; sin embargo, en la etapa de investigación que se manifiesta en este trabajo lo tomamos bajo dos ideas totalmente intuitivas y parciales. Por una parte, la riqueza del léxico será un recuento de los vocablos distintos que el corpus identifica, esperando que la propia extensión del corpus garantice una suficiente aproximación en el número identificado. Por otra parte, será la distribución de frecuencias con las cuales se utilizan esos vocablos ya identificados. El primer sentido es bastante natural y comprensible. El segundo es una conceptualización de mayor minuciosidad y que requiere de mayor explicación. Se espera que con las explicaciones y cuadros numéricos que siguen se logre un esclarecimiento al respecto[13].

Al hablar de la riqueza de un léxico debemos, antes que cualquier otra cosa, acordar de qué léxico se trata. El CEMC nos da la oportunidad

[13] En este punto es necesaria una aclaración. Para el trabajo lexicográfico, el CEMC facilita dos instrumentos principales: uno es el de proporcionar los contextos dentro de los cuales se utilizan los vocablos que lo componen, y el otro es el diccionario de frecuencias y medidas estadísticas que provee hechos numéricos en cuanto al uso de los vocablos. Este segundo instrumento es el que abre las posibilidades de investigación, como es el caso de la cuantificación de la importancia de un término, tal y como se trató en los puntos anteriores, y también los dos últimos temas citados sobre número de vocablos distintos y sobre la distribución de frecuencia de uso. En tal sentido se piensa que es más útil para la lexicografía un diccionario de frecuencias de vocablos más que de tipos, sin que esto indique que uno de tipos careciera de utilidad en investigaciones léxicas. Desafortunadamente, el estado actual del diccionario de frecuencias y medidas estadísticas producto del CEMC, es tal que nos coloca en una posición in-

de trabajarlo en distintas agregaciones, dada su clasificación en niveles de lengua y en géneros, aunque siempre sea dentro del español de México. De tal suerte, podemos referirnos a la riqueza de cada uno de los géneros, o de cada nivel de lengua, o del propio español de México como totalidad. También existe la factibilidad de efectuar exploraciones al respecto dentro de ciertos agregados, como podían ser por ejemplo varios géneros que en conjunto presenten un interés particular. En este trabajo presentamos únicamente posiciones relativas sobre riqueza léxica entre géneros y entre niveles de la lengua, tal y como las pone en evidencia el CEMC.

Nos encontramos en el cuadro 5 con los datos del tamaño del vocabulario encontrado y de la extensión del corpus utilizado. Las cifras se pre-

termedia entre uno puramente de vocablos y otro puramente de tipos. La razón de la irregularidad estriba en el hecho de que el procesamiento numérico que efectúa el conteo de frecuencias y el cálculo de porcentaje e índices estadísticos, intenta avanzar en la producción del diccionario de frecuencias de vocablos utilizando un analizador gramatical (véase en este mismo volumen el artículo de Isabel García Hidalgo "La formalización del analizador gramatical del DEM") el cual permite agrupar la mayor parte de las ocurrencias en la forma correspondiente del vocablo. Sin embargo, esto no se logra en todos los casos debido a ambigüedades difíciles de resolver en forma automática, como es el caso de una ocurrencia como "fue", que no se sabe fuera del contexto en que se encuentra si se refiere al verbo "ser" o al verbo "ir". Tampoco se logra el agrupamiento en otras situaciones como es el caso de los tipos de aumentativos o diminutivos de un adjetivo determinado, respecto de los cuales este último constituye el vocablo. Bajo estas circunstancias temporales, se presentan los resultados acerca del número de vocablos distintos encontrados y sobre las distribuciones de frecuencia de uso, con la advertencia de que no son los definitivos. No obstante lo anterior, sí podemos anticipar que estos resultados, cuando se llegue a la agrupación total de tipos en vocablos, no tendrán diferencias notables de las que aquí se presentan. En cualquier caso, los productos definitivos serán siempre en el sentido de mostrar "menos rico" el léxico en estudio, por la razón sencilla de que se identificarán menos vocablos y se concentrarán las distribuciones de uso, debido precisamente a que habrá tipos antes dispersos agrupados ahora en vocablos.

sentan para el corpus en su totalidad, para los tres niveles de lengua y para los catorce géneros clasificados. Tanto en el caso de los niveles de la lengua como en el de los géneros, se han puesto en orden decreciente respecto al tamaño del corpus en cada clasificación. En el mismo cuadro, en las dos últimas columnas, se especifica el número de orden que la clasificación tiene, atendiendo precisamente al tamaño del corpus y enseguida a la extensión del vocabulario en cada caso.

Al leer las dos últimas columnas del cuadro 5, notamos alta correlación entre el orden marcado por el tamaño del corpus y el orden asignado por la extensión del vocabulario encontrado. En el caso de los tres niveles de la lengua esta correlación entre rangos es perfecta. Cuando se trata de los géneros, si bien esta correlación no es total, sí la encontramos bastante alta. Esto de inmediato sugiere el hecho, por demás natural y que podía incluso anticiparse, de que a mayor tamaño del corpus, mayor oportunidad tiene el vocabulario de ser más extenso. Este hecho viene simplemente de que con un corpus más grande hay mayor probabilidad de identificar nuevos vocablos. Sin embargo, al tratar la clasificación en géneros nos encontramos con el hecho de que no siendo perfecta la correlación entre rangos, nos permite discernir que algunos de los géneros son de mayor riqueza léxica que otros. La manera de llevar a cabo ese ordenamiento parcial entre los géneros es mediante el sencillo procedimiento de calificar a un género como de mayor riqueza léxica que otro, cuando teniendo el primero un tamaño de corpus menor que el segundo, no obstante se da que la extensión del vocabulario del primer género es mayor que el segundo. En otras

palabras, cuando encontramos más vocabulario a pesar de que sea un género con un número menor de ocurrencias.

Con la norma anterior y adoptando la notación RL(G_i) > RL (G_j) para expresar el resultado de que la riqueza léxica del género i-ésimo es mayor que la riqueza léxica del género j-ésimo, nos encontramos con la siguiente ordenación parcial, utilizando los resultados expresados en el cuadro 5:

RL (Literatura) > RL (Periodismo)
RL (Técnicas) > RL (Textos dialectales)
RL (Discursos políticos) > RL (Jergas)
RL (Discursos políticos) > RL (Lírica popular)
RL (Jergas) > RL (Lírica popular)
RL (Jergas) > RL (Habla popular)

Cuadro 5

Clasificación	Tamaño del corpus	Tamaño del vocabulario	Lugar respecto al Corpus	Lugar respecto al Vocab.
Español de México	1 891 045	64 183	—	—
Lengua Culta	1 241 313	53 714	1	1
Lengua no estándar	417 557	18 250	2	2
Lengua sub-culta	232.175	16 982	3	3
Ciencias	346 313	26 487	1	1
Periodismo	299 775	21 072	2	3
Literatura	269 788	24 483	3	2
Textos dialectales	258 881	13 095	4	5
Técnicas	202 716	17 836	5	4
Literatura popular	127 459	12 876	6	6
Habla culta	69 473	6 291	7	7
Docts. antropológicos	68 376	6 082	8	8
Habla media	59 567	5 870	9	9
Habla popular	54 461	4 548	10	12
Lírica popular	45 149	4 172	11	13
Jergas	34 839	4 710	12	11
Discursos políticos	31 971	5 516	13	10
Religión	21 277	3 979	14	14

Esta ordenación propuesta es bastante limitada y no permite muchas otras comparaciones de posible interés entre parejas de géneros. Sin embargo, esta limitación puede subsanarse haciendo uso del concepto que nos resta por explorar y que es el de la frecuencia de utilización de los vocablos.

La idea que sustenta al empleo de la frecuencia de uso de los vocablos para fines de indicio sobre la riqueza léxica es la de que no basta tan sólo el número de vocablos distintos existentes en la lengua o en una subdivisión de ésta, sino que también hay que tomar en cuenta la frecuencia con la que son utilizados. Aunque en una lengua o género existan muchos vocablos disponibles no podemos decir que realmente existe gran riqueza léxica si al momento de la producción de textos la frecuencia de uso de estos se concentra en una minoría, haciéndose uso no significante de los demás. Por el contrario, si en otra lengua o género que tuviera o no menos vocablos, aunque no sustancialmente menor, se hiciera un uso más disperso de los vocablos, tendríamos en realidad una situación de mayor riqueza léxica manifestada en el hecho del mayor aprovechamiento del vocabulario disponible.

La manera de aprovechar el concepto expresado es a través de la frecuencia acumulada de uso de los vocablos. Supongamos que f_i denota la frecuencia de uso del vocablo V_i. Con el ordenamiento de las palabras por frecuencia de uso mayor a menor, de manera que $f_i \geqslant f_2 \geqslant \ldots \geqslant f\,\Omega$, donde Ω resulta ser el número de vocablos distintos encontrados, estamos en posibilidades de destacar cuál es el número de vocablos distintos necesarios para cubrir un porcentaje determinado de ocurrencias dentro del corpus. La forma

de lograrlo es a través de la distribución acumulativa de frecuencia de uso de los vocablos respetando la ordenación anterior indicada. Esto es, denotando por $F_i = f_1 = f_2 \ldots + f_i$, con $1 \leqslant i \leqslant \Omega$. F_i indica el porcentaje de ocurrencias dentro del corpus de las i palabras de mayor frecuencia. Con el cálculo de la distribución acumulativa de frecuencias F_i, también se puede proceder a la inversa en el sentido de preguntarnos, para un porcentaje de ocurrencias dentro del corpus, cuál es el número de palabras de mayor frecuencia que se ocupan. Esto es lo que en estadística se conoce por deciles de una distribución.

Cuando se tiene construida la distribución acumulada de frecuencias, es interesante conocer el valor de los deciles en ciertos porcentajes, que indiquen la forma de la distribución. De los más utilizados son los denominados cuartiles y que son los valores para los cuales la distribución acumula 25, 50, 75 y 100% de las observaciones.

En el cuadro 6 se consignan los valores aproximados[14] de los cuartiles para el corpus total, para cada nivel de la lengua y para cada género estudiado. En el corpus completo los valores encontrados significan que son solamente 9 palabras las que se utilizan el 25% de las veces; con 74 ya ocupamos la mitad de las ocurrencias; con 1 131 se tiene el 75% y hay un total de 64 183 palabras distintas en la totalidad del corpus. El mismo tipo de lectura se hace para cada nivel de

[14] Los valores de los cuartiles, expresados en números enteros de palabras no coinciden exactamente con los porcentajes señalados, sobre todo en el caso de los dos primeros, correspondientes al 25 y 50%. Por facilidad de interpretación se toman las aproximaciones citadas.

la lengua y cada género representado en el CEMC.

En esta parte tenemos que hacer una advertencia. El tamaño del corpus como total y en sus subdivisiones tiene desde luego una influencia directa en los valores para los cuales se completa un determinado porcentaje de ocurrencias. La idea es desde luego, que el valor al 100%, el del número de palabras distintas encontradas en todo el corpus, se incrementa cuando el tamaño del corpus crece, dada la posibilidad de identificar nuevos vocablos tal y como ya se ha comentado, aunque también es necesario comentar que la probabilidad de identificar vocablos nuevos decrece notablemente conforme aumenta el tamaño del corpus. Simplemente cada vez se requiere de mayores longitudes de texto para localizar nuevos vocablos. Sin embargo, en los demás valores, aquellos correspondientes a los porcentajes 25, 50 y 75 no es muy claro qué sucede si se incrementa el tamaño del corpus después de cierto límite.

Más ocurrencias ciertamente pueden aumentar el número de vocablos distintos, pero no necesariamente modifican el número de vocablos para completar las primeras 25, 50 o 75% de las ocurrencias. De lo único que sí podemos estar seguros es de que un mayor corpus reforzará la estimación de la frecuencia de uso de los vocablos más comunes, los que tienen la mayor probabilidad de aparecer y que son los que dan lugar a los valores de los cuartiles. Este reforzamiento en las estimaciones no implica desplazamientos sustanciales en los valores. Así por ejemplo, de acuerdo con la experiencia observada, se tiene que en el género del Habla Culta, con 11 palabras se ocupan un 25% de las ocurrencias, y no hay razón que nos diga que el hecho de au-

mentar el corpus en este género modificará a 10 o a 12 palabras, las necesarias para tener el 25% de las ocurrencias. Lo que más bien parece ser predecible es que volverá a ser 11 el valor calculado, sólo que con mayor confiabilidad en la estimación.

Admitiendo en este momento que en todos los casos el corpus proporciona suficiente tamaño como para hacer estables las estimaciones de los cuartiles correspondientes a los porcentajes 25, 50 y 75, podremos utilizar esos valores como indicadores de la riqueza del léxico bajo la siguiente interpretación: un género o nivel de la lengua presenta evidencias de mayor riqueza léxica si los valores de esos cuartiles son mayores, pues esto último indicaría un mayor uso de los vocablos disponibles.

En el cuadro 6 notamos que el primer cuartil es —como se adelantó— 9 para el corpus total y entre 9 y 10 para los tres niveles de la lengua. Esto nos dice que del total de palabras que producimos son sólo 9 o 10 las que utilizamos hasta en un 25% de las ocurrencias. Cuando se trata de los géneros, hay un mayor rango de este valor pues va desde 9 hasta 11, aunque en realidad no es una gran dispersión. Este primer valor para el primer 25%, tiene poco poder discriminador. Los valores que más parecen servir para nuestros propósitos de apuntar hacia comparaciones sobre riqueza léxica son los correspondientes a los porcentajes 50 y 75, puesto que su variabilidad es mayor y por lo tanto su poder como indicadores también lo es.

Al revisar los valores correspondientes a la distribución acumulada en el 50 y 75% resalta la siguiente ordenación de la riqueza léxica en cada nivel de la lengua:

RL (lengua culta) > RL (lengua subculta) >
> RL (lengua no estándar).

De igual manera, aprovechando el mismo cuadro 6 es posible hacer una ordenación entre géneros respecto a la mayor o menor riqueza léxica relativa, resultado de la lectura del número de palabras distintas para alcanzar el 50 y el 75% de las ocurrencias. El resultado de las comparaciones se expresa en el cuadro 7 en forma matricial con la siguiente interpretación: se compara cada género con todos los demás y esta comparación se hace del género que se lee en un renglón con el que se lee en una columna. Si en el cruce se consigna un símbolo + esto indica que para el primer género, el del renglón, hay evidencias que apuntan a que tiene una mayor riqueza

Cuadro No. 6

	25%	50%	75%	100%
CEMC	9	74	1 131	64 183
L. Culta	9	100	1 451	53 714
L. Sub-culta	10	60	560	16 982
L. no estándar	10	47	298	18 350
Literatura	9	81	1 118	24 483
Periodismo	8	85	1 195	21 072
Ciencias	9	110	1 382	26 487
Técnicas	9	117	1 308	17 836
Discursos políticos	8	83	780	5 516
Religión	8	62	565	3 979
Habla culta	11	51	330	6 291
Literatura popular	11	65	631	12 876
Habla media	9	46	308	5 870
Lírica popular	10	49	332	4 172
Textos dialectales	9	44	279	13 095
Docts. antropológicos	11	52	282	6 082
Jergas	11	55	345	4 710
Habla popular	10	45	216	4 548

Cuadro 7

	Literatura	Periodismo	Ciencias	Técnicas	Discursos Políticos	Religión	Habla Culta	Literatura Popular	Habla Media	Lírica Popular	Textos Dialectales	Documentos Antropológicos	Jergas	Habla Popular
Literatura		=	—	—	+	+	+	+	+	+	+	+	+	+
Periodismo	=		—	—	+	+	+	+	+	+	+	+	+	+
Ciencias	+	+		+	+	+	+	+	+	+	+	+	+	+
Técnicas	+	+	—		+	+	+	+	+	+	+	+	+	+
Discursos Políticos	—	—	—	—		+	+	+	+	+	+	+	+	+
Religión	—	—	—	—	—		+	—	+	+	+	+	+	+
Habla Culta	—	—	—	—	—	—		—	+	=	+	+	—	+
Literatura Popular	—	—	—	—	—	+	+		+	+	+	+	+	+
Habla Media	—	—	—	—	—	—	—	—		—	+	+	—	+
Lírica Popular	—	—	—	—	—	—	=	—	+		+	+	—	+
Textos Dialectales	—	—	—	—	—	—	—	—	—	—		—	—	+
Documentos Antropológicos	—	—	—	—	—	—	—	—	—	—	+		—	+
Jergas	—	—	—	—	—	—	+	—	+	+	+	+		+
Habla Popular	—	—	—	—	—	—	—	—	—	—	—	—	—	

léxica que el segundo género, el de la columna. Cuando la situación es la opuesta y la riqueza léxica parece ser menor, el símbolo que aparece es —. En una situación en la que no hay evidencia clara en ningún sentido se apunta el símbolo =.

Como se indica en el título de este trabajo, se han presentado los primeros resultados estadísticos, del 1 al 100 dentro de una gran enumeración por venir. Por ello mismo los hallazgos comentados también tienen mucho de preliminares y esto último es especialmente cierto acerca de los resultados del cuadro 7. En él se apunta simplemente qué género puede tener mayor riqueza léxica que otro, sin decir qué tan mayor. También es posible que algunas de las conclusiones deban revisarse, ya sea por mayor precisión en los conceptos o por mejoras en el corpus. Un ejemplo parece ser la situación del *habla culta* que presenta menor riqueza que la *literatura popular* y que las *jergas*. La explicación de esta situación parece provenir de la manera como fue obtenido el corpus, pues obedeció más a una situación de disponibilidad inmediata que a la deseable situación de selección aleatoria como fue en el caso de los textos escritos. Quizá la explicación pudiera verse al revés, en el sentido de que este resultado ilógico es realmente ilógico y que la razón de él estriba en el hecho de que por alguna razón las cintas que componen esta sección del CEMC se encuentran circunscritas a un diálogo restringido.

La Formalización del Analizador Gramatical del DEM

María Isabel García Hidalgo

1. A fines de 1973 se comenzó a discutir la posibilidad de hacer uso de la computadora electrónica en la investigación para el *Diccionario del español de México*. Las proporciones del corpus que deseaba manejar el equipo de lingüistas del DEM, las necesidades de objetividad y regularidad del análisis de los textos hablados y escritos en español mexicano, y la serie de datos y medidas estadísticas que planteaban como condiciones primarias para la elaboración del diccionario, aconsejaban la aplicación de la computación electrónica en forma intensiva. Así se estableció la colaboración entre ellos y el Centro de Procesamiento y Evaluación "Dr. Arturo Rosenblueth" de la Secretaría de Educación Pública. Desde entonces se ha trabajado en forma interdisciplinaria. Ha correspondido al Centro Rosenblueth hacer el análisis de los problemas lingüísticos y no-lingüísticos planteados por el equipo del DEM, diseñar los algoritmos computacionales necesarios, llevar a cabo la programación de todo el sistema y producir los resultados requeridos. También ha sido allí donde se ha hecho el procesamiento de todo el material, según los límites de tiempo y de equipo humano y electrónico del Centro. Lo que sigue es una exposición de la forma como se ha realizado este trabajo hasta hoy, en que el corpus coleccionado ha sido analizado y se han obtenido los resultados estadísticos y lingüísticos requeridos.

El deseo original del equipo de lingüistas era que, con la ayuda de la computadora, se produjeran las listas de las diferentes palabras que se encontraran en el CEMC y se asociaran a cada palabra su frecuencia de aparición y algunos valores estadísticos que sirvieran para medir su representatividad[1]. Además de esto, se deseaba que, para cada una de las palabras encontradas en el corpus, se pudiera recuperar un número razonable de contextos (*concordancias*) con los diferentes usos del vocablo en cuestión; esto último se consideraba documento imprescindible para llevar a cabo la redacción de los artículos del DEM. Considerando el hecho de que se carece de un diccionario de la lengua española que pueda ser la base de un diccionario de máquina[2], iniciamos un diálogo entre los miembros de los dos grupos (lingüistas y matemáticos) del cual saldría como resultado el planteamiento de las bases lingüísticas del sistema de análisis gramatical cuya descripción me ocupará en lo sucesivo.

Nota: Este trabajo es una versión completa de uno presentado en la International Conference on Computing in the Humanities, realizada en Waterloo, Canadá, en julio de 1977.

[1] Cf. en este mismo volumen los trabajos de Roberto Ham y Luis Fernando Lara sobre el análisis estadístico realizado para el DEM.

[2] La elaboración de un diccionario de máquina que contenga el mayor número de vocablos pertenecientes a una lengua es un procedimiento común en lingüística computacional. Cuando por cada vocablo se han asentado sus posibles morfemas de género, número, persona, tiempo, modo y la serie de sus derivativos, el reconocimiento automático se hace más fácil aunque produce dificultades computacionales de otra índole. En cualquier forma ese procedimiento no podía aplicarse en nuestro caso particular en que, precisamente, lo que se intenta construir es un diccionario. Sobre esto y los motivos por los que el Diccionario de la Real Academia Española no puede constituir una base para un diccionario de máquina, cf. Luis Fernando Lara, "Méthode en lexicographie: valeur et modalité du dictionnaire de machine", *CLex*, 29,2 (1976), 103-128.

Los problemas básicos que debían ser resueltos para que en la computadora se realizara el conteo de frecuencia de los vocablos en forma adecuada eran la diferenciación entre homógrafos y la agrupación de todas las diferentes formas de una palabra. La diferenciación entre homógrafos podría llevarse a cabo si las palabras del corpus tuvieran asociada su categoría gramatical. La solución planteada contemplaba como requisito indispensable la asociación automática de categoría gramatical a cada ocurrencia del CEMC. Se decidió que algunas reglas de reconocimiento de la morfología de los vocablos, así como algunas otras basadas en las relaciones de precedencia entre vocablos de un sintagma y las reglas de concordancia que permitían manejar las morfológicas y las de precedencia simultáneamente, resultaban ser suficientemente consistentes como para pensar en diseñar algoritmos computacionales que, haciendo uso de tales reglas, produjeran el análisis sintáctico del corpus, es decir, el etiquetamiento gramatical de todos los vocablos contenidos en él.

A continuación describiré cómo se implementaron las reglas morfológicas y de precedencia para su utilización por la computadora. No daré detalles que se refieran a la justificación teórica lingüística del uso de esas reglas[3] sino que la descripción será desde el punto de vista computacional.

1.1. *Un diccionario de máquina mínimo.*

Sea $\psi = \{ X \mid X$ es un vocablo español que ocurre en el CEMC o X es un símbolo de puntuación$\}$

[3] Cf. L.F. Lara, "Méthode en lexicographie. . ." y Ma. Angeles

y sea $\psi^* = \{ x \mid x = \ldots PXQ \ldots$ es una cadena de elementos de ψ, de longitud arbitraria, pero que ocurre en el CEMC$\} =$

$= \{$todos los contextos en el CEMC para cada vocablo en $\psi\}$

Formemos un conjunto $D \subset \psi$ con 360 vocablos de los paradigmas de <artículo>, <preposición>, <conjunción>, <pronombre>, <contracción>, y <adverbio>, así como con algunas formas muy frecuentes de verbos auxiliares y todos los símbolos de puntuación. Otro conjunto $D' \subset D$ está formado por todos aquellos elementos que pueden funcionar con dos categorías gramaticales pero para los cuales, en un contexto dado, se puede decidir con qué categoría están siendo usados, si se sabe solamente el vocablo que les precede en el corpus, o si se tiene información específica sobre la categoría gramatical del vocablo que las sigue; D' contiene 93 elementos. Los elementos de D y D' se enlistarán posteriormente, cuando se haga la descripción de la implementación computacional de todo el sistema.

1.2. *Una función* **cg** *para asociar categoría gramatical a los elementos del diccionario de máquina mínimo.*

1.2.1 Las categorías gramaticales.

Sea G = ⟨ <artículo>, <preposición>, <con-

Soler, "Problemas en la elaboración de un diccionario de máquina" próximo a publicarse en las Actas del IV Congreso de la Asociación de Lingüística y Filología de América Latina, Lima, 1974.

junción>, <pronombre>, <contracción>, <adverbio>, <nominal>, <verbo>, <ambigua>}

y sea cg* : D → G la relación que asocia a cada vocablo X∈D todas las posibles categorías gramaticales con que X puede funcionar en un contexto arbitrario de ψ^*. Si denotamos como $\overline{c}$ al conjunto de valores que puede tomar cg*(X), entonces $\overline{c}$ puede estar formado por un solo elemento, por ejemplo:

cg*(y) = {<conjunción>},
cg*(de) = {<preposición>},
cg*(él) = {<pronombre>},
cg*(soy) = {<verbo>},
cg*(aquél) = {<pronombre>}, etc.

Sin embargo, $\overline{c}$ puede contener dos elementos como en los siguientes casos:

cg*(cual) = {<adverbio>, <pronombre>},
cg*(la) = {<artículo>, <pronombre>},
cg*(cuyo) = {<adjetivo>, <nominal>},
cg*(una) = {<artículo>, <pronombre>}.

En el caso en que cg*(X) contiene tres o más elementos, como por ejemplo cg*(cerca) = {<verbo>, <nominal>, <adjetivo>} convenimos en que cg*(X) = {<ambigua>}.

Sea $C_Y = \{XYZ \mid Y \in D;\ X, Y, Z \in \psi\}$ para cada Y∈D y sea $C = C_{Y\ 1} \cup C_{Y\ 2} \cup \ldots \cup C_{Y\ 360}$ entonces podemos construir la función:

$$cg : D \times C \rightarrow G$$

en donde cg(Y,XYZ) es la categoría gramatical de Y en el contexto XYZ∈C_Y. Hacemos notar

que los elementos $Y \epsilon D$ pueden ser de tres tipos: aquellos que funcionan con una categoría fija, independientemente del contexto en que aparezcan, aquellos que pudiendo funcionar con exactamente dos categorías, se desambiguan con un contexto XYZ y aquellos a los que se asocia convenientemente la categoría <ambigua>; por lo consiguiente, la función cg está bien definida.

1.2.2. *La relación de equivalencia.*

El conjunto D fue dividido en 31 clases de equivalencia $I_1, \ldots I_{30}, I_{32}$ de acuerdo con el siguiente criterio: si X y X' están en D, entonces X es equivalente a X', si y sólo si, para toda Y en D la categoría gramatical de Y en el contexto XYZ es igual a la categoría gramatical de Y en el contexto X'YZ', con Z y Z' elementos arbitrarios de ψ, es decir:
si $X,X' \epsilon D$ entonces $X \sim X' \Leftrightarrow$ para toda $Y \epsilon D$, dados $XYZ,X'YZ' \epsilon C_Y$ se cumple que $cg(Y,XYZ) = cg(Y,X'YZ')$.

Para aclarar la anterior definición consideremos las siguientes oraciones tomadas del corpus:

"tirando de ellas *cual* si fuesen toros bravos" y
"y del *cual* saqué la aclaración",
si hacemos X = *ellas,* X' = *del,* Y = *cual,* Z = *si,* Z' = *saqué,* podemos observar que cg(Y, XYZ) = cg(cual, ellas cual si) = <adverbio> en tanto que cg(Y,X'YZ') = cg(cual, del cual saqué) = <pronombre>, por lo tanto, existe por lo menos una Y = *cual* en D para la cual dados los contextos XYZ = *ellas cual si* y X'YZ' = *del cual saqué* en C_{cual} no se cumple que cg(Y,XYZ) sea igual a

cg(Y,X'YZ'), concluimos que *ellas* γ *del*. Consideremos, sin embargo, las siguientes frases:

"cada *cual* tomó su camino" y
"y del *cual* saqué la aclaración",

en este caso cg(cual, cada cual tomó) = cg(cual, del cual saqué) = <pronombre> lo que indica que *cada* y *del* cumplen la condición de equivalencia. Tomemos ahora los siguientes ejemplos:

"cada *cuyo* se apareó con su pareja" y
"el cerebelo del *cuyo* fue estudiado histológicamente",

cg(cuyo, cada cuyo se) = cg(cuyo, del cuyo fue) = <nominal>, es decir, una vez más *cada* y *del* cumplen la condición de equivalencia. La conclusión final es que *cada* $\sim$ *del*.

La presentación de los ejemplos anteriores tiene el propósito de indicar al lector los razonamientos lingüísticos que se utilizaron en la selección y partición en clases de equivalencia de los elementos de D y D'. Se buscaba clasificar los elementos de D y D' por sus comportamientos homogéneos; por supuesto, no se verificó si se cumplía o no la condición de equivalencia para cada contexto del corpus sino que los argumentos fueron de tipo gramatical[4]. Estamos seguros de que la partición de equivalencia funciona para un gran porcentaje de contextos reales de la len-

[4] El equipo de lingüistas hizo una inspección empírica de las relaciones gramaticales que podrían darse entre los elementos de ambos conjuntos. Esta inspección consistió, por una parte, en la revisión de los materiales que aportan las gramáticas usuales del español y, por la otra, en la aplicación de pruebas de conmutación y sustitución, respecto de las cuales muchas veces se tenían que tomar decisiones probabilísticas.

gua española en México, aunque estamos conscientes de que habrá ejemplos en los que la aplicación de la relación sea errónea.

1.2.3. *La definición de la función* cg.

Si un elemento $Y \epsilon D$ pertenece a la clase I_j decimos que el indicador gramatical *(ingra)* de Y es j; en esta forma el *ingra* de cualquier elemento en D puede valer desde 1 hasta 30 y 32. Además, si asociamos a cada paradigma un número de <0> a <9> como sigue:

<0> = <ambigua>
<1> = <adverbio>
<2> = <adjetivo>
<3> = <conjunción>
<4> = <preposición>
<5> = <pronombre>
<6> = <artículo>
<7> = <contracción>
<8> = <nominal>
<9> = <verbo>

y llamamos $N = \{ [1], [2], [3], \ldots, [30], [32] \}$ y $M = \{<0>, <1>, \ldots, <9>\}$, los contextos $XYZ \epsilon C_Y$ cuando X, $Y \epsilon D$ y $Z \epsilon \psi$, pueden ser vistos como ternas ordenadas de elementos de la forma [i] [j]Z donde i es el *ingra* de X y j es el *ingra* de Y. De todo esto, la función cg puede ser vista como una función

$$cg : N \times (N \frown N \frown \psi) \rightarrow M$$

que opera ya no sobre los elementos de D sino sobre las clases de equivalencia $I_1 = [1]$, $I_2 = [2]$, . . ., $I_{30} = [30]$, $I_{32} = [32]$. Así, lo que inicialmente expresábamos como cg(cual, del cual

saqué) = <pronombre> donde *cual* ϵI_6, *del* ϵI_2 se convierte en cg([6], [2] [6]Z) = <5>; del mismo modo y ya que *ellas* ϵI_3, cg(cual, ellas cual si) = <adverbio> equivale a cg([6], [3] [6] Z') = <1>. La generalización de elementos a clases que acabo de describir es posible debido al comportamiento lingüístico homogéneo de todos los elementos de una clase dada.

1.3.*Una función* cgs *para asociar categoría gramatical mediante relaciones de precedencia.*

Consideremos ahora el siguiente hecho lingüístico: cuando se ha resuelto que cg(cual, del cual saqué) = <pronombre> podemos afirmar que la categoría gramatical de *saqué* en el contexto *del cual saqué la* es <verbo>. Asimismo cg(cual, cada cual tomó) = <pronombre> indica que la categoría de *tomó* en el contexto *cada cual tomó su* es también <verbo>. En cambio en contextos como *ellas cual Z* donde cg(cual, ellas cual Z) = <adverbio>, es mucho menos probable que la categoría del elemento Z sea una fija y determinada. De lo anterior se ve que a través de la aplicación de la función cg se puede definir otra función cgs que nos da el valor de la categoría gramatical de Z o Z' en expresiones como cg([6], [2] [6] Z) = <5> o cg([6], [3] [6] Z') = <1>, es decir, cuando cg([6], [2] [6] Z) = <5> entonces cgs (Z) = <9> = <verbo> o si cg([6], [3] [6] Z') = <1> entonces cgs(Z) = <0> = <ambigua>.

Hasta aquí, la función cgs está definida si y sólo si el contexto $XYZ\epsilon C_Y$ cumple con que $X,Y\epsilon D$. Para extender la función cgs a todo C se hizo lo siguiente: se definieron dos clases I_0 e I_{31}; a la clase I_0 se hicieron pertenecer todas las ocurrencias $Z\epsilon\psi$ para las cuales se obtuviera que

cgs(Z) = <0> = <ambigua> y a la clase I_{31} todas aquellas para las que cgs(Z) = <8> = <nominal>; también se fueron añadiendo a la clase I_{30} aquellas Z tales que cgs(Z) = <9> = <verbo>. El hacer pertenecer una ocurrencia $Z \epsilon \psi$-D a la clase I_0, I_{31} o el añadirla a la clase I_{30}, es un proceso temporal, es decir, el elemento Z no se incluye en forma definitiva en la clase, pues cgs(Z) puede tomar valores diferentes según el contexto que se analice. Esto da lugar a la posible separación de homógrafos para una $Z \epsilon \psi$-D en dos contextos diferentes. Lo que interesa en el proceso de asignación temporal de clase de equivalencia es utilizar el valor adecuado del *ingra* –para seguir manejando las funciones cg y cgs durante la asociación automática lineal– de izquierda a derecha de categorías gramaticales a todas las ocurrencias del corpus.

Con todo lo anterior se podían admitir contextos $XYZ \epsilon C_Y$ con $Y \epsilon D$ y $X \epsilon \psi$ arbitrario (nótese que I_{27} = { , }, e I_{28} = {todos los símbolos de puntuación} – I_{27}), para los que cgs quedara definida. Sin embargo, si $Y \epsilon \psi$-D, dado un contexto RXYZ, cgs(Y) es muchas veces <ambigua>, así que la información que cg(Y,XYZ) nos da para cgs(Z) es generalmente poca y obliga a definir cgs(Z) = <ambigua> toda vez que $Z \epsilon \psi$-D. Fue aquí donde se decidió hacer uso de las posibilidades de reconocimiento morfológico de vocablos españoles, para disminuir el porcentaje de palabras ambiguas.

El algoritmo de reconocimiento morfológico, así como su utilización asociada a las funciones de precedencia cg y cgs serán descritos posteriormente. La definición completa de las funciones cg y cgs se dará al hacer la descripción computacional del sistema. Antes, describiré algunas ex-

tensiones de dichas funciones de precedencia que también se aplican para incrementar el rendimiento del análisis.

1.4. *Extensiones de la función de precedencia.*

1.4.1. *La función* ig *para guardar memoria de algunas relaciones de precedencia.*

Consideremos el ejemplo ya utilizado:

"y del cual saqué la aclaración"

Habíamos dicho que el hecho de que cg(cual, del cual saqué) fuera igual a <pronombre>, permitía definir cgs(saqué) igual a <verbo>; sin embargo, casos como:

"y del cual no saqué la aclaración" o
"y del cual nunca saqué la aclaración" o
"y del cual nunca más saqué la aclaración" o
"y del cual tal vez saqué la aclaración", etc.,

son muy frecuentes; ahora bien, cuando se ha logrado reconocer una ocurrencia de *cual* como <pronombre> se tiene la certeza de que un <verbo> ocurrirá tarde o temprano a la derecha en el corpus; es decir, la información obtenida de que cg(cual, del cual Z) es un <pronombre> debe ser conservada si Z está en D (y no es <verbo>), pues en este caso Z no es el <verbo> esperado. Debemos por lo tanto extender el manejo de cg como sigue: si cg(cual, del cual $Z_1 Z_2 \ldots Z_n$) = <pronombre>, entonces cgs(Z_i) = <verbo> para la mínima i tal que $Z_i \notin$ D. Formalmente, y transcribiendo al lenguaje de las clases de equivalencia para hacer la apropiada generalización, vemos que esta extensión de cg puede definirse a través de una nueva función

$$ig : N \times (N \frown N \frown \psi) \rightarrow N$$

que se aplique antes que cg y obligue a no modificar la clase I_e a la que se haya hecho pertenecer la ocurrencia de Y en el contexto XYZ_1. Así, si i es tal que $Z_1, Z_2, \ldots, Z_{i-1} \in D$ pero $Z_i \notin D$ y si $Z_j \in I_{kj}$ y $\langle p_{kj} \rangle$ es la categoría gramatical de Z_j en el contexto $Z_{j-1}\ Z_j Z_{j+1}$ para j = 1, . . . , i-l, entonces

$cg([e], [f]\ [e]\ Z_1) = \langle$pronombre$\rangle$
y $ig([e], [f]\ [e]\ Z_1) = [e]$
$cg([k_1], [e]\ [k_1]\ Z_2) = \langle p_{k1} \rangle$
y $ig([k_1], [i]\ [k_1]\ Z_2) = [e]$
$cg([k_2], [k_1]\ [k_2]\ Z_3) = \langle p_{k2} \rangle$
y $ig([k_2], [k_1]\ [k_2]\ Z_3) = [e]\ldots$
$cg([k_{i-1}], [k_{i-2}]\ [k_{i-1}]\ Z_i) = \langle p_{k\,i-1} \rangle$
y $ig([k_{i-1}], [k_{i-2}]\ [k_{i-1}]\ Z_i) = [e]$

es la secuencia de aplicaciones de las funciones cg e ig durante el rastreo de los elementos Z_1, Z_2, . . ., Z_i; en la última etapa, ya que $Z_i \notin D$, el valor [e] de ig ($[k_{i-1}]$, $[K_{i-2}]$ $[k_{i-1}]$ Z_i) que recordaba la existencia del <pronombre> a la izquierda, determina que cgs (Z_i) sea igual a <verbo>, es decir, permite definir $cgs(Z_i) = \langle 9 \rangle$, para el ejemplo en cuestión. Una vez más la definición completa de la función ig se pospone hasta la descripción computacional.

1.4.2. *La función* **pg** *de postcedencia.*

Consideremos ahora el siguiente ejemplo:

"cuando una cantaba, la otra"

Los vocablos *cuando* y *una* pertenecen a D

pero *cantaba* $\notin$ D; *cuando* funciona con una categoría gramatical fija que es <1> = <adverbio> y además *cuando* ϵI_1. *Una* puede funcionar como <pronombre> = <5> o como <artículo> = <6> y *una* ϵI_{19}. Sin embargo, la función cg(una, cuando una cantaba) no puede tener el valor fijo de <pronombre> pues existen contextos como

"cuando una puerta se abre"

caso en el que *una* funciona como <artículo>. Ahora bien, suponiendo que ya somos capaces de manejar el algoritmo morfológico, surge la posibilidad de fijar la categoría gramatical de *una* a través de la información que tengamos sobre la categoría gramatical de *cantaba.* Esto es, cg(una, cuando una cantaba) no está definida y por lo tanto cgs(cantaba) tampoco lo está, pero si morfológicamente reconocemos que cantaba es <verbo>, deducimos inmediatamente que *una* es <pronombre>. Denotemos como am(cantaba) al resultado de aplicar el algoritmo morfológico a *cantaba;* entonces, cuando am(cantaba) = <verbo>, la ocurrencia de *cantaba* en el contexto *cuando una cantaba* se hace pertenecer a la clase I_{30}; así cg(una, cuando una cantaba) se puede definir a través de una nueva función pg del siguiente modo: cg(una, cuando una cantaba) = pg([19], [19] [30]) = <pronombre>. Generalizando, sea

$$pg : N \times (N \frown N) \rightarrow M$$

y sea XYZ un contexto tal que cg(Y,XYZ) no está definida; supongamos que $Y \epsilon$ D, $Y \epsilon I_i$ y $Z \epsilon \psi$ es tal que su categoría gramatical ha sido

determinada mediante la función am y la ocurrencia de Z en el contexto YZ se ha hecho pertenecer a la clase I_j, entonces pg([i], [i] [j]) puede ser definida. Una vez definida la función pg podemos hacer cg(Y,XYZ) = pg([i], [i] [j]). La definición de la función pg se dará durante la descripción computacional.

1.5 *Algoritmo morfológico.*

Pasaré ahora a describir lo que hemos llamado algoritmo morfológico. El reconocimiento de la categoría gramatical de un vocablo a través de la comparación de sus caracteres terminales con los morfemas de tiempo, número, persona y modo de los verbos españoles, así como con algunas terminaciones numerales o adverbiales, resultaba ser de mucha eficacia; sin embargo, el proceso de comparación debía ser hecho para todos los vocablos $X \epsilon \psi$-D para los cuales la función cgs(X) fuera <ambigua>, y además, cuando cgs(X) = <9> = <verbo>, en cuyo caso el reconocimiento de la terminación se debía hacer, ya no para identificar la categoría gramatical de X, sino para identificar la raíz respectiva del <verbo> y poder así agrupar con ella todas las ocurrencias del <verbo>, independientemente de su diversidad de forma. Una vez que los lingüistas concluyeron la inspección empírica necesaria para dar la lista de las terminaciones, se procedió a buscar una estructura computacional que permitiera realizar la comparación de los caracteres terminales de los vocablos de la muestra con la lista de terminaciones.

Con el fin de proporcionar al lector un punto de vista sobre la eficiencia de la estructura resultante, tomaremos una pequeña lista de termina-

ciones y construiremos la estructura asociada. Sea, entonces,

LT = [–ad, –are, –mente, –ira/n, –endo, –cho, –so, –to, –er, –se/is, –o/] [5]

y supongamos que queremos comparar los caracteres terminales del vocablo *fuertemente* con las terminaciones en LT. Naturalmente, la comparación se efectúa de derecha a izquierda. LLamemos u a la función que asocia a cada vocablo o cadena de caracteres concatenados, el último carácter de la concatenación, así por ejemplo, u(fuertemente) = e; y r a la función que asocia a cada vocablo todos excepto el último carácter, por ejemplo r(fuertemente) = fuertement. Comparamos u(fuertemente) = e con u(–ad) = d y encontramos que e ≠ d; es suficiente esta comparación para concluir que *fuertemente* no tiene la terminación *–ad.* Pasamos por lo tanto al siguiente elemento de LT: u(fuertemente) = e es igual a u(–are) = e, por lo cual pasamos a comparar los caracteres que siguen hacia la izquierda, tanto en el vocablo analizado, como en la terminación elegida y tenemos que $u(r$(fuertemente)) = u (fuertement) = t ≠ r = u (–ar) = $u(r$(–are)) lo que indica que *fuertemente* tampoco tiene la terminación *–are.*

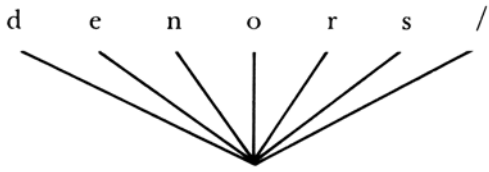
Vemos hasta aquí, que el algoritmo de reconocimiento morfológico debe incluir las siguientes reglas: si la comparación por igualdad de caracteres correspondientes en la terminación y el vocablo resulta falsa, debe tomarse la siguiente

5 La importancia del acento para el proceso de reconocimiento y la imposibilidad de manejar caracteres acentuados en nuestro equipo de cómputo nos obligó a codificar el acento como un carácter '/' inmediatamente después de la vocal acentuada, por ejemplo: número =nu/mero e –irán = –ira/n.

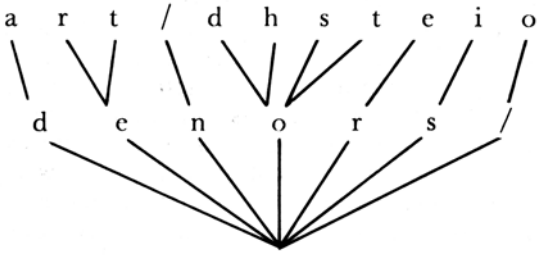
terminación, pero si los caracteres correspondientes son iguales, deben compararse los caracteres correspondientes que siguen a la izquierda.

Volviendo a nuestro ejemplo, *fuertemente* no termina en *–are,* así que hay que tomar la siguiente terminación en LT, que es *–mente;* de acuerdo con el procedimiento que acabamos de describir, debíamos analizar nuevamente si u(fuertemente) = e es igual o no a u(–mente) = e, pero podríamos ahorrarnos esta comparación, ya que en el paso anterior se había encontrado como cierto que u(fuertemente) = e era igual a u(–are) = e. Efectivamente, basta comparar $u(r$(fuertemente)) = t con $u(r$(–mente)) = t para seguir adelante, siempre y cuando las terminaciones en LT tengan un orden adecuado. Entonces, como $u(r$(fuertemente)) = $u(r$(–mente)) continuamos comparando $u(r(r$(fuertemente))) = n con $u(r(r$(–mente))) = n lo que nos da nuevamente el valor cierto, al igual que $u(r(r(r$(fuertemente)))) = e = $u(r(r(r$(–mente)))) y $u(r(r(r(r-$(fuertemente))))) = m = $u(r(r(r(r$-mente))))); aquí, la marca '–' en la terminación –mente nos indica que hemos terminado el proceso comparativo y que *fuertemente* tiene la terminación *–mente.*

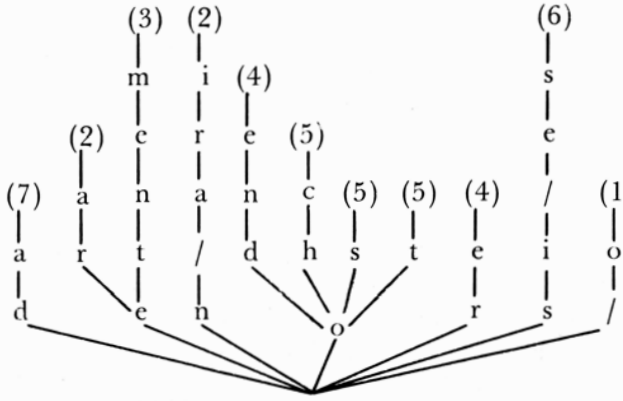
El ordenamiento de los elementos en LT, para hacer el uso adecuado de las reglas de comparación que acabamos de describir, debió ser el de una estructura arborescente. Esta estructura es como sigue: de la raíz del árbol emergen tantas ramas como caracteres diferentes pueda haber en el lugar de más a la derecha de cada terminación; las ramas se ordenan correspondiendo a un ordenamiento alfabético de los caracteres seleccionados:



en esta figura, la rama con la etiqueta 'd' corresponde a la terminación *–ad,* pero la rama etiquetada con 'e' corresponde tanto a *–are* como a *–mente* y la 'o' a *–endo, –cho, –so* y *–to*, y así para todas las ramas y terminaciones. De la misma manera, tomando todas las terminaciones que corresponden a cada clase (todas las que pertenecen a una misma rama en la figura anterior), construimos subarborescencias en cada nodo del primer nivel y obtenemos un segundo nivel en el árbol:



Siguiendo este proceso llegará el momento en que todas las terminaciones de LT estén representadas en la arborescencia mediante una trayectoria desde la raíz hasta un nodo terminal:



La arborescencia es utilizada de la siguiente manera: se inicia el rastreo desde la raíz del árbol y se sigue la trayectoria que une la raíz con el nodo terminal de más a la izquierda. El cambio de un nivel dado en el árbol hacia un nivel más alto corresponde a pasar al carácter inmediato a la izquierda en el vocablo analizado; notará el lector que el algoritmo no permite bajar a un nivel inferior en la arborescencia, es decir, nunca se analizan de nuevo los caracteres del vocablo que ya han sido rastreados. La trayectoria es seguida hacia arriba siempre que los caracteres correspondientes del vocablo y la arborescencia sean iguales; cuando los caracteres correspondientes no son iguales, se cambia de trayectoria tomando la rama inmediata a la derecha que emerge del mismo nodo que la que se estaba utilizando, y nuevamente se va comparando hacia las hojas. Eventualmente sucederá que la rama que se acaba de analizar es la de más a la derecha de un nodo fijo y que los caracteres correspondientes son desiguales; el procedimiento de com-

paración acaba aquí y se concluye que el vocablo analizado no tiene ninguna de las terminaciones de LT, en cuyo caso se le asocia una categoría gramatical <ambigua>. Cuando el vocablo tiene alguna de las terminaciones previstas en LT, durante el rastreo se alcanza el nodo terminal correspondiente. Estos nodos terminales están etiquetados para que pueda efectuarse la operación adecuada según la terminación reconocida. En la arborescencia que hemos usado como ejemplo, las etiquetas de la (1) a la (7) corresponden a:

(1) La terminación reconocida es verbal (*cantó, comió, oyó*), pero excepcionalmente puede haber un nominal con la misma terminación (*chacó, rococó*). El operador (1) inicia una comparación con la lista de excepciones (que se anexa posteriormente). Si el vocablo X coincide con alguna excepción, se marca con la categoría que se asignó previamente a la excepción; generalmente es <nominal> pero en algunos casos es <ambigua> como en *sed;* además, la ocurrencia del vocablo X se hace pertenecer a la clase I_{31} o I_0 (para <nominal> o <ambigua> respectivamente) para que se pueda seguir el proceso de aplicación de las funciones de precedencia descritas antes. Si, en cambio, X no es ninguna de las excepciones, el resultado del algoritmo morfológico am(X) es <verbo> y X se hace pertenecer a la clase I_{30}.

(2) am(X) = <verbo> y la ocurrencia de X en el contexto analizado se incluye en la clase I_{30}.

(3) am(X) = <adverbio> y X se incluye definitivamente en la clase I_1.

(4) am(X) = <ambigua> y la ocurrencia de X en el contexto analizado se hace pertenecer a la clase I_0.

(5) La terminación puede ser de <verbo>, de <nominal> o <ambigua>, sin embargo, una decisión puede obtenerse a través de la aplicación de la función tv que describo a continuación:

Sea $L = \{ t \mid t \in LT$ y su nodo terminal asociado está etiquetado con (5)$\}$

y sea $\psi_L = \{X \mid X = yt$ con $t \in L$ y $X \epsilon \psi\}$ entonces

$$tv : L \times N * L \rightarrow M$$

es tal que tv(t,[j]t) asocia al vocablo $X \epsilon \psi_L$ una categoría gramatical que depende de la clase I_j a la que pertenezca el vocablo que precede a X en el corpus. La definición de la función tv, como en el caso de las funciones de precedencia, fue el resultado de un análisis empírico de la lengua natural. Consideremos los ejemplos:

"contra vien*to* y marea" y
"ampliamente escri*to*"

viento y *escrito* están en ψ_L, *contra* ϵI_{20}, además después de aplicar el algoritmo de reconocimiento morfológico a *ampliamente* se obtiene que am(ampli*amente*) = <adverbio> y según el caso (2) anterior *ampliamente* ϵI_1; entonces tv(to,[20]to) = <nominal> en tanto que tv(to, [1]to) = <verbo>.

(6) am(X) = <numeral> y X se incluye definitivamente en I_{21}.

(7) am(X) = <verbo> excepto si está precedido de algunas palabras específicas. Como en (1) se inicia la comparación del vocablo que precede a X con la lista de excepciones. Cada excep-

ción tiene en este caso asociada la categoría gramatical que debe ser asignada a X.

2. *El sistema computacional.*

A continuación describiré la implementación computacional del sistema:

2.1. *Almacenamiento de los datos.*

El CEMC quedó almacenado en un archivo llamado CORPUS, en donde cada línea tiene una clave de 9 caracteres (tttpppllll) según el número del texto en cuestión (de 000 a 999), la página y la línea correspondientes a la publicación original[6]. Todas las líneas de todos los textos fueron ordenadas de acuerdo con dicha clave en forma creciente dentro de cada género y siguiendo este ordenamiento se les asignó un número secuencial de sólo 6 caracteres. La ilustración 1 es una muestra del CORPUS. La primera clave (entre la primera y la segunda marcas de control representadas por @) es la clave secuencial, y la segunda (entre la segunda y la tercera marcas de control) es la de texto, página y línea. Después de la tercera marca de control puede verse el contenido del texto; la '/' es el acento, el '+' corresponde a la 'ñ', la marca'@'al final de una palabra indica que la palabra es un nombre propio, en tanto que el 'Δ' indica que la palabra estaba escrita con mayúscula en el texto original.

2.2. *El diccionario de máquina mínimo se almacena en los archivos DICCIONARIO y DECISIONES.*

Los elementos del conjunto D definido al inicio de este artículo, se almacenaron en un archi-

[6] Véase "Base estadística. . .", § 3.3.2.

Ilustración 1

CORPUS

@003812 @017013010 @LAS MERAS DEFUNCIONES RESULTARON YA INSUFICIENTES. EL MINISTRO△DE
@003813 @017013011 @SALUDA PU/BLICA△SE SINTIO/ SINCERO, Y UNA NOCHE CALIGINOSA, CON LA LUZ
@003814 @017013012 @APAGADA, DESPUE/S DE ACARICIARLE UN RATITO EL PECHO COMO POR NO DEJAR,
@003815 @017013013 @LE CONFESO/ A SU MUJER QUE SE CONSIDERABA INCAPAZ DE ELEVAR LA
@003816 @017013014 @MORTALIDAD A UN NIVEL GRATO A LOS INTERESES DE LA COMPA+I/A△ A LO QUE
@003817 @017013015 @ELLA LE CONTESTO/ QUE NO SE PREOCUPARA, QUE YA VERI/A CO/MO TODO IBA A
@003818 @017013016 @SALIR BIEN, Y QUE MEJOR SE DURMIERAN.
@003819 @017014004 @DE ACUERDO CON ESA MEMORABLE LEGISLACIO/N, A LOS ENFERMOS GRAVES SE
@003820 @017014005 @LES CONCEDI/AN VEINTICUATRO HORAS PARA PONER EN ORDEN SUS PAPELES Y
@003821 @017014006 @MORIRSE; PERO SI EN ESTE TIEMPO TENI/AN SUERTE Y LOGRABAN CANTAGIAR A
@003822 @017014007 @LA FAMILIA, OBTENI/AN TANTOS PLAZOS DE UN MES COMO PARIENTES FUERAN
@003823 @017014008 @CONTAMINADOS. LAS VI/CTIMAS DE ENFERMEDADES LEVES Y LOS SIMPLEMENTE
@003824 @017014009 @INDISPUESTOS MERECI/AN EL DESPRECIO DE LA PATRIA Y, EN LA CALLE,
@003825 @017014010 @CUALQUIERA PODI/A ESCUPIRLES EL ROSTRO. POR PRIMERA VEZ EN LA HISTORIA
@003826 @017014011 @FUE RECONOCIDA LA IMPORTANCIA DE LOS ME/DICOS (HUBO VARIOS CANDIDATOS
@003827 @017014012 @AL PREMIO NOBEL @) QUE NO CURABAN A NADIE. FALLECER SE CONVIRTIO/ EN
@003828 @017014013 @EJEMPLO DEL MA/S EXALTADO PATRIOTISMO, NO SO/LO EN EL ORDEN NACIONAL,
@003829 @017014014 @SINO EN EL MA/S GLORIOSO, EN EL CONTINENTAL

vo llamado DICCIONARIO; cada elemento quedó asociado a dos códigos: *catgra* de categoría gramatical e *ingra* de clase de equivalencia; estos códigos corresponden a los que se han manejado durante el desarrollo de este trabajo. En otro archivo llamado DECISIONES se almacenaron todos los elementos del conjunto D', que como recordará el lector es el subconjunto de D que contiene aquellos vocablos que pueden funcionar con dos categorías gramaticales, pero que, dado un contexto, se puede decidir con qué categoría están siendo usados a través de la aplicación de las funciones antes descritas. Cada elemento de DECISIONES quedó asociado también a sus códigos *catgra* e *ingra,* pero el valor de *catgra* corresponde a la otra categoría con que la palabra puede ser usada; es decir, las palabras de D' quedan en ambos archivos pero con diferentes códigos *catgra.* En la ilustración 2 se enlistan todas las palabras de los conjuntos D (DICCIONARIO) y D' (DECISIONES); cada palabra está asociada con un número secuencial a la izquierda; el primer dígito de la derecha corresponde al código *catgra* y puede valer, como había dicho antes, desde 0 hasta 9; el segundo y tercer dígitos corresponden al código *ingra* que puede tomar valores desde 0 hasta 32.

2.3. *Las definiciones de las funciones* **cg** *y* **cgs** *se almacenan en la TABLA de PRECEDENCIA.*

Recordará el lector que la función

$$cg: N \times (N \frown N \frown \psi) \rightarrow M$$

permite asociar una categoría gramatical (un elemento de M) a cada $Y \in \psi$ en un contexto dado $XYZ \in C$, a través de las clases de equiva-

Ilustración 2

DICCIONARIO

1:ABAJO	11	39:ENDENANTES	11
2:ACA/	11	40:ENTONCES	11
3:ACASO	11	41:HOY	11
4:ADEMA/S	11	42:JAMA/S	11
5:AFUERA	11	43:LEJOS	11
6:AHI/	11	44:MIENTRAS	11
7:AHORA	11	45:MUY	11
8:AHORITA	11	46:NOMA/S	11
9:ALLA/	11	47:NUNCA	11
10:ALLI/	11	48:ORITA	11
11:ANOCHE	11	49:PRONTO	11
12:ANTENOCHE	11	50:QUIZA/	11
13:ANTES	11	51:QUIZA/S	11
14:ANTEAYER	11	52:SIEMPRE	11
15;ANTIER	11	53:TAMBIE/N	11
16:APENAS	11	54:TAMPOCO	11
17:APRISA	11	55:TAN	11
18:AQUI/	11	56:TEMPRANO	11
19:ARRIBA	11	57:TODAVI/A	11
20:ASI/	11	58:YA	11
21:ATRA/S	11	59:AUNQUE	31
22:AU/N	11	60:CONQUE	31
23:AYER	11	61:EMPERO	31
24:CASI	11	62:MAS	31
25:CO/MO	11	63:NI	31
26:CUANDO	11	64:O	31
27:CUA/NDO	11	65:U	31
28:DEBAJO	11	66:PERO	31
29:DELANTE	11	67:PORQUE	31
30:DENTRO	11	68:SI	31
31:DEPRISA	11	69:Y	31
32:DESPACIO	11	70:E	31
33;DESPUE/S	11	71:CONMIGO	51
34:DETRA/S	11	72:CONTIGO	51
35:DONDE	11	73:MI/	51
36:DO/NDE	11	74:TI	51
37:DONDEQUIERA	11	75:ALLENDE	41
38:ENANTES	11	76:AQUENDE	41

77:CON 41
78:DESDE 41
79:DURANTE 41
80:MEDIANTE 41
81:HASTA 41
82:PA 41
83:SEGU/N 41
84:SIN 41
85:SO 41
86:TRAS 41
87:ALGU/N 22
88:AQUEL 22
89:AQUELLA 22
90:AQUELLOS 22
91:AQUELLAS 22
92:CADA 22
93:CUALQUIER 22
94:CUYA 22
95:CUYAS 22
96:MI 22
97:MIS 22
98:NINGU/N 22
99:SENDOS 22
100:SU 22
101:SUS 22
102:TU 22
103:TUS 22
104:UN 62
105:AL 72
106:DEL 72
107:AQUE/L 53
108:AQUE/LLA 53
109:AQUE/LLAS 53
110:AQUE/LLOS 53
111:AQUELLO 53
112:E/SA 53
113:E/SAS 53
114:E/SE 53
115:ESO 53
116:E/SOS 53
117:E/STA 53
118:E/STE 53
119:ESTO 53
120:E/STAS 53
121:E/STOS 53
122:YO 53
123:TU/ 53
124:E/L 53
125:ELLA 53
126:ELLO 53
127:ELLAS 53
128:ELLOS 53
129:NOSOTROS 53
130:NOSOTRAS 53
131:VOSOTROS 53
132:VOSOTRAS 53
133:USTED 53
134:USTEDES 53
135:ALGUIEN 53
136:ALGUNO 53
137:QUIEN 53
138:QUIE/N 53
139:QUIENES 53
140:QUIE/NES 53
141:LE 53
142:LES 53
143:CUALQUIERA 53
144:CUALESQUIERA 53
145:NINGUNO 53
146:NADIE 53
147:ME 53
148:TE 53
149:SE 53
150:NOS 53
151:OS 53
152:LA 64
153:LO 64
154:LOS 64
155:LAS 64
156:EL 65
157:CUAL 56
158:CUALES 56

159:MI/O 532
160:MI/A 532
161:MI/OS 532
162:MI/AS 532
163:TUYO 532
164:TUYA 532
165:TUYOS 532
166:TUYAS 532
167:SUYO 532
168:SUYA 532
169:SUYOS 532
170:SUYAS 532
171:NUESTRO 532
172:NUESTRA 532
173:NUESTROS 532
174:NUESTRAS 532
175:VUESTRO 532
176:VUESTRA 532
177:VUESTROS 532
178:VUESTRAS 532
179:NO 17
180:ESA 28
181:ESE 28
182:ESAS 28
183:ESOS 28
184:ESTA 28
185:ESTAS 28
186:ESTE 28
187:ESTOS 28
188:CUYO 89
189:BIEN 810
190:MAL 810
191:MA+ANA 810
192:HORA 811
193:HORITA 811
194:QUE/ 212
195:A 413
196:ANTE 413
197:DE 413
198:EN 413
199:HACIA 413
200:POR 413
201:SI/ 514
202:LUEGO 315
203:PUES 316
204:SINO 817
205:CONFORME 218
206:UNA 619
207:CONTRA 820
208:ALGUNA 521
209:ALGUNAS 521
210:ALGUNOS 521
211:AMBOS 521
212:AMBAS 521
213:BASTANTES 521
214:CUANTA 521
215:CUANTAS 521
216:CUANTOS 521
217:CUA/NTA 521
218:CUA/NTAS 521
219:CUA/NTOS 521
220:CUA/L 521
221:CUA/LES 521
222:DEMASIADA 521
223:DEMASIADOS 521
224:DEMASIADAS 521
225:DEMA/S 521
226:MUCHOS 521
227:MUCHAS 521
228:MUCHA 521
229:MUCHI/SIMOS 521
230:MUCHI/SIMAS 521
231:MUCHI/SIMA 521
232:NINGUNA 521
233:NINGUNAS 521
234:NINGUNOS 521
235:OTRO 521
236:OTRA 521
237:OTROS 521
238:OTRAS 521
239:POCOS 521
240:POCAS 521

241:POCA	521	282:TAL	025
242:POQUI/SIMOS	521	283:MEDIO	025
243:POQUI/SIMAS	521	284:ALGO	025
244:POQUI/SIMA	521	285:NADA	025
245:VARIOS	521	286:CUYOS	025
246:VARIAS	521	287:QUE	026
247:TODOS	521	288:HABER	929
248:TODAS	521	289:H-	929
249:TODA	521	290:HAB-	929
250:UNO	622	291:HAY;	929
251:UNOS	622	292:HUB-	929
252:UNAS	622	293:HABR-	929
253:BASTANTE	223	294:ESTAR	929
254:CUANTO	223	295:ESTOY	929
255:CUA/NTO	223	296:ESTA/S	929
256:DEMASIADO	223	297:ESTA/	929
257:MA/S	223	298:ESTA/N	929
258:MENOS	223	299:ESTE/S	929
259:MEJOR	223	300:ESTUV-	929
260:MUCHO	223	301:EST-	929
261:MUCHI/SIMO	223	302:ESTE/N	929
262:PEOR	223	303:TENER	929
263:POCO	223	304:TEN-	929
264:POQUI/SIMO	223	305:TIEN-	929
265:TANTO	223	306:TENG-	929
266:TODO	223	307:TENDR-	929
267:COMO	924	308:TUV-	929
268:FUERA	924	309:SER	025
269:FUERAS	924	310:ES	929
270:ORA	924	311:ER-	929
271:ADELANTE	924	312:SO-	929
272:ADENTRO	924	313:FUI	929
273:ENCIMA	924	314:FUE	929
274:ENFRENTE	924	315:SER-	929
275:CABE	924	316:S-	929
276:PARA	924	317:SOY	929
277:SOBRE	025	318:SOIS	929
278:BAJO	025	319:SON	929
279:ENTRE	025	320:SOMOS	929
280:CERCA	025	321:LLEVAR	929
281:CLARO	025	322:LLEV-	929

323:DEJAR	929	342:VEN-	929
324:DEJ-	929	343:SEGUIR	929
325:QUEDAR	929	344:SIG-	929
326:QUED-	929	345:SEGU-	929
327:IR	929	346:SIGU-	929
328:VOY	929	347:ANDAR	929
329:V-	929	348:AND-	929
330:VAY-	929	349:ANDUV-	929
331:ID	929	350:CONTINUAR	929
332:IB-	929	351:CONTINU-	929
333:I/B-	929	352:DOY	930
334:YENDO	929	353:DI/	930
335:FU-	929	354:HAZ	930
336:IR-	929	355:OI/MOS	930
337:VENIR	929	356:OI/D	930
338:VEN	929	357:OI/STE	930
339:VIEN-	929	358:OI/STES	930
340:VENDR-	929	359:CUAN	01
341:VENG-	929	360:SEIS	221

DECISIONES *

1:LA	54	19:NUESTRO	232
2:LO	54	20:NUESTRA	232
3:LOS	54	21:NUESTROS	232
4:LAS	54	22:NUESTRAS	232
5:CUAL	16	23:VUESTRO	232
6:CUALES	16	24:VUESTRA	232
7:MI/O	232	25:VUESTROS	232
8:MI/A	232	26:VUESTRAS	232
9:MI/OS	232	27:CUYO	29
10:MI/AS	232	28:BIEN	110
11:TUYO	232	29:MAL	110
12:TUYA	232	30:MA+ANA	110
13:TUYOS	232	31:HORA	111
14:TUYAS	232	32:HORITA	111
15:SUYO	232	33:SI/	114
16:SUYA	232	34:LUEGO	115
17:SUYOS	232	35:PUES	116
18:SUYAS	232	36:SINO	317

37:CONFORME	318
38:UNA	519
39:CONTRA	420
40:ALGUNA	221
41:ALGUNAS	221
42:ALGUNOS	221
43:AMBOS	221
44:AMBAS	221
45:BASTANTES	221
46:CUANTA	221
47CUANTAS	221
48:CUANTOS	221
49:CUA/NTAS	221
50:CUA/NTA	221
51:CUA/NTOS	221
52:CUA/L	221
53:CUA/LES	221
54:DEMASIADA	221
55:DEMASIADOS	221
56:DEMA/S	221
57:MUCHOS	221
58:MUCHAS	221
59:MUCHA	221
60:MUCHI/SIMOS	221
61:MUCHI/SIMAS	221
62:MUCHI/SIMA	221
63:NINGUNA	221
64:NINGUNAS	221
65:NINGUNOS	221
66:OTRO	221
67:OTRA	221
68:OTROS	221
69:OTRAS	221
70:POCOS	221
71:POCAS	221
72:POCA	221
73;POQUI/SIMOS	221
74:POQUI/SIMAS	221
75:POQUI/SIMA	221
76:VARIOS	221
77:VARIAS	221
78:TODOS	221
79:TODAS	221
80:TODA	221
81:UNO	522
82:UNOS	522
83:UNAS	522
84:COMO	124
85:FUERA	124
86:FUERAS	124
87:ORA	124
88:ADELANTE	124
89:ADENTRO	124
90:ENCIMA	124
91:ENFRENTE	124
92:CABE	124
93:PARA	124

lencia a las que pertenecen X y Y; es decir, si $X \epsilon I_i$ y $Y \epsilon I_j$ entonces:

$$cg(Y, XYZ) = cg([j], [i] [j]Z)$$

Por otra parte, en el momento de decidir la categoría gramatical de X en el contexto XYZ, se puede decidir la categoría gramatical de Z en ese mismo contexto, es decir, se puede decidir el valor de cgs(Z). Estas dos funciones se almacenaron en la computadora en forma de una tabla bidimensional llamada PRECEDENCIA, que contiene 33 renglones y 33 columnas correspondientes a los 33 posibles valores (0, . . .,32) de las clases de equivalencia. Para aclarar cómo es utilizada esta tabla supongamos que estamos analizando la expresión:

". . .tirando de ellas
cual si fuesen toros bravos. . ."

El lector puede hacer uso de la siguiente ilustración 3, así como de la tabla de PRECEDENCIA que se anexa a continuación (ilustración 4), para seguir este análisis. El archivo DICCIONARIO tiene almacenadas previamente las palabras *de, ellas, cual* y *si,* con sus códigos *catgra* e *ingra* como en la ilustración, y el archivo DECISIONES almacena la palabra *cual* con 1 para *catgra* y 6 para *ingra.* En el momento en que la palabra *ellas* es leída, se busca en el archivo DICCIONARIO y se obtiene el valor 3 de su *ingra.* El valor del *ingra* de la palabra *de* fue 13; entonces se busca en el décimo tercer renglón y la tercera columna de PRECEDENCIA y se obtiene una pareja de valo-

Ilustración 3

DICCIONARIO				DECISIONES			
	Catgra	Ingra	Frec		Catgra	Ingra	Frec
de	4	13		cual	1	6	+1
ellas	5	3	+1				
cual	5	6					
si	3	1	+1				
*fuesen	9	30	1				

. . . tirando de ellas cual si fuesen toros bravos. . .
5 1 3 9

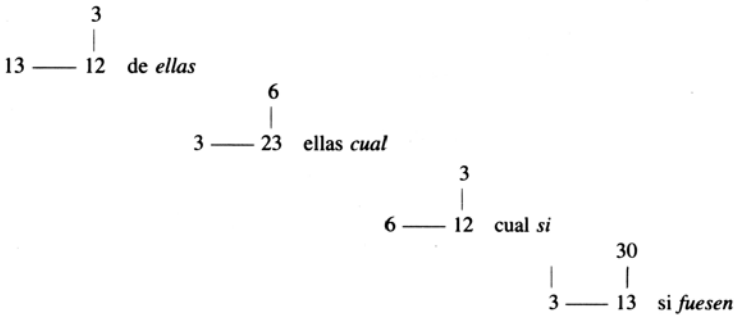


Ilustración 4

TABLA DE PRECEDENCIA

	0.	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.	21.	22.	23.	24.	25.	26.	27.	28.	29.	30.	31.	32.
0.	33	13	11	12	43	11	23	13	11	21	23	23	43	13	23	23	13	23	23	43	23	43	43	43	43	33	33	3	3	13	13	33	23
1.	33	13	11	12	43	11	23	13	11	21	23	23	43	13	23	23	13	23	23	43	23	43	43	43	43	33	33	3	3	13	13	33	23
2.	33	13	11	12	43	11	12	13	11	13	13	13	43	13	23	23	13	13	23	43	13	43	43	43	43	33	33	3	3	13	13	33	12
3.	33	13	11	12	43	11	23	13	11	21	23	23	43	13	23	23	13	23	23	43	23	43	43	43	13	33	33	3	3	13	13	33	23
4.	33	13	11	12	43	11	12	13	11	21	23	13	43	13	23	23	13	23	23	43	13	43	43	43	43	33	33	3	3	13	13	33	12
5.	33	13	11	12	43	11	12	13	11	13	13	23	43	13	23	23	13	23	23	43	23	43	43	43	43	33	33	3	3	13	13	33	12
6.	33	13	11	12	43	11	23	13	11	21	23	13	43	13	23	23	13	23	23	43	23	43	43	43	43	33	33	3	3	13	13	33	23
7.	33	13	11	12	43	11	23	13	11	21	23	23	43	13	23	23	13	23	23	43	23	43	43	43	43	33	33	3	3	13	13	33	23
8.	33	13	11	12	43	11	23	13	11	13	13	13	43	13	23	23	13	23	23	43	13	43	43	43	43	33	33	3	3	13	13	33	23
9.	33	13	11	12	43	11	23	13	11	13	13	23	43	13	23	23	13	23	23	43	23	43	43	43	43	33	33	3	3	13	13	33	23
10.	33	13	11	12	43	11	23	13	11	21	23	23	43	13	23	23	13	23	23	43	23	43	43	43	43	33	33	3	3	13	13	33	23
11.	33	13	11	12	43	11	23	13	11	21	23	23	43	13	23	23	13	23	23	43	23	43	43	43	43	33	33	3	3	13	13	33	23
12.	33	13	11	12	43	11	23	13	11	21	23	13	43	13	23	23	13	23	23	43	23	43	43	43	43	33	33	3	3	13	13	33	23

13 33 13 11 12 43 11 23 13 11 21 23 23 43 13 13 23 13 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23
14. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 23 13 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23
15. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 23 13 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23
16. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 23 13 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23
17. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 23 13 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23
18. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 23 13 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23
19. 33 13 11 12 43 11 23 13 11 21 23 13 43 13 23 23 13 23 23 43 13 43 43 43 43 33 33 3 3 13 13 33 23
20. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 23 13 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23
21. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 23 13 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23
22. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 23 13 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23
23. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 23 13 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23
24. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 23 13 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23
25. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 23 13 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23
26. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 23 13 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23
27. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 13 43 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23
28. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 23 13 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23
29. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 23 13 23 13 43 23 43 43 43 43 33 33 3 3 13 13 33 23
30. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 23 13 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23
31. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 23 13 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23
32. 33 13 11 12 43 11 23 13 11 21 23 23 43 13 23 23 13 23 23 43 23 43 43 43 43 33 33 3 3 13 13 33 23

res (1,2). El primer valor *1* indica que *ellas* está siendo usada con la categoría que fue previamente almacenada en el DICCIONARIO y entonces, el contador de frecuencia para *ellas* es incrementado en este archivo. El segundo valor de la pareja *2* indica que la siguiente palabra en el corpus es o un elemento del conjunto D o un <verbo>. Note el lector que el primer valor de la pareja nos permite asociar su categoría gramatical a *ellas* en el contexto *de ellas cual;* y el segundo valor nos da la posibilidad de asociar una categoría a *cual* en ese mismo contexto (pues si no es <verbo>, entonces es un elemento de D y en la siguiente aplicación de la PRECEDENCIA se decidirá definitivamente su categoría). Podemos entonces decir que el primer valor de la pareja simula a la función cg(ellas, de ellas cual) y que el segundo simula a cgs(cual). Continuando el análisis de nuestro ejemplo, leemos la palabra *cual,* se busca en DICCIONARIO y se obtiene un 6 para su *ingra.* Una nueva pareja de valores (2,3) se obtiene de la tabla PRECEDENCIA; el primer valor *2* indica que *cual* está siendo utilizado con la categoría gramatical que fue previamente almacenada en DECISIONES y su frecuencia se incrementa ahí. El segundo valor *3* indica que la siguiente palabra en el texto o es una palabra del conjunto D o es una palabra que no puede ser categorizada automáticamente por el uso de las reglas de precedencia, esto es, se trata de una palabra <ambigua>. Ahora se lee el vocablo *si* y los valores (1,2) obtenidos en PRECEDENCIA indican que se incremente la frecuencia de *si* en DICCIONARIO y que la siguiente palabra *fuesen* es un <verbo>; entonces, esta palabra es almacenada en DICCIONARIO con 9 como *catgra,* 30 como *ingra* y 1 de frecuencia.

2.4. *La definición de la función* ig *se almacena en la TABLA de ACTUALIZACION de PRECEDENCIA.*

Análogamente, la función

$$ig: N \times (N \frown N \frown \psi) \rightarrow N$$

permite, como en el ejemplo que utilizamos al describirla, recordar la existencia de un <pronombre>, para marcar como <verbo> el primer elemento a la derecha en el corpus que no esté en D. Esta función es simulada en la computadora como una tabla bidimensional llamada ACTUALIZACION DE PRECEDENCIA, con 33 renglones y 33 columnas como la de PRECEDENCIA. Lo que se logra con esta tabla es modificar (actualizar) o conservar el valor del *ingra* (clase de equivalencia) que sea significativo en un momento dado del análisis, de acuerdo a si se debe o no recordar la existencia de, por ejemplo, el <pronombre>. A continuación se anexa la tabla de ACTUALIZACION de PRECEDENCIA (ilustración 5), note el lector que en general el valor del *ingra* de un vocablo dado (número de renglón) se cambia por el valor respectivo del *ingra* (número de columna) del vocablo que le sigue en el corpus.

2.5. *La definición de la función* pg *se almacena en la TABLA de POSTCEDENCIA.*

La función pg : $N \times (N \frown N) \rightarrow M$ nos permite decidir la categoría gramatical de un vocablo Y en un contexto XYZ, cuando $Y \epsilon I_i$ y $Z \epsilon I_j$, y el vocablo X no nos dio suficiente información con respecto a la categoría de Y. Z puede estar en I_j ya sea porque Z está en D o porque al aplicarle el algoritmo morfológico se le asigne el *in-*

Ilustración 5

TABLA DE ACTUALIZACION DE PRECEDENCIA

	0.	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.	21.	22.	23.	24.	25.	26.	27.	28.	29.	30.	31.	32.
0.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
2.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	2	22	23	24	25	26	27	28	29	30	31	32
3.	0	3	2	3	4	5	6	3	8	9	10	11	12	13	14	15	16	17	18	19	20	3	22	23	24	25	23	27	28	29	30	31	32
4.	0	1	2	3	4	5	3	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
5.	0	1	2	3	4	5	3	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
6.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
7.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
8.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
9.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
10.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
11.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
12.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32

13. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
14. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
15. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
16. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
17. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
18. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
19. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
20. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
21. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
22. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
23. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
24. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
25. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
26. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
27. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
28. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
29. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
30. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
31. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
32. 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32

Ilustración 6

TABLA DE POSTCEDENCIA

	0.	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.	21.	22.	23.	24.	25.	26.	27.	28.	29.	30.	31.	32.
0.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
1.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
2.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
3.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
4.	3	3	3	3	3	3	3	3	3	3	3	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	2	1	3
5.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
6.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
7.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
8.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
9.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
9.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
10.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
11.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
12.	3	3	3	3	3	3	3	3	3	3	3	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	1	3
13.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3

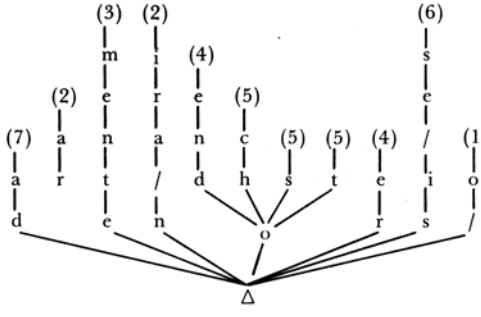
14.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
15.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
16.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1
17.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
18.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
19.	3	3	3	3	3	3	3	3	3	3	3	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	2	1	3
20.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
21.	3	3	3	3	3	3	3	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	1	1	1	1	2	3
22.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	2	1	3
23.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	1	3
24.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	2	3	3
25.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
26.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
27.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
28.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
29.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
30.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
31.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
32.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3

gra igual a j. La tabla de POSTCEDENCIA, anexa anteriormente (ilustración 6), es también un arreglo bidimensional de 33 renglones y 33 columnas, y contiene los valores que nos permiten asociar su categoría gramatical al vocablo Y, buscando en el i-ésimo renglón de la j-ésima columna.

2.6. *Almacenamiento y método de utilización de las TERMINACIONES VERBALES.*

La arborescencia de terminaciones verbales se almacena en la computadora como una tabla con 568 registros, en donde cada registro corresponde a un nodo en el árbol. La tabla está organizada por niveles desde la raíz hacia las hojas y de izquierda a derecha, de manera que todos los nodos que emergen de otro nodo del nivel anterior, están en la tabla en registros secuenciales. Para que el lector entienda cómo se construye esta tabla repetiré la arborescencia asociada al ejemplo de LT que se presentó al principio y construiré la tabla que corresponde a dicha arborescencia:

LT = [–ad, –are, –mente, –ira/n, –endo, –cho, –so, –to, –er, –se/is, –o/]



La tabla correspondiente es la siguiente:

```
 1. * Δ 7 0 0 2
    ---------
 2. * d 1 0 0 9
 3. * e 2 0 1 0
 4. * n 1 0 1 2
 5. * o 4 0 1 3
 6. * r 1 0 1 7
 7. * s 1 0 1 8
 8. * / 1 0 1 9
    ---------
 9. * a 1 0 2 0
10. * r 1 0 2 1
11. * t 1 0 2 2
12. * / 1 0 2 3
13. * d 1 0 2 4
14. * h 1 0 2 5
15. * s 1 0 2 6
16. * t 1 0 2 7
17. * e 1 0 2 8
18. * i 1 0 2 9
19. * o 1 0 3 0
    ----------
20. 7 0 8 9 0 6
21. * a 1 0 3 1
22. * n 1 0 3 2
23. * a 1 0 3 3
24. * n 1 0 3 4
25. * c 1 0 3 5
26. 5 3 4
27. 5 3 4
28. 4
29. * / 1 0 3 6
30. 1 0 8 1 0 8
    ---------
31. 2
32. * e 1 0 3 7
33. * r 1 0 3 8
34. * e 1 0 3 9
35. 5 3 4
36. * e 1 0 4 0
    ---------
37. * m 1 0 4 1
38. * i 1 0 4 2
39. 4
40. * s 1 0 4 3
    ---------
41. 3
42. 2
43. 6
    ---------
```

Como dije antes la tabla está organizada por niveles; cada línea horizontal en la tabla anterior separa un nivel del siguiente, y el primer registro corresponde a la raíz del árbol. Existen dos tipos de registros: los que corresponden a un nodo terminal y los que no son terminales; el primer carácter de cada registro contiene un operador; el * es el único operador no terminal y los opera-

dores terminales son 1, 2, 3, 4, 5, 6 y 7 de acuerdo con las posibles decisiones del algoritmo morfológico (estas posibles decisiones fueron descritas cuando se presentó la arborescencia).

Cuando el operador es * el registro se interpreta como formado por 4 campos: el primero, de longitud uno, es el operador *; el segundo, también de longitud uno, almacena el carácter que debe ser comparado con el correspondiente del vocablo al que se está aplicando el algoritmo morfológico; el tercer campo, de longitud uno, indica el número de ramas que emergen del nodo correspondiente al registro en que estamos; por último, el cuarto campo, de longitud tres, indica el número de registro a partir del cual están secuencialmente almacenados los nodos que emergen del nodo en que estamos. Así por ejemplo, el primer registro de la tabla anterior se interpreta como *, Δ, 7, 002 y nos está diciendo que de la raíz del árbol emergen 7 ramas cuyos nodos están a partir del registro 2, es decir, son los registros 2, 3, 4, 5, 6, 7 y 8.

Esta tabla es la que dirige el reconocimiento morfológico que se inicia en el registro 1. Si el carácter del registro o nodo que se está usando es * se hace la comparación entre el segundo campo en el registro y el carácter correspondiente del vocablo analizado; si coinciden, se lee el número almacenado en el cuarto campo del registro y ése será el registro que dirija ahora el análisis; si los caracteres correspondientes no coinciden se toma el registro inmediatamente siguiente. Cada palabra del corpus que entra al algoritmo morfológico se marca con un carácter Δ al final para que siempre coincida con el carácter Δ del campo dos del registro 1 y se inicie

así la búsqueda en el árbol, siempre empezando en el registro 2.

Siguiendo este procedimiento, se llegará eventualmente a un nodo terminal, es decir, a un registro que no tenga * como operador. Cuando el operador es terminal (1, . . .,7) el registro almacena la información necesaria para realizar la operación adecuada según la terminación reconocida. Estos operadores corresponden a las posibles decisiones descritas cuando se presentó la arborescencia:

(1) Cuando el operador es 1 el registro se interpreta como formado por 3 campos, por ejemplo, el registro número 30 de la tabla anterior será 1, 081, 08; este registro es el nodo terminal que se alcanza siguiendo la trayectoria del árbol que corresponde a la terminación ' o/ '; esta terminación es de <verbo> excepto si el vocablo coincide con determinadas excepciones. El segundo y tercer campos del registro indican que desde el registro 81 de la tabla EXCEPCIONES deben compararse las 8 posibles excepciones. Como ya dije, si el vocablo coincide con la excepción se marca con la categoría que fue previamente almacenada en la tabla EXCEPCIONES, y si no coincide, le corresponde la categoría de <verbo>. En la tabla EXCEPCIONES (ilustración 7), el 1 corresponde a <nominal>, el 2 a <ambigua> y el 3 a <verbo>.

(2) Cuando el operador es 2, el vocablo se marca como <verbo>.

3) Cuando el operador es 3, el vocablo se marca como <adverbio>.

(4) Cuando el operador es 4, la palabra analizada se marca como <ambigua>.

(5) Cuando el operador es 5, el registro se in-

Ilustración 7

EXCEPCIONES

1:AUXILIARI/A	1	45:CHICHICASTE	1	89:NI/	2
2:JUGLARI/A	1	46:CODASTE	1	90:O	2
3:AVEMARI/A	1	47:PASTE	1	91:Y·	2
4:COMISARI/A	1	48:ENGASTE	3	92:CON	1
5:SECRETARI/A	1	49:CHISTE	1	93:SIN	1
6:SUBSECRETARI/A	1	50:ALPISTE	1	94:CUALQUIERA	2
7:TESTAMENTARI/A	1	51:TRISTE	1	95:PEOR	2
8:NOTARI/A	1	52:QUISTE	1	96:MEJOR	2
9:VICARI/A	1	53:SAN	1	97:COMO	2
10:PENITENCIARI/A	1	54:CAN	1	98:MUY	2
11:DEPOSITARI/A	1	55:PAN	1	99:TAN	2
12:DI/A	1	56:CLAN	1	100:CONFORME	1
13:MERCED	1	57:GRAN	1	101:NI	1
14:RED	1	58:FLAN	1	102:O	2
15:PARED	1	59:PLAN	1	103:Y	2
16:SED	2	60:CUAN	1	104:CON	2
17:CESPED	1	61:CRAN	1	105:SIN	2
18:HUESPED	1	62:CHAN	1	106:MEDIANTE	2
19:CID	1	63:ORDEN	1	107:TRAS	2
20:ARDID	1	64:DESGASTES	2	108:CUALQUIERA	2
21:LID	1	65:EMPASTES	2	109:CONFORME	2
22:ADALID	1	66:TRASTES	1	110:NI	2
23:VID	1	67:CONTRASTES	2	111:O	2
24:ASPID	1	68:DESBASTES	2	112:Y	2
25:QUID	1	69:CHICHICASTES	1	113:CON	2
26:BASE	1	70:CODASTES	1	114:SIN	2
27:CLASE	1	71:PASTES	1	115:MEDIANTE	2
28:SUBCLASE	1	72:ENGASTES	2	116TAMBIEN	2
29:PASE	3	73:CHISTES	1	117:MATASIETE	1
30:FRASE	1	74:ALPISTES	1	118:BEREBERE	1
31:FASE	1	75:TRISTES	1	119:CE/LERE	1
32:ENGRASE	3	76:QUISTES	1	120:CHE/VERE	1
33:DESENGRASE	3	77:PAGARE/	2	121:CONGE/NERE	1
34:DESGRASE	3	78:MOARE/	1	122:MISERERE	1
35:ENVASE	3	79:YACARE/	1	123:TI/TERE	1
36:TRASVASE	3	80:MUARE/	1	124:COMO	2
37:UCASE	1	81:CALO/	1	125:FUERA	2
38:CESE	3	82:CHAPO/	1	126:FUERAS	2
39:MAESE	1	83:SANSEACABO/	1	127:PARA	2
40:DESGASTE	3	84:ROCOCO/	1	128:FRESES	2
41:EMPASTE	3	85:RONDO/	2	129:APRESES	2
42:TRASTE	1	86:LANDO/	1	130:TA+ERES	2
43:CONTRASTE	3	87:COCO/	1		
44:DESBASTE	3	88:CHACO/	1		

terpreta como formado por dos campos. El segundo campo, de longitud dos, es el número de un renglón de la tabla TERVER. Esta tabla bidimensional con 37 renglones y 33 columnas (las 33 clases de equivalencia) es la que simula la función tv : L x N * L → M.

Tomemos de nuevo aquel ejemplo:

"contra viento y marea"

contra ϵI_{20} y al analizar morfológicamente el vocablo *viento* se llega al registro 27 de la ARBORESCENCIA que usamos de ejemplo, que contiene un 34; entonces, se busca en el trigésimo cuarto renglón y la vigésima columna de TERVER y se obtiene un 31, que nos indica que *viento* es <nominal>; así hemos calculado tv(to, [20] to). En TERVER el 31 indica <nominal>; el 30 indica <verbo>; el 4 indica que una secuencia de palabras así no puede darse, y tanto el 0 como el 1 nos dan decisión <ambigua>; el 1 se refiere a un tipo especial de ambigüedad producida por la confusión de las terminaciones verbales *–ase, –aste, –iste, –os, –amos* y *–emos* con los pronombres enclíticos *–se, –te* y *–os.*

(6) Cuando el operador es 6 el vocablo analizado se marca como <numeral>.

(7) Cuando el operador es 7, el registro se interpreta como formado por 3 campos, por ejemplo, el registro 20 de la tabla anterior será 7, 089, 06; este registro corresponde al nodo terminal durante el reconocimiento de la terminación *–ad* que es de <verbo> excepto si el vocablo con dicha terminación está precedido de palabras específicas. Como en (1), el segundo y tercer campos del registro indican que desde el registro 89 deben compararse las 6 posibles excep-

Ilustración 8

TERVER

	0.	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.	21.	22.	23.	24.	25.	26.	27.	28.	29.	30.	31.	32.	
1.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	0	31	0	30	30	30	0	0	0	0	31	31	0	31	-aba
2.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	0	31	0	30	30	30	0	0	0	0	31	31	0	31	-i/a
3.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	0	31	0	30	30	30	0	0	0	0	31	31	0	31	-ari/*a*
4.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	0	31	0	30	30	30	0	0	0	0	31	31	0	31	-eri/a
5.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	0	31	0	30	30	30	0	0	0	0	31	31	0	31	-ara
6.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	0	31	0	30	30	30	0	0	0	0	31	31	0	31	-iera
7.	0	30	31	30	31	31	31	31	31	31	30	4	31	31	30	0	0	0	31	31	31	31	30	31	31	31	31	0	0	31	31	0	31	-ad
8.	0	30	31	30	31	31	31	31	31	31	30	4	31	31	30	0	0	0	31	31	31	31	30	31	31	31	31	0	0	31	31	0	31	-ed
9.	0	30	31	30	31	31	31	31	31	31	30	4	31	31	30	0	0	0	31	31	31	31	30	31	31	31	31	0	0	31	31	0	31	-id
10.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	0	31	0	30	0	0	0	0	0	0	31	31	0	31	-e
11.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	0	31	0	0	0	0	0	0	0	0	31	31	0	31	-e/
12.	0	30	31	30	30	31	0	30	31	31	30	4	0	31	30	0	30	0	30	30	31	0	30	0	0	0	0	0	0	31	31	0	31	-are/
13.	1	30	31	30	1	31	1	1	31	31	30	4	1	31	30	1	30	1	30	1	31	1	30	1	1	1	1	1	1	31	31	1	31	-ase/
14.	1	30	31	30	30	31	1	1	31	31	30	4	1	31	30	1	30	1	30	30	31	1	30	1	1	1	1	1	1	31	31	1	31	-aste
15.	1	30	31	30	1	31	1	1	31	31	30	4	1	31	30	1	30	1	30	30	31	1	30	1	1	1	1	1	1	31	31	1	31	-iste

16.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	0	31	0	30	0	0	0	0	0	0	31	31	0	31	-i/
17.	1	1	31	30	31	31	31	31	31	31	30	4	31	31	30	1	30	1	4	31	31	31	31	31	31	31	31	1	1	31	31	1	31	-os
18.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	0	31	0	30	0	0	0	0	0	0	31	31	0	31	-en
19.	0	30	31	30	0	31	0	0	31	31	0	4	0	31	30	0	30	0	30	0	31	0	30	0	0	0	0	0	0	31	31	0	31	-o
20.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	30	31	0	30	0	0	0	0	0	0	31	31	0	31	-as
21.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	30	31	0	30	0	0	0	0	0	0	31	31	0	31	-abas
22.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	30	31	0	30	0	0	0	0	0	0	31	31	0	31	-i/as
23.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	30	31	0	30	0	0	0	0	0	0	31	31	0	31	-ari/as
24.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	30	31	0	30	0	0	0	0	0	0	31	31	0	31	-eri/*as*
25.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	30	31	0	30	0	0	0	0	0	0	31	31	0	31	-aras
26.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	30	31	0	30	0	0	0	0	0	0	31	31	0	31	-es
27.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	30	31	0	30	0	0	0	0	0	0	31	31	0	31	-ares
28.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	30	31	0	30	0	0	0	0	0	0	31	31	0	31	-ases
29.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	30	31	0	30	0	0	0	0	0	0	31	31	0	31	-astes
30.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	30	31	0	30	0	0	0	0	0	0	31	31	0	31	-istes
31.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	0	31	0	30	0	0	0	0	0	0	31	31	0	31	-i/s
32.	1	30	31	30	1	31	1	1	31	31	30	4	1	31	30	1	30	1	30	30	31	1	30	1	1	1	1	1	1	31	31	1	31	-amos
33.	1	30	31	30	1	31	1	1	31	31	30	4	1	31	30	1	30	1	30	30	31	1	30	1	1	1	1	1	1	31	31	1	31	-emos •
34.	0	0	31	30	0	31	0	0	31	31	0	4	0	0	30	0	0	0	30	0	31	0	0	0	0	0	0	0	0	0	31	0	31	-to,-so,-cho,...
35.	0	0	31	30	0	31	0	0	31	31	0	4	0	0	30	0	0	0	30	0	31	0	0	0	0	0	0	0	0	0	31	0	31	-ando,-endo,-iendo,...
36.	0	30	31	30	0	31	0	0	31	31	30	4	0	0	30	0	30	0	30	30	31	0	0	0	0	0	0	0	0	31	31	0	31	-eras
37.	0	30	31	30	0	31	0	0	31	31	30	4	0	31	30	0	30	0	30	0	31	0	30	0	30	0	0	0	0	31	31	0	31	-era

ciones, con la palabra precedente a la que se ha analizado morfológicamente. Si el vocablo precedente coincide con alguna excepción, la tabla EXCEPCIONES contiene la categoría gramatical que debe ser asignada al vocablo analizado.

A continuación doy los registros de la tabla asociada al ejemplo de LT, que se seleccionan durante el reconocimiento de la terminación del vocablo *fuertemente:*

1. * Δ 7 0 0 2	fuertemente Δ
2. * d 1 0 0 9	fuerteme*nte* Δ
3. * e 2 0 1 0	fuerteme*nte* Δ
10. * r 1 0 2 1	fuerteme*nte* Δ
11. * t 1 0 2 2	fuerteme*nte* Δ
22. * n 1 0 3 2	fuerteme*nte* Δ
32. * e 1 0 3 7	fuerteme*nte* Δ
37. * m 1 0 4 1	fuerte*mente* Δ
41. 3	am(fuerte*mente*) = <adverbio>

A continuación se anexan la tabla de EXCEPCIONES, la tabla TERVER (ilustración 8) y la ARBORESCENCIA DE TERMINACIONES VERBALES (ilustración 9) completa. En la ARBORESCENCIA, los operadores (primer carácter de cada registro) no son como en nuestro ejemplo sino que corresponden a lo que sigue:

1. Es el único operador no terminal, es decir,*.
2. Corresponde al 1 de nuestro ejemplo.
3. Corresponde al 7 de nuestro ejemplo.
4. Corresponde al 5 de nuestro ejemplo.
5. Corresponde al 2 de nuestro ejemplo.
6. Correspondía a una terminación <nominal> que fue excluida en la versión definitiva.
7. Corresponde al 3 de nuestro ejemplo.

Ilustración 9

ARBORESCENCIA DE TERMINACIONES VERBALES

1:1Δ8002
2:1A8010
3:1D3018
4:1E8021
5:1N4029
6:109033
7:1R3042
8:1S5045
9:1/4050
10:1B1054
11:1H1055
12:1L3056
13:1R2059
14:1S1061
15:1T1062
16:1/1063
17:0
18:1A2064
19:1E3066
20:1I3069
21:1L3072
22:1M2075
23:1R2077
24:1S3079
25:1T4082
26:1V1086
27:308906
28:410
29:1A5087
30:1E5092
31:101097
32:1/1098
33:1C1099
34:1H1100
35:1L3101
36:1D3104
37:1R1107
38:1S1108
39:1T1109
40:310108
41:419
42:1A1110
43:1E1111
44:1I1112
45:1A9113
46:1E6122
47:1I3128
48:109131
49:1/2140
50:1A1142
51:1E3143
52:1I2146
53:103148
54:1A2151
55:1C1153
56:1E3154
57:1S1157
58:9
59:1A2158
60:1E2160
61:434
62:434
63:1I4162
64:308906
65:407
66:201306
67:308906
68:408
69:201907
70:308906
71:409
72:1E3166
73:1S1169
74:9
75:1E2170
76:9
77:1A1172
78:1E3173
79:1A3176
80:1E4179
81:9
82:1E2183
83:1N1185
84:1S2186
85:9
86:1E1188
87:1B1189
88:1R2190
89:1/1192
90:205310
91:5
92:1R2193
93:1S2195
94:206301
95:310108
96:418
97:1R2197
98:1A1199
99:1N1200
100:1C1201
101:1E3202
102:1S1205
103:9
104:1A1206
105:1I1207
106:1N2208
107:1T1210
108:434
109:434
110:0
111:0
112:0
113:1B1211
114:1H1212
115:1L3213
116:1R2216
117:1S1218
118:1T1219
119:1/1220
120:310107
121:420
122:1L3221
123:1R2224
124:1S2226
125:1T1228
126:310107
127;426
128:1A4229
129:1E3233
130:1/1236
131:1A1237
132:1E3258
133:1H1241
134:1L3242
135:1M3245
136:1N5248
137:1S4253
138:1T1257
139:417
140:1A1258
141:1I2259
142:1R4261
143:1R3265
144:308906
145:411
146:310107
147:416
148:1I1268
149:208108
150:5
151:308909
152:401
153:434
154:1M1269
155:1S1270
156:1T1271
157:102272
158:308909
159:405
160:1T2274
161:212404
162:1R3276
163:201201
164:308909
165:402
166:1M1279
167:1S1280
168:1T1281
169:102282
170:1S1284
171:1T1285
172:5
173:1T1286
174:311806
175:5
176:202612
177:310108
178:413
179:1T1287
180:1T1288
181:203802
182:5
183:1T1289
184:1S1290
185:1E1291

186:1A3292
187:1I3295
188:1U1298
189:1A1299
190:1A1300
191:1E2301
192:1I2303
193:1A1305
194:1E2306
195:1A1308
196:1E2309
197:1A1311
198:1E2312
199:1R3314
200:1I1317
201:434
202:1M1318
203:1S1319
204:1T1320
205:102321
206:0
207:0
208:1A1323
209:1E2324
210:1A1326
211:1A2327
212:1C1329
213:1E3330
214:1S1333
215:9
216:1A2334
217:1E3336
218:434
219:434
220:1I3339
221:1E3342
222:1S1345
223:9
224:1A2346
225:1E3348
226:1A2351
227:1E3353
228:1S2356
229:1B1358
230:1R2359
231:1/1361
232:5
233:1R4362
234:1S2366
235:1T1368
236:1E3369
237:1L1372
238:1L1373
239:1S1374
240:1T1375
241:1C1376
242:1E3377
243:1S1380
244:9
245:1A5381
246:1E4386
247:1I2390
248:1A1392
249:1E3393
250:101396
251:1S3397
252:9
253:1A1400
254:1E1401
255:101402
256:434
257:434
258:1R4403
259:310007
260:431
261:1A1407
262:1E1408
263:1I1409
264:5
265:1A3410
266:1E1413
267:1I1414
268:5
269:9
270:9
271:9
272:1N1415
273:9
274:308911
275:406
276:1A3416
277:1E2419
278:1I1421
279:9
280:9
281:9
282:1N1422
283:9
284:9
285:9
286:5
287:5
288.9
289:1S2423
290:9
291:1M1425
292:204009
293:310108
294:414
295:204904
296:1N1426
297:309401
298:415
299:5
300:5
301:1I1427
302:5
303:1R3428
304:5
305:5
306:1I1431
307:5
308:5
309:1I1432
310:5
311:5
312:1I1433
313:5
314:1A1434
315:1E1435
316:1I1436
317:1C1437
318:9
319:9
320:9
321:1N1438
322:9
323:435
324:1I1439
325:435
326:1U1440
327:310107
328:421
329:434
330:1M1441
331:1S1442
332:1T1443
333:102444
334:310107
335:425
336:1I1446
337:311007
338:436
339:1R3447
340:310107
341:422
342:1M1450
343:1S1451
344:1T1452
345:102453
346:310107
347:427
348:1I1455
349:213001
350:A
351:310107
352:428
353:1I1456
354:212802
355:A
356:1A3457
357:1I3460
358:1A1463
359:1A1464
360:1E2465
361:1I2467
362:1A1469
363:1E1470
364:1I1471
365:1/1472
366:1A1473
367:1E2474
368:1S1476
369:1R1477
370:1S1478
371:5
372:9
373:9
374:9
375:9
376:434
377:1M1479
378:1S1480
379:1T1481
380:102482
381:1B1484
382:1R1485
383:/1486
384:310007
385:432

386:1R4487
387:1S1491
388:310007
389:433
390:1S1492
391:5
392:1L1493
393:1I1494
394:1S1495
395:1T1496
396:1L1497
397:1A1498
398:1E1499
399:101500
400:1L1501
401:1L1502
402:1L1503
403:1A1504
404:1E1505
404:1I1506
406:5
407:5
408:5
409.5
410:207704
411:310108
412:412
413:5
414:5
415:9
416:200111
417:308909
418:403
419:308909
420:404
421:5
422:9
423:211701
424:8
425:7
426:8
427:5
428:1A1507
429:1E1508
430:1I1509
431:5
432:5
433:5
434:5
435:5
436:5
437:8
438:9
439:435
440:1C1510
441:9
442:9
443:9
444:1N1511
445:9
446:5
447:1A2512
448:1E2514
449:1I116
450:9
451:9
452:9
453:1N1517
454:9
455:5
456:5
457:206:409
458:310107
459:429
460:207304
461:310107
462:430
463:5
464:5
465:1I1518
466:5
467:1R3519
468:5
469:5
470:5
471:5
472:1E1522
473:5
474:1I1523
475:5
476:1/2524
477:1A1526
478:8
479:9
480:9
481:9
482:1N1527
483:9
484:1/1528
485:1/2529
486:1I2531
487:1A1533
488:1E1534
489:1I1535
490:1/2536
491:1/2538
492:1/1540
493:9
494:9
495:9
496:9
497:9
498:1L1541
499:1L1542
500:1L1543
501:9
502:9
503:9
504:5
505:5
506:5
507:5
508:5
509:5
510:8
511:9
512:310107
513:423
514:310107
515:424
516:5
517:9
518:5
519:1A1544
520:1E1545
521:1I1546
522:1I1547
523:5
524:1A1548
525:1I1549
526:5
527:9
528:1A1550
529:1A1551
530:1E2552
531:1R3554
532:5
533:5
534:5
535:5
536:1A1557
537:1E1558
538:1A1559
539:1E2560
540:1I1562
541:9
542:9
543:9
544:5
545:5
546:5
547:5
548:5
549:5
550:5
551:5
552:1I1563
553:5
554:1A1564
555:1E1565
556:1I1566
557:5
558:1I1567
559:5
560:1I1568
561:5
562:0
563:5
564:5
565:5
566:5
567:5
568:5

8. Corresponde al 6 de nuestro ejemplo.

9. Da como decisión <ambigua> pero habla precisamente del caso en que es posible que la terminación reconocida esté afectada por la presencia de un pronombre enclítico.

A. Se usa en el caso de terminaciones que pueden ser de dos categorías distintas pero en las que no es posible una homografía, es decir, las palabras con esa terminación se separan en dos grupos ajenos. En este caso, si la categoría gramatical de un vocablo ha sido determinada en un contexto específico a través de las funciones de precedencia, se sabe que esa categoría debe ser preservada; de modo que si en otro contexto no se puede decidir la categoría del vocablo mediante las funciones de precedencia y es necesario aplicarle el algoritmo morfológico, se llegará a la solución gramatical tomada en el primer contexto.

0. Corresponde al 4 de nuestro ejemplo.

2.7. *Otros archivos necesarios.*

Dos archivos más: AMBIGÜEDADES y LIGAS son usados como sigue: en AMBIGÜEDADES se van almacenando todas las ocurrencias que se deciden como <ambiguas>, para un solución posterior manual; LIGAS se usa para almacenar todos los apuntadores que permiten desde DICCIONARIO o DECISIONES recuperar en CORPUS todas sus concordancias para cualquier vocablo. Existe además un archivo para el control de procesos.

2.8. *El proceso de análisis gramatical.*

Todas las ocurrencias que se reconocen como no <ambiguas> en CORPUS, se van incorporando a DICCIONARIO a través de una función de *hash*, pero si un homógrafo de un vocablo ya en

DICCIONARIO se reconoce en CORPUS, es almacenado en DECISIONES con su *catgra* adecuado.

En términos generales, el proceso de análisis es como sigue: Supongamos que RXYZS ϵ C con YϵD y que XϵI_i y YϵI_j y RϵI_k y estamos analizando la ocurrencia Y. Como Y está en D se puede obtener el valor de j. Aplicamos la función ig en el contexto RXY para determinar el *ingra*, que asociado a X, debe ponerse en funcionamiento. Digamos que ig([i], [k] [i] Y) = e. El valor e se obtiene buscando en el k-ésimo renglón y la i-ésima columna de ACTUALIZACION DE PRECEDENCIA. Esto significa que en el contexto RXY, X debe funcionar como un elemento de la clase I_e (generalmente e = i). Ahora se calculan cg([j], [e] [j]Z) y cgs(Z) en el contexto XYZ para decidir la categoría de Y y obtener información sobre Z; estos valores de cg y cgs se obtienen buscando en el e-ésimo renglón y la j-ésima columna de PRECEDENCIA de donde se saca una pareja de valores (a,b). El elemento *a* de la pareja (a,b), según tome los valores 1,2,3 o 4 indica respectivamente: que Y se está usando con el *catgra* de DICCIONARIO (se incrementa en DICCIONARIO su frecuencia); o que Y se está usando con el *catgra* de DECISIONES (se hace el incremento de frecuencia en DECISIONES); o que Y es <ambigua>, es decir, el contexto no fue suficiente para decidir la categoría gramatical de esa ocurrencia de Y (se escribe en AMBIGÜEDADES); o que debe aplicarse la función pg en el contexto XYZ. El elemento *b* de la pareja (a,b) que puede valer 1, 2 o 3 indica que Z puede ser <nominal>, <verbal> o <ambigua> respectivamente. Si *a* es igual a 4 se aplicará a Z el

algoritmo morfológico, o si Z está en D se recuperará de DICCIONARIO el valor de su *ingra*, cualquiera que sea la solución sobre el *catgra* de Z, un valor digamos *m* para su *ingra* es obtenido; de manera que se busca en el j-ésimo renglón y la m-ésima columna de la tabla POSTCEDENCIA que guarda los valores de la función pg y obtendremos la decisión adecuada correspondiente para Y (incrementar frecuencia en DICCIONARIO, en DECISIONES o escribirla en AMBIGÜEDADES, según se obtenga 1, 2 o 3 respectivamente). Cualquiera que sea la decisión para Y, se obtiene el valor adecuado de la clase a la que la ocurrencia analizada debe pertenecer y así se puede avanzar a analizar Z en el contexto XYZSTϵ C, teniendo los requeridos valores del *ingra* de X y de Y.

Para el caso en que Y $\notin$ D en el contexto RXYZ se aplica el algoritmo morfológico a Y. Este algoritmo, cuando da la solución de <verbo>, separa la ocurrencia de Y en su raíz y su terminación para que la raíz identificada con un '–' al final (*cant–* para *cantar*, *s–* para *ser*, *o–* y *oy–* para *oir*) sea almacenada en DICCIONARIO. Esta forma de almacenar los <verbos> permite la agrupación de ocurrencias diferentes que correspondan al mismo <verbo>. Asimismo se imposibilita que al leer de nuevo una ocurrencia que se ha marcado como <verbo> y agregado a DICCIONARIO, se encuentre ahí (siempre serán diferentes las ocurrencias y los <verbos> almacenados); al no encontrarse la ocurrencia en DICCIONARIO se efectúa de nuevo el proceso de reconocimiento de su categoría gramatical, permitiendo la separación de un posible homógrafo. El criterio probabilístico arbitra-

rio que hemos usado para separar una ocurrencia etiquetada como <verbo> en su raíz y su terminación, es el siguiente: se considera terminación a la cadena más larga de caracteres terminales de la ocurrencia dada que coincidan con una terminación en LT, por ejemplo, *cantari/a* se separa en *cant–* y *–ari/a* correctamente, pero también se dan casos en que la ocurrencia quedará separada incorrectamente como por ejemplo, *aumente* se separa en *au–* y *–mente*.

3. *La aplicación del sistema computacional.*

Los archivos DICCIONARIO y DECISIONES, así como las matrices correspondientes a todas las funciones, y las tablas EXCEPCIONES y ARBORESCENCIA, son almacenados en memoria central, en cambio, CORPUS, AMBIGÜEDADES y LIGAS son manejados siempre en disco. La idea de tener DICCIONARIO y DECISIONES en memoria central fue para hacer eficiente la búsqueda, mediante un *hash*, de cada ocurrencia de CORPUS; esto trajo como consecuencia que DICCIONARIO pudiera contener un máximo de 2000 vocablos diferentes, pues disponíamos de 55 K palabras de 36 bits en memoria central. El proceso, por lo tanto, tuvo que ser en secciones de más o menos 10 textos cada vez y los resultados parciales fueron mezclados para producir una estructura definitiva de archivos DICCIONARIO, DECISIONES, LIGAS y CORPUS, además de AMBIGÜEDADES, todos en disco.

A continuación (ilustración 10) se presenta una muestra de los resultados obtenidos mediante el analizador gramatical que he descrito.

Las palabras <ambiguas> que van llenando AMBIGÜEDADES, se enlistan para que el equi-

GEMIDO	INGRA: 31	CATGRA:NOMINAL	FRECUENCIA: 1
SORDO	INGRA: 0	CATGRA:AMBIGUA	FRECUENCIA: 0
Y	INGRA: 1	CATGRA:CONJUNCI	FRECUENCIA: 1
NO	INGRA: 7	CATGRA:ADVERBIO	FRECUENCIA: 1
LADR-	INGRA: 30	CATGRA:VERBO	FRECUENCIA: 1
MA/S	INGRA: 23	CATGRA:AMBIGUA	FRECUENCIA: 0
DESOERT-	INGRA: 30	CATGRA:VERBO	FRECUENCIA: 1
SOBRESALTADO	INGRA: 0	CATGRA:AMBIGUA	FRECUENCIA: 0
VADE-	INGRA: 30	CATGRA:VERBO	FRECUENCIA: 1
EL	INGRA: 5	CATGRA:ARTICULO	FRECUENCIA: 2
RI/O	INGRA: 31	CATGRA:NOMINAL	FRECUENCIA: 1
Y	INGRA: 1	CATGRA:CONJUNCI	FRECUENCIA: 2
TOM-	INGRA: 30	CATGRA:VERBO	FRECUENCIA: 1
LA	INGRA: 4	CATGRA:AMBIGUA	FRECUENCIA: 0
VERTIENTE	INGRA: 0	CATGRA:AMBIGUA	FRECUENCIA: 0
OPUESTA	INGRA: 0	CATGRA:AMBIGUA	FRECUENCIA: 0

Ilustración 10

ANALISIS GRAMATICAL

@000194 @000007003 @DE PRONTO SE OYO/ UN DISPARO, EL PERRO LANZO/ UN GEMIDO SORDO Y NO
@000195 @000007004 @LADRO/ MA/S.
@000196 @000013014 @DEMETRIO @ DESPERTO/ SOBRESALTADO, VADEO/ EL RI/O Y TOMO/ LA VERTIENTE
@000197 @000013015 @OPUESTA DEL CA+O/N. COMO HORMIGA ARRIERA ASCENDIO/ LA CRESTERI/A,
@000198 @000013016 @CRISPADAS LAS MANOS EN LAS PE+AS Y RAMAZONES, CRISPADAS LAS PLANTAS
@000199 @000013017 @SOBRE LAS GUIJAS DE LA VEREDA.
@000200 @000020007 @LOS FEDERALES GRITABAN A LOS ENEMIGOS, QUE, OCULTOS, QUIETOS Y

DE	INGRA: 13	CATGRA:PREPOSIC	FRECUENCIA: 1
PRONTO	INGRA: 1	CATGRA:ADVERBIO	FRECUENCIA: 1
SE	INGRA: 3	CATGRA:PRONOMBR	FRECUENCIA: 2
OY–	INGRA: 30	CATGRA:VERBO	FRECUENCIA: 1
UN	INGRA: 2	CATGRA:ARTICULO	FRECUENCIA: 1
DISPARO	INGRA: 31	CATGRA:NOMINAL	FRECUENCIA: 1
EL	INGRA: 5	CATGRA:ARTICULO	FRECUENCIA: 1
PERRO	INGRA: 31	CATGRA:NOMINAL	FRECUENCIA: 1
LANZ–	INGRA: 30	CATGRA:VERBO	FRECUENCIA: 1
UN	INGRA: 2	CATGRA:ARTICULO	FRECUENCIA: 2

12	CRISPADAS	000013016	NM	38	HABI/A	000020009	VE HAB-
13	LAS	000013016	AR	39	HECHO	000020009	VE HAC-
14	MANOS	000013016	NM	40	FAMOSOS	000020009	NM
15	LAS	000013016	AR	41	TIRANDO	000023012	VE TIR-
16	PE+AS	000013016	NM	42	ADVIRTIENDO	000023012	VE ADVERT-
17	RAMAZONES	000013016	NM	43	PELIGRO	000023012	NM
18	CRISPADAS	000013016	NM	44	LOS	000023012	AR
19	LAS	000013017	AR	45	OTROS	000023012	PN
20	PLANTAS	000013016	NM	46	DESESPERADA	000023013	NM
21	SOBRE	000013017	PR	47	QUE	000023013	CO
22	LAS	000013017	AR	48	LAS	000023014	HR
23	GÜIJAS	000013017	NM	49	BALAS	000023014	NM
24	LA	000013017	AR	50	UNO	000023014	PN
25	VEREDA	000013017	NM	51	LOS	000027003	AR
26	LOS	000020007	AR	52	SERRANOS	000027003	NM

Ilustración 11

AMBIGÜEDADES

1	SORDO	000007003	NM	27	FEDERALES	000020007	NM
2	MA/S	000007004	AV	28	QUE	000020007	PN
3	SOBRESALTADO	000013014	AR	29	OCULTOS	000020007	NM
4	LA	000013014	AR	30	QUIETOS	000020007	NM
5	VERTIENTE	000013014	NM	31	CALLADOS	000020008	NM
6	OPUESTA	000013015	NM	32	HACIENDO	000020008	VE HAC-
7	COMO	000013015	AV	33	GALA	000020008	NM
8	HORMIGA	000013015	NM	34	UNA	000020008	AR
9	ARRIERA	000013015	NM	35	PUNTERI/A	000020008	NM
10	LA	000013015	AR	36	QUE	000020008	PN
11	CRESTERI/A	000013015	NM	37	LOS	000020009	PN

po de lingüistas, refiriéndose a los textos originales, las etiqueten gramaticalmente, como se puede observar en la ilustración 11.

Hecho lo anterior, las etiquetas gramaticales de las palabras <ambiguas> se incorporan a la estructura de archivos definitiva DICCIONARIO, DECISIONES, LIGAS, CORPUS, mediante el uso de programas interactivos entre la computadora y los lingüistas.

El número exacto de ocurrencias en CORPUS, con excepción de nombres propios, abreviaturas y fechas, fue de 1,973,151; de éstas, 889,006 permanecieron <ambiguas>, lo que da un 55% de eficiencia de nuestro analizador gramatical. Quiero hacer notar que la eficiencia del analizador en el género de Literatura fue de 67.5%.

A continuación presento una muestra del DICCIONARIO durante el análisis de los primeros textos. En él pueden verse todos los vocablos que se incluyeron en DICCIONARIO debido a que fueron gramaticalmente etiquetados en forma automática. La información almacenada en este DICCIONARIO PROCESADO es la que sigue: por ejemplo, el registro 742 contiene el vocablo *so/tano* que tiene 8 como *catgra,* es decir, es <nominal>, tiene un 31 como *ingra* y un 1 como frecuencia en ese momento. El siguiente número en ese registro, 914, es la liga de dispersión y significa que la palabra del registro 914 tiene el mismo *hash* que ésta del 742 (el *hash* es una función que asocia a cada vocablo un número que permite su rápida localización en el archivo). La liga de dispersión existe solamente en el DICCIONARIO en memoria, pues el DICCIONARIO en disco está ordenado alfabéticamente; también, el DICCIONARIO en memoria contie-

Ilustración 12

* DICCIONARIO PROCESADO *

740.TROZO	8311		1500	1500
741.US-	9301		1503	1503
742.SO/TANO	8311	914	1510	1510
743.SUE+-	9301		1511	1511
744.EJERCIT-	9301		1515	1515
745:INICI-	9301		1518	1518
746:EQUIVOCADOS	8311		1520	1520
747:OPRIM-	9301		1523	1523
748:PECHO	8311		1525	1525
749:PUED-	9301		1527	1527
750.CISNE	8311	922	1533	1533
751.VISIT-	9301	958	1534	1534
752.CUENT-	9301		1537	1537
753.COSAS	8311		1538	1538
754.COMID-	9301		1542	1542
755.ALMUERZOS	8311		1544	1544
756.DESCRIBEN-	9301	761	1548	1548
757.NATURALMENTE	11 1		1551	1551
758.AUGE	8311		1555	1555
759.ANTIGUAS	8311	913	1558	1558
760.PLATOS	8311		1560	1560
761.METO/DICAMENTE	11 1	866	1562	1562
762.PATAGRA/S	8311		1567	1567
763.GORGONZOLA	8311	839	1570	1570
764.ROBLE	8311	878	1574	1574
765.COMEDOR	8311		1576	1576
766.UNTUOSA	8311		1579	1579
767.INSI/PIDAS	8311	881	1582	1582
768.SILENCIO	8311		1585	1585
769.NI+O	8314		1587	2459
770.HABANO	8311		1591	1591
771.TABACO	8311	777	1597	1597
772.PIPA	8311		1600	1600
773.WHISKY	8311		1603	1603
774.DEUDAS	8311	830	1606	1606
775.CRI/TICOS	8311		1609	1609
776.FUTURO	8311		1611	1611
777.TALEGO	8311		1614	1614

778.IMA/GENES	8311		1616	1616
779.GE/NERO	8311		1618	1618
780.HIJOS	8311		1621	1621
781.MALGAST-	9301	870	1622	1622
782.DINERO	8311	815	1624	1624
783.ANTEPASADOS	8311		1627	1627
784.PROPIAS	8311	918	1632	1632
785.NOCHES	8311		1644	1644
786.SOLITARIO	8311	817	1646	1646
787.GERMEN	8311		1649	1649
788.DIOS	8311	949	1651	1651
789.SOFA/	8311	900	1662	1662
790.MUROS	8311	962	1667	1667
791.ARMONIOSAMENTE	11 1		1668	1668
792.DESNUDO	8311	895	1670	1670
793.FONDO	8313		1678	2154
794.PARED	8311		1680	1680
795.INTEGRAMENTE	11 1		1683	1683
796.GRAN	8311	809	1686	1686
797.ALUMINIO	8311		1689	1689
798.ESCRITORIO	8316		1691	2339
799.SILLO/N	8311		1694	1694
800.A/NGULO	8311		1698	1698
801.JUICIO	8311		1703	1703
802.DEDO	8311	876	1705	1705
803.BRONCE	8311	935	1707	1707
804.ESCUCH-	9301		1714	1714
805.TOQUE	8311		1719	1719
806.LEVAN-	9301		1724	1724
807.TELO/N	8311		1726	1726
808.BATERI/A	8311		1729	1729
809.PROGRAMA	8312	869	1736	1743
810.APARATO	8313		1748	1770
811.ENCHUFE	8311	836	1753	1753
812.OY-	9301		1761	1761
813.FRASES	8311		1762	1762
814.LARGO	8311		1767	1767
815.DIRIG-	9301		1773	1773
816.PU/BLICO	8311		1775	1775
817.SOBRIEDAD	8311	856	1778	1778
818.VULGARIDAD	8311		1785	1785

819.ANUNCIADOR	8311		1789	1789
820.RADIO	8311		1791	1791
821.IMPORTANTE	8311		1795	1795
822.AGUARDA-	9301	884	1800	1800
823.INMEDIATAMENTE	11 1		1802	1802
824.PRECIO	8312		1805	1856
825.AZU/CAR	8312		1807	1858
826.MARCADO	8311		1810	1810
827.DECL-	9301		1813	1813
828.MISM-	9301		1817	1817
829.PERIODO	8311	834	1820	1820
830.DECIDE	8311		1822	1822
831.ACCIONES	8311		1824	1824
832.CARPETA	8311	891	1826	1826
833.LADO	8311	926	1836	1836
834.PEQUE+O	8312		1843	2350
835.OBREROS	8311		1860	1860
836.NORMALMENTE	11 1		1863	1863
837.PAR	8311		1866	1866
838.SEMANAS	8311	863	1868	1868
839.POD-	9301		1871	1871
840.REAJUSTE	8311		1873	1873
841.ALTER-	9301		1877	1877
842.RITMO	8311		1881	1881
843.ALZA	8311		1886	1886
844.ESTA+O	8311		1888	1888
845.DESCUIDADOS	8311	906	1896	1896
846.REU/N-	9301	917	1899	1899
847.RINCO/N	8311		1903	1903
848.FUM-	9301		1912	1912
849.LENTAMENTE	11 1		1913	1913
850.OSCUREC-	9302		1915	2478
851.LIGERI/SIMO	8311		1920	1920
852.ABRE-	9301	933	1923	1923
853.CHIRRIDO	8311		1926	1926

@005500	@025108013	@ METERSE. SON PELIGROSOS. SIEMPRE ANDAN ARMADOS. DICEN QUE EL PARQUE ES	1109019800689
@005501	@025108014	@SO/LO PARA BLANCOS Y CUALQUIER COCHINO NEGRO QUE PONGA AQUI/ LAS PATAS	12008328000100
@005502	@025108015	@SUFRIRA/ LAS CONSECUENCIAS.	3 900
@005503	@025108016	@–PERO NO HAY DERECHO, ESO NO PUEDE HACERSE.	8 31005190
@005504	@025108017	@–¿LO DICE EN SERIO? ASI/ HABLA LA GENTE DEL BARRIO. PERO CUANDO LLEGA	130048100078310
@005505	@025108018	@EL MOMENTO NO ACEPTA NEGROS EN SUS CASAS NI DEJA QUE SE SIENTEN EN SUS	15681004283005942
@005506	@025108019	@BARES.	1 8
@005507	@025111015	@MR WAUCH @SE APROXIMO/ AL VENTANAL: DESDE EL PISO NU/MERO DIECINUEVE SE	120B5978468005
@005508	@025111016	@OBSERVA UNA CIUDAD QUE NO VEN QUIENES LA MIRAN DESDE LOS RASCACIELOS	12968019500400
@005509	@025111017	@MA/S ALTOS. ALLI/ SE TIENE LA CONFIANZA DE POSEER TODAS LAS VENTAJAS DE	130015900400004
@005510	@025111018	@LA ALTURA Y TODOS LOS BENEFICIOS DE LA CERCANI/A.	9 003068400
@005511	@025117001	@LOS ANTERIORES OCUPANTES TUVIERON QUE ABANDONAR APRESURADAMENTE LA	8 00090010
@005512	@025117002	@CASA. HALLAMOS MUEBLES EN DESORDEN, ROPA ESPARCIDA, CARTAS Y PAPELES	100004800030
@005513	@025117003	@PRIVADOS, ALIMENTOS A MEDIOCONSUMIR EN EL REFRIGERADOR YA CUBIERTO DE	100040468104
@005514	@025117004	@MOHO. EN CAMBIO LA DESPENSA SO/LO GUARDABA UNA LATA DE COMIDA PARA	12048000000400
@005515	@025117005	@ANIMALES. HABI/A UNA CASITA DE MADERA EN EL TRASPATIO PERO NO HUELLAS	12000040468310
@005516	@025117006	@DE GATOS NI DE PERROS. EL TELE/FONO FUE ARRANCADO DE CUAJO, LAS	12403486890480
@005517	@025117007	@CONEXIONES CERCENADAS.	2 00
@005518	@025120019	@PASE/ EL DI/A EN LA FA/BRICA. TODO ANDUVO MENOS MAL DE LO QUE PENSABA.	1406840000014000
@005519	@025120020	@A FIN DE CUENTAS ERA UN EXPERTO Y ME CONTRATARON PORQUE ME NECESITABAN.	134040068359359
@005520	@025120021	@DE REGRESO ENCONTRE/ A ESTER @MUY INQUIETA. NO QUISO HABLARME DE LO QUE	134004B10100400

Ilustración 13

CORPUS PROCESADO

@005483	@025082001	@ANSELMO @PRENDIO/EL CIGARRO DE HOJA Y SE VOLVIO/ PARA MIRARME. EL SOL	13B968404590068
@005484	@025082002	@QUEMABA LOS CAMPOS SECOS, PERO LOS QUE REZABAN CERCA DE LA CHOZA	12068030090400
@005485	@025082003	@PARECI/AN NO SENTIR EL CALOR. ANSELMO @RECARGO/ LA SILLA CONTRA EL MURO	1291068B900468
@005486	@025082004	@DE ADOBES. PIDIO/ QUE ACERCARA MI ASIENTO Y COMENZO/ SU NARRACIO/N:	1148900283928
@005487	@025082005	@–QUESQUE FUE AURORITA @LA QUE PRIMERO VIO A LA VIRGEN @. UNA MA+ANA AL	1309B000040B017
@005488	@025082006	@CRUZAR LA HUERTA/ LA APARICION/N EN EL TRONCO DE UN A/RBOL DEL	138009004684687
@005489	@025082007	@PARAI/SO. Y LUEGO DIZQUE CORRIO/ A DECIRLE A SU ESPOSO QUE SE LE	138310940428055
@005490	@025082008	@ACABA DE APARECER LA MADRE∆DEL CIELO∆LORENZO @LLAMO/ A LOS	119400078B946
@005491	@025082009	@EJIDATARIOS PA QUE FUERAN TESTIGOS DEL MILAGRO. NO SE/ BIEN CO/MO	1184098781011
@005492	@025082010	@ESTUVO; EL CASO ES QUE CUANDO LLEGUE/ AL RANCHO LA GENTE DE LOS	139689019780040
@005493	@025082011	@ALREDEDORES LLEVABA MESES DE VENERAR A LA SANTI/SIMA VIRGEN∆	9 0084040C0
@005494	@025082012	@–¿Y USTED CO/MO SE ENTERO?	5 35159
@005494	@025082013	@–LA HISTORIA ES UN POCO LARGA, PERO YA QUE INSISTE CON MUCHO GUSTO SE	1400960031004005
@005496	@025082014	@LA CUENTO. AL FIN Y AL CABO USTE/ NO PUEDE ANDAR DE HOCICO/N AI NOMA/S	15007837801094001
@005497	@025082015	@CHIVATEA/NDOME, PORQUE TIENE TAMBIE/N SUS PENDIENTES CON LA AUDITORIDA/,	9 039128400
@005498	@025082016	@¿NO?	1 1
@005499	@025108012	@ES IRREMEDIABLE. PASA TODAS LAS NOCHES. USTED HIZO MUY BIEN EN NO	12900000591141

ne un sólo contador de frecuencia, en cambio el DICCIONARIO en disco contiene contadores de frecuencia para cada género. Por último, los valores 1510 corresponden al primero y al último registro de LIGAS en donde están almacenados los números de las líneas del CORPUS donde la palabra apareció (en este caso son iguales porque la frecuencia es uno). Véase ilustración 12.

Las 1,084,145 ocurrencias etiquetadas automáticamente se redujeron a 31 022 diferentes tipos en DICCIONARIO y, una vez incorporadas todas las palabras <ambiguas>, el DICCIONARIO creció hasta tener 65,200 tipos diferentes.

Anteriormente se presentó una muestra del CORPUS ya procesado. En la serie de numeros de la derecha, los dos primeros dígitos corresponden al contador de palabras en la línea y los siguientes son las categorías gramaticales respectivas de cada palabra en la línea (ilustración 13).

Una vez procesado todo el CORPUS y construido el DICCIONARIO en disco ordenado alfabéticamente (el DICCIONARIO definitivo es la mezcla de los últimos DICCIONARIO y DECISIONES), se calculan algunos valores estadísticos[7] Estas frecuencias, estadísticamente valoradas, se usan como un criterio para elegir las entradas del DEM. Para cada entrada seleccionada en DICCIONARIO, se produce la lista de sus contextos en CORPUS (recuperados a través de LIGAS) para que los lingüistas las empleen como concordancias en el momento de la composición

[7] Véase Roberto Ham, "Del 1 al 100 en lexicografía" en este mismo volumen sobre el funcionamiento de las medidas estadísticas obtenidas.

de artículo. A continuación se presenta la lista de concordancias del vocablo *apellido* ya analizada por uno de los lingüistas (ilustración 14).

El sistema se programó en la versión ALGOL 60 de la Universidad de Noruega (NUALGOL) para la computadora UNIVAC 1106 del Centro de Procesamiento "Dr. Arturo Rosenblueth", que tiene 262 K palabras de 36 bits en memoria central y 800 millones de caracteres en disco.

El sistema que he descrito es indudablemente simple. La labor de pre-edición que tuvieron que hacer los lingüistas se redujo, de hecho, a la selección y clasificación del corpus, ya que la codificación quedó en manos de personal no especializado. El trabajo esencial se refiere al estudio de toda la heurística de precedencia, pero la información que alimenta al sistema es poca. La solución manual de las palabras <ambiguas> es el trabajo más lento y tedioso, sin embargo, quedó concluida en un lapso de año y medio. Tomando en cuenta los resultados obtenidos sobre el porcentaje de solución, pienso que para aumentar el rendimiento habría que alcanzar un nivel de complejidad lingüístico-computacional mayor, cuyos resultados hubiesen tardado mucho más. Por supuesto, los resultados obtenidos son probabilísticos, pero altamente correctos; aunque no se haya realizado una evaluación detallada todavía, me parece que hemos cubierto las necesidades que nos proponíamos.

Agradezco profundamente la colaboración de Jorge Serrano y Javier Becerra tanto en la programación del sistema como en el procesamiento de los datos y la producción de los resultados estadísticos y lingüísticos.

APELLIDO
030239011 SITUACIO/N PAREJA.
1.- 030253025 NI NOMBRE ES NICOMACOD; MI APELLIDO, FLOR
030253026 EN MITAD DEL INVIERNO; CUANDO CALLADA, A
APELLIDO
044075007 VOLUNTAD AVANZAS Y TROPIEZAS Y SIN DECIDI
2.- 044075009 FUERZA DEL APELLIDO, EL U/LTIMO DE LA EST
044075009 TODAS LAS MUJERES PERDIDO, Y EN TODOS LOS
APELLIDO
134038087 DESATA MALES MAYORES QUE LOS QUE SE PROPO
3.- 134038089 PORQUE PARALELAMENTE A LA EMPRESA DE PANT
134038089 UNA SUGERENCIA SONORA, QUIZA/ INCONSCIENTE
APELLIDO
192013014 E/L CON DOS PALABRAS: 'REYA LOPITOS'. 'L
4.- 192013015 APELLIDO COMU/N, QUE PODI/A PERTENECER AL
192013016 TAN BIEN FUGA+O/ A LA MUJER QUE LO HABI/A
APELLIDO
494118023 SE+ORA DE SEPU/LVEDAD. SO/LO IBA/ DE MAYU
5.- 494118024 CORRESPONDA AL APELLIDO DEL MARIDO: 'SE+O
494118025 DELO RI/OO, SE+ORA DE DEO LA PE+AO'.
APELLIDO
603038189 DE NO ESTAR CASADOS...¿QUE/ LEGALIDAD LOS
6.- 603038190 SIN NO ESTA/N RECONOCIDOS COMO TALES, NI
603038191 HABLAR DE LA PARTE ESPIRITUAL, TAN IMPORT
APELLIDO
672077284 PAULO, PERTENECIENTE A LA DIFERENCIA DE CLASE
7.- 672077285 MUCHA IMPORTANCIA. POR OTRA PARTE, EL APE
672077286 COSA. SALVO, COMO ME OCURRI/A A MI/, QUE
APELLIDO
672077292 AQUELLA MISMA TARDE ESTUVE A PUNTO DE DEC
8.- 672077293 RICOS Y POBRES, ¿NO? MI PADRE ES RIQUI/SI
672077294 TIENE OTRO QUE CASI ESTREMECE. NUNCA TE A
APELLIDO
672077294 TIENE OTRO QUE CASI ESTREMECE. NUNCA TE A
9.- 672077295 MENOS QUE SEAS UN POTENTADO, O TU APELLIDO
672077296 CLARO, NO SE LO DIJE.
APELLIDO
745003019 ENTONCES ARTUROO ME DICE, LO QUE SI/ ME A
10.- 745003020 SU PRIMER NOMBRE, PERO ME ACUERDO DE SU A
745003021 SA/NCHEZO... BUENO, PUES A/I ME TIENES, Q
APELLIDO
829003095 ERA UNO DE UNA TIENDA MUY GRANDE; VIVI/A
11.- 829003096 ESTE... ¿DE QUE/? NO ME ACUERDO DE SU APE
829003097 TIENDA TAMBIE/N GRANDE. AQUI/ ENTRABA TOD
APELLIDO
917103040 HIJO. PA POCO ME CORRIO/ CUANDO ME FUI A
12.- 917103041 QUINCE DI/AS Y YO ESTABA MA/S CHICO. ¡POR
917103042 ¡SI E/L ES CAPAZ DE NEGARME COMO SU HIJO
APELLIDO
917103040 HIJO. PA POCO ME CORRIO/ CUANDO ME FUI A
13.- 917103041 QUINCE DI/AS Y YO ESTABA MA/S CHICO. ¡POR
917103042 ¡SI E/L ES CAPAZ DE NEGARME COMO SU HIJO

ITASO. HACI/ A LA MEDIANOCHE,
'CURAS, CON GRAN DESOLACIO/N

PROSIGUES GUIADA POR LA
PE Y EL QUE TERMINO/, EN
OMBRES CLAUSURADO. NO, NO

'/A SOMETER, STNO ADEMA/S
AQ (¿NO HAY EN EL APELLIDO
DE PANDORA?) SE DESAPROLLA

ITOSO, PORQUE ERA UN
ADRE QUE NUNCA TUVO Y QUE
AUTIZADO COMO JOELO OMARO Y

/CULA LA 'ARTI/CULA QUE
DE DEO SALESO, SE+ORA DE

MPARA? NI SIQUIERA A VECES,
APELLIDO... ESTO, SIN
TE EN UN HIJO DESDE QUE

AQUELLO NO DEBI/A TENER
IDO SHOREO NO ERA GRAN
TUVIERA CUBIERTO DE ORO EN

LE: 'PUEDE HABER SHOREO
, Y DETRA/S DE ESE APELLIDO
ITTIRA/ COMO MARIDO MI/O, A

ITTIRA/ COMO MARIDO MI/O, A
SUPERE TU FALTA DE FORTUNA'.

ERDO DICE, NO ME ACUERDO DE
LLIDO, SE APELLIDA PE/REZO
AL DIRECTORIO Y PE/REZO

NO VICENTEO, ESTE... QUE...
IDO, E/STE TENI/A OTRA
LA AZU/CAR DE LA HACIENDA

APULCOO?... Y ESO QUE FUERON
SO ME QUITE/ SU APELLIDO!
LO NIEGO COMO PADRE!'

APULCOO?... Y ESO QUE FUERON
SO ME QUITE/ SU APELLIDO!
LO NIEGO COMO PADRE!'

identificación después del nombre

casta, familia

-nombre, identificación

identificación, padre

del marido, mayúscula

nombre, legal

nombre

representación

representación, casta

1ºs nombre/apellido; identificación

identificación

casta, pertenece

Del Análisis Semántico en Lexicografía

Luis Fernando Lara

"Donde la rosa no es ya sino el nombre sin rosa de la rosa. . ."

*Gilberto Owen**

0. La práctica del análisis semántico es uno de los atolladeros típicos de la lingüística moderna. Las concepciones teóricas que ofrece la lingüística interesada por el tema –pues evidentemente las posiciones bloomfieldianas no tienen nada que ofrecer a un estudio del significado declarado inexistente– han sido más de principio que realmente desarrolladas y se enfrentan a un gran número de dificultades. Pareciera como que el instrumental estructuralista, de probado éxito en fonología, requiere o bien de una serie de reducciones de la complejidad del significado, o bien de un replanteamiento general de sus tesis y de sus métodos.

Las reducciones operan, sobre todo, cuando de la significación total de un signo lingüístico[1]

* Este epígrafe se lo debo a Francisco Segovia.

[1] Entiendo por "signo lingüístico" cualquier segmento de la lengua natural, aislado mediante la concurrencia de varios criterios tanto del plano de la expresión (sustitutibilidad, permutabilidad, unidad fonológica, etc.), como del contenido (cohesión, conectividad, unidad autosémica, etc.); por lo que las "palabras" son solamente una clase particular de signo (cf. K. Heger, *Monem, Wort, Satz und Text*, Tübingen, 1976; en adelante, *Heger 76a)*. Sin embargo, en la gran mayoría de las obras filosóficas, lexicológicas y semánticas, "signo lingüístico" es ambiguo: aunque se pueda entender en la misma forma en que lo hago, por razones teóricas –como se verá a lo largo de este trabajo– así como por razones prácticas de manejo y ejemplificación, "signo" se confunde con "palabra". Puesto que en este artículo se trata de lexicografía, en varios lugares es posible entender por "signo" la "palabra", pero con la salvedad de que el criterio general sigue siendo válido.

se aisla exclusivamente su función referencial de acuerdo con el modelo de Bühler, para lograr así una semántica sistemática y referencial que cuidadosamente eluda los problemas de la connotación y la metáfora. Operan también cuando, de la variación corriente de los signos en el habla diaria, se discrimina metódicamente todo lo que no pueda constituirse en *lengua funcional,* es decir, todo lo que pueda responder a diferencias de nivel de lengua, de estrato social, de región geográfica o de estadio histórico de una lengua. Operan, por último, cuando se distingue una estructuración interna de los significados de los signos respecto de estructuraciones atribuibles a determinaciones extralingüísticas, como los términos de parentesco y otras terminologías. Estas reducciones no son de un orden limitadamente metodológico, sino que obedecen al planteamiento de objetivos propio de la lingüística del *sistema,* fin último del estructuralismo.

Si es relativamente fácil cumplir con ellas en análisis bien controlados de ciertos signos lingüísticos, desde el momento en que, como en lexicografía, el problema práctico se plantea con todos sus detalles, al lexicógrafo no le queda más que retirarse tristemente del campo acotado por la lingüística moderna, o ignorarla por completo con justificaciones estrictamente prácticas.

Pero ninguna de estas salidas puede satisfacernos. Todavía queda en pie la creencia de que, si la semántica lingüística y la lexicografía tienen en común una cierta clase de signos llamados "palabras", debe ser posible tratarlos coherentemente en la teoría y en la práctica de ambas disciplinas.

Este trabajo tiene ese objetivo. Sobre una dis-

cusión de principio acerca de las causas de las reducciones estructuralistas, se intenta replantear la semántica con la ayuda de observaciones procedentes de la práctica lexicográfica, para así relacionar teoría y ejercicio.

1. Algunas hipótesis de partida.

1.1. La autocontención del sistema.

Una de las enseñanzas que ahora ya se pueden sacar de la historia cercana de la lingüística es de importancia crucial para el desarrollo de este trabajo. Si se intenta comprender las características profundas del estructuralismo lingüístico, dentro del cual se ha constituido la lingüística que aún hoy ejercemos, se verá que mediante la noción de "sistema" que con tanto éxito introdujeron en la lingüística del primer cuarto del siglo XX los iniciadores Saussure y Bloomfield, también se hacía necesario un postulado epistemológico de orden mayor, consistente en la necesidad, defendida por Saussure, de que la lengua se estudiara en forma autónoma, en contraposición con las maneras como hasta su época se había hecho. Pienso que es posible sostener esta afirmación si se toman en cuenta las implicaciones que conlleva la noción de "sistema" en el pensamiento saussureano (que, respecto de las características generales del estructuralismo lingüístico, pueden también adjudicarse al descriptivismo bloomfieldiano).

El concepto de "sistema" se presentaba como una solución verdaderamente revolucionaria a los más difíciles problemas de la lingüística neogramática; por un lado, liberaba a los lingüistas de sus ataduras a la documentación histórica de las entidades lingüísticas, es decir, abría el paso entre una metodología típicamente filológica y

la utilización de métodos más modernos de la ciencia, como la abstracción de invariantes a partir de registros fragmentarios y variables, o como la inferencia de elementos necesarios al sistema, independientemente de que se hubieran atestiguado en la historia (creo que ese es el sentido de la tesis saussureana sobre las vocales indoeuropeas, tanto para la totalidad de su obra, como para la lingüística que él inició); por otro lado, también permitía superar las limitaciones de una lingüística encadenada a la sustancia (en el sentido hjelmsleviano del término) de los sonidos y las letras, pasando a la noción de "fonema" y por lo tanto al interés por los aspectos formales de las lenguas[2]. El concepto de "sistema", en consecuencia, habría de ser el meollo de la teoría lingüística estructuralista. Así puede entenderse el porqué epistemológico de la dicotomía de *lengua* y *habla.* Respecto de la realidad histórica y social del habla –pero también realidad fragmentaria y variable– se hacía necesario teóricamente construir un modelo invariante y abstracto dentro del cual se pudiera operar en independencia de aquellas viejas concepciones historicistas o psicologistas que, en palabras de Hjelmslev, trascendían de la lengua al resto de las ciencias del hombre. Pero para lograrlo, diría Saussure, habría de estudiarse la lengua como objeto autónomo, la lengua por ella misma, como un sistema "où tout se tient". A partir de ese momento habría de ser el sistema mismo el que definiera sus elementos; estos habrían de concebirse como términos de una red de relacio-

[2] En un artículo de divulgación sobre "De Saussure, Chomsky, el ajedrez y los toros" aparecido en *Diálogos* 72 (1976), 14-19 hice un esbozo de las diferencias entre las revoluciones de Saussure y Chomsky en que se detalla más este punto.

nes de las cuales obtendrían su valor y, por lo tanto, su identidad teórica.

Bajo tal concepción del sistema, entonces, se aisla la lengua y se postula la autocontención de ella misma. Las tareas de la lingüística quedarían desde entonces definidas por ese postulado.

1.2. Las doctrinas sobre la naturaleza del signo[3]

Hay una segunda consecuencia que se saca de lo anterior, aunque no se explique solamente a partir del postulado antes indicado, sino especialmente a partir de otra de las contribuciones decisivas del pensamiento saussureano. Me refiero a su teoría del signo y a sus apuntes para aquella ciencia que habría de llamar semiología.

Los signos lingüísticos hasta su época —y todavía hoy— se habían concebido dentro de los cauces del pensamiento filosófico. Los antiguos debates heracliteanos y sofísticos respecto de la relación entre los signos y las cosas del mundo se habían perpetuado a lo largo de la historia de la filosofía hasta llegar a oponer dos doctrinas que, en cierta forma, los resumen: aquella que sostenía la participación de los signos en la esencia de los objetos que representan —simbolizada por Cratilo en el diálogo platónico del mismo nombre—[4] y aquella que, por el contrario, propone la

[3] Una primera versión de esta parte está en mi artículo "Una base semántica para la lexicografía: la conceptualización del signo lingüístico", *NRFH*, 26, 2 (1977), 261-275.

[4] La bibliografía que pude consultar sobre la obra de Platón no se ocupa casi del *Cratilo;* generalmente sólo destaca la ironía de Sócrates, pero no analiza el sentido del diálogo respecto de la totalidad de las obras platónicas. Cf. K. Jaspers, *Plato and Augustine, The great philosophers I,* Harvest Books, New York, 1957, P. Friedländer, *Plato. An introduction,* Princeton U. Press, 1958, F. Copleston, *A history of philosophy,* t. I, 1, Image Books, Maryland y B. Parain (dir.), *La filosofía griega. Historia de la filosofía,* Siglo XXI, México, 1972.

naturaleza convencional de los signos respecto de las cosas[5] —representada por Hermógenes en el diálogo, en una versión un tanto diferente de la que sostuvo más tarde Aristóteles y que es la que guía estas reflexiones.

Cabe aclarar previamente que de la gran complejidad de cuestiones que aparecen ligadas con la de la naturaleza del signo lingüístico y que tocan desde la metafísica y la ontología hasta la lógica y la gramática, sólo me referiré a los aspectos más generales de la discusión. Al mismo tiempo también hay que señalar que, evidentemente, el pensamiento filosófico posterior al de la antigüedad clásica ha enriquecido sustanciosamente el tema, por lo que atribuir lo que sigue a "la filosofía" no es más que una referencia de lugar, que no hace realmente justicia a la filosofía.

La doctrina de la motivación natural del signo (que es como corrientemente se denomina en los cursos de semántica) parece fácilmente superable y superada. Las razones y las burlas etimológicas de Sócrates en el diálogo hacen obvio de inmediato que no hay nada en los signos que pertenezca a la esencia de las cosas que representan[6]. En comparación con ella la doctrina aristotélica aparece como una ganancia definitiva en la conceptualización de la naturaleza del signo y,

[5] No es mi intención precisar términos como "cosa", "objeto", "referente" o "mundo" en el sentido de las discusiones metafísicas u ontológicas. Al hablar de cualquier objeto conocible, sea una idea, un concepto, una cosa material, etc. con esas palabras lo hago en la forma más neutral posible.

[6] En un libro sobre semiología y crítica literaria que prepara actualmente Tomás Segovia, explora profundamentete un sentido de esta relación que no hemos tomado en cuenta. A él le debo muchas de las ideas aquí reunidas. Más tarde la confrontación entre su pensamiento y este trabajo me permitirá definir mejor este punto.

con las diferencias necesarias, como una posición todavía vigente en términos generales.

Pero lo que me interesa resaltar es una característica común a las dos doctrinas en la forma como plantean la relación entre el signo y su referente. Para ellas (como para tendencias contemporáneas, como el positivismo lógico, la filosofía analítica y la filosofía del lenguaje ordinario) el signo solamente interesa en cuanto vehículo para formar enunciados sobre el mundo conocible. Los signos son nombres, etiquetas de las cosas. En la doctrina de la motivación la etiqueta participa del ser de la cosa; en la de la convencionalidad la etiqueta se asocia con la cosa según la convención social que estipule la relación. En casos extremos, como los de ciertas corrientes empiricistas, los signos apenas son cadenas sonoras que prestan una forma material a los estímulos, o a las imágenes, o a las percepciones, o a los conceptos de las cosas. En todas estas concepciones el referente o la cosa es lo central, mientras que el signo es secundario y depende del referente.

Cuando el objeto central de las doctrinas es el mundo, y el lenguaje (el signo) no vale sino en cuanto vehículo para hablar del mundo, es natural que se quiera ver en los signos un reflejo inmediato de sus referentes y, por lo tanto, que al reparar en el uso cotidiano de los signos, se les vea como causantes permanentes de equívocos sobre el conocimiento de sus referentes. La lengua natural, en consecuencia, rápidamente pasa a ser materia de crítica y se buscan los mecanismos necesarios para eliminar de ella los errores que suele producir en los enunciados sobre el mundo. En algunas líneas del *Cratilo* se llega al extremo de proponer que, puesto que los signos

falsean el conocimiento, habrá que eliminarlos y hablar de las cosas con las cosas mismas.

Se diría que es esa constatación la que origina la lógica. La misión de la lógica es precisamente la de eliminar las posibilidades de error que aparecen en los enunciados hechos con lengua natural, sometiéndolos a un control "racional" adecuado a ciertas categorías universales del pensamiento.

Así es que una de las principales necesidades de la lógica es la de definir bien y de manera unívoca los términos que aparezcan en los enunciados. Los signos del enunciado lógico deben corresponder uno a uno a los objetos de los que se habla y no deben alterar sus relaciones sin antes anunciar el cambio. Dicho en forma esquemática, a cada signo debe corresponder un solo objeto y a cada objeto un solo signo. Este es el ideal de la denotación y se propone hacer desaparecer la contradicción en la lengua, impedir la variación de los significados y, a fin de cuentas, hacer del léxico una nomenclatura bien estipulada en la que todo el grupo social esté de acuerdo. Es la visión de la lengua que Engler llama nomenclaturista[7].

[7] Cf. R. Engler, "European structuralism: Saussure" en T.A. Sebeok (éd.), *Current trends in linguistics, 13: Historiography of linguistics,* 1975, pp. 829-886 (En adelante Engler 75). En n. 12, p. 835 formula la diferencia entre el punto de vista nomenclaturista y el saussureano como sigue: "Nomenclatory view:

Objet →Concept
Concept →Sign (= Name, or = Name + concept)
Sign + Sign →System.

Saussurian view:

System → Sign x Sign
Sign →Signifiant x Signifié (or: Concept x Name)
Sign ~Object.

I.e. The signs are dependent on the system, *signifiant* and *signifié* as elements dependent on systemic signs and the signs (not the concepts) are applied to (not derived from) extralinguistic objects."

La doctrina de la convencionalidad de los signos respecto de sus referentes se puede explicar dentro de esa característica del pensamiento ontológico y lógico. Tiene la gran ventaja de que separa el signo del referente totalmente; establece una diferencia determinante entre los objetos y los signos que los representan. En ese sentido difícilmente se podría alguien oponer a ella y en ese mismo sentido hay que considerarla como una de las tesis fundamentales para la lingüística anterior a Saussure.

El problema queda, en cambio, en el carácter *convencional* de la relación signo-referente. En apariencia elimina el origen misterioso de los nombres, atribuyéndolos a una especie de contrato social entre los miembros de una comunidad. La idea de que la naturaleza del signo es convencional es una rápida respuesta para las interrogantes que plantea la relación entre las cosas y sus nombres. Si hay convención, la relación es pensable y razonable; se comprende que, por ejemplo, en lenguas distintas haya nombres distintos para las mismas cosas; se comprende que haya cambios de la relación, aunque no sea tan fácil comprender por qué y cómo se dan esos cambios. En la posibilidad del cambio es en donde la doctrina de la convencionalidad muestra sus debilidades. Si los signos fueran convencionales sería igualmente posible o no cambiarlos en absoluto –pues se podría convenir en ello– o cambiarlos de inmediato y totalmente. Sucedería lo que con otras convenciones sociales: que nada en ellas se muestra necesario y basta un pequeño acuerdo para sustituirlas, como en el caso de las leyes o de las modas.

Ahora bien, los cambios se dan pero en una forma oscura y sin que se vea la actuación de

una convención clara; aun cuando se desee conservar un signo sin cambio, a la convención inicial se opone un uso social que la transgrede poco a poco hasta cambiarla. Igualmente, si se desea provocar un cambio, no basta un acuerdo para que se realice, sino que parece haber tendencias más fuertes y más inexplicables hacia la conservación. Estos fenómenos de la lengua natural son inexplicables a partir de la teoría de la convencionalidad del signo.

Pero en realidad hay que comprender la teoría de la convencionalidad del signo en términos de la exigencia nomenclaturista del pensamiento lógico. Bajo esa concepción los signos lingüísticos se miden de acuerdo con un patrón ideal de lengua, universal, ordenada y lógica, como la "characteristica universalis" de Leibniz o como los repetidos intentos por construir lenguas artificiales como el volapük, el esperanto o la interlingua. Lo que hay que recordar ahora es que tal ideal no es una imagen de la lengua natural sino lo contrario, es una imagen de una lengua que no sea como la natural, pues lo que se busca es librarse de la variabilidad incontrolada y del aparente capricho que se encierra bajo cada metáfora. La doctrina convencionalista no puede, por eso, concebir una lengua histórica para la que el cambio sea inherente a su naturaleza, pues el movimiento de la tradición se opone a la sucesión puntual de convenios lógicamente necesarios pero, como dice Saussure, nunca reales.

Saussure trataba de salir al paso de la concepción convencionalista con su concepto de la *arbitrariedad*. Evidentemente, también en el convencionalismo hay una arbitrariedad inicial, pues nada obliga a la comunidad a nombrar un objeto

en una forma y no en otra. La diferencia entre ambas posiciones no radica en ello, sino en el hecho de que, para los convencionalistas, tras la arbitrariedad inicial aparece el contrato social que *estipula* las relaciones entre el signo y el referente y que, de esa manera, asegura la biunivocidad de la relación. En cambio, bajo la teoría saussureana de la arbitrariedad, esta se conserva siempre en forma radical en el horizonte semiológico de la lengua. Quiero decir con ello que la arbitrariedad saussureana no deja de actuar en un momento dado, sino que está siempre mediando todas las relaciones entre la lengua y el mundo sensible. Es un motor en permanente funcionamiento para la significación y es el pivote respecto del cual se explica, tanto por qué cambian las lenguas, como por qué no cambian.

Parece que la enseñanza saussureana sobre este tema no ha sido entendida así por la vulgata que se ha venido haciendo a partir del *Curso de lingüística general.* Ha tocado a Tullio de Mauro y Rudolf Engler[8] comenzar la exploración de las fuentes saussureanas en ese sentido. Pero basta leer los capítulos y párrafos del *Curso* en que se trata la arbitrariedad, para notar la manera decidida como Saussure se oponía a la visión convencionalista y nomenclaturista del signo.

En realidad, el concepto de la arbitrariedad del signo tiene varios niveles de funcionamiento. Lo primero que se obtiene con él es la crítica de las dos posiciones simbolizadas por el *Cratilo* —como también señala Nicola Abbagnano en su

[8] R. Engler en su inmensa edición crítica del *Cours,* publicada en O. Harrassowitz, Wiesbaden, 1967, y muchos artículos publicados en diferentes lugares; Tullio de Mauro en su edición del *Cours* publicada por Payot, 1975, en traducción francesa.

Diccionario de filosofía[9] –; después, se convierte en un concepto clave para explicar la manera como actúan la historia y la tradición sobre la lengua, más allá del valor teórico de la dicotomía *sincronía/diacronía*[10]; pero lo que más interesa en este trabajo es la forma como se presta –en una interpretación básicamente saussureana pero ya no literalmente suya– para comprender la relación signo-referente.

Tanto la doctrina de la motivación natural de los signos en sus referentes como la de la convencionalidad exigen el establecimiento de una relación de uno a uno entre ambos. Si en el primer caso la relación está dada por la esencia de las cosas, en el segundo es fruto de un trabajo humano y racional que antes ha establecido sus necesidades de conocimiento y luego las imputa a la lengua natural. Bajo cualquiera de los dos puntos de vista la relación entre signo y referente se tiene que concebir como dada a partir del momento en que míticamente se estableció *por primera vez*. La semántica, por lo tanto, habrá de contar con que la relación existe siempre, pues nunca le será dado registrar o al menos especular sobre ese primer momento de la denominación.

Para la teoría de la arbitrariedad eso no es necesario. Al negar la convención y al negar la motivación como orígenes de los signos, propone

[9] Cf. N. Abbagnano, *Diccionario de filosofía*, México, 1966, s.v. *lenguaje*, donde propone la posición saussureana como una alternativa al debate clásico.

[10] Tengo la impresión de que ha sido la vulgata saussureana la causante del radical antagonismo entre *diacronía* y *sincronía* y no la obra misma de Saussure donde, además de que la dicotomía radical es del orden de la segunda metalengua y no del de la teoría del lenguaje, hay pasajes en que Saussure ofrece luminosas conceptualizaciones de la historia en su relación con la teoría que estaba desarrollando.

una aleatoriedad permanente en las relaciones con los objetos del mundo sensible. No es que haya un primer momento de denominación de las cosas mediante signos, como implícitamente sostienen las doctrinas de la motivación y de la convencionalidad; aunque un cierto pensamiento racional exija la existencia de razones, de causas o de orígenes, no es posible documentar el "nacimiento" de *la* lengua ni como el acto de denominación de los objetos en el *Génesis*, ni como los instantes en que, dice Sócrates, los legisladores que pueden dar nombres a las cosas establecen los signos originarios, ni como un acuerdo entre los miembros de una sociedad. Por el contrario, "à n'importe quelle époque et si haut que nous remontions, la langue apparaît toujours comme une héritage de l'époque précédente. L'acte par lequel, à un moment donné, les noms seraient distribués aux choses, par lequel un contrat serait passé entre les concepts et les images acoustiques —cet acte, nous pouvons le concevoir, mais il n'a jamais été constaté. L'idée que les choses auraient pu se passer ainsi nous est suggerée par notre sentiment très vif de l'arbitraire du signe." (*CLG*, p. 105) Es, según Saussure, la presencia permanente de la arbitrariedad del signo la que, una vez victoriosa en la especulación filosófica sobre la creencia de que los signos tienen un naturaleza participatoria en sus referentes, obliga a la razón a proponerse otra causa posible de la relación signo-objeto, cuando, como se ve, esa causa no es como la imagina la lógica, sino que es arbitrariedad pura. La biunivocidad de la relación que propone la doctrina de la convencionalidad es un resultado de las necesidades de la ontología en la representación verbal de sus objetos de conocimiento, pero no

un fenómeno que se compruebe en el uso diario de las lenguas. Lo que se comprueba es lo otro: que los signos tienen relaciones muy variadas y variables con los objetos representados; que siempre, a toda relación rígida que uno estipule, se le podrán encontrar, al cabo del tiempo, transgresiones, transformaciones, cambios. Así, la convencionalidad del signo es solamente una teoría sobre su naturaleza imputada desde el horizonte de la lógica y la ontología. Es —me parece— el resultado de un largo proceso en la historia de la filosofía empeñado en aclarar la relación entre los nombres y las cosas, pero no por reconocer la particularidad de los nombres en sí mismos (o sea, la naturaleza de la lengua natural), sino por someter su arbitrariedad a la razón lógica.

Lo que propone Saussure con su concepto de arbitrariedad no es, entonces, una explicación del porqué de los signos, ni de la causa de su nacimiento, sino una no-causa: la simple comprobación de que los signos se relacionan al azar y arbitrariamente con sus referentes. De esta manera no es posible proponer un primer momento de denominación de las cosas, ni tampoco imputar a la lengua la rigidez de las nomenclaturas, sino que habrá que aceptar que hay un proceso permanente de relación entre signos y objetos —el proceso de la *significación*—, que es aleatorio y que gravita sobre la radical arbitrariedad de los signos lingüísticos.

1.3. Arbitrariedad y sistema.

Esta concepción de la arbitrariedad no podría darse sin que el postulado anteriormente explicado de la autocontención del sistema lingüístico hubiera preparado el terreno. Solamente después de que se hubo aislado la lengua como obje-

to de estudio autónomo, solamente después de que se la desligó de sus "causas" históricas o psicológicas, se habría creado el espacio necesario entre lengua y mundo (signo y referente) que permitiera un trabajo crítico sobre su relación.

Es verdad que, históricamente, la posición estructuralista ha sido parte del movimiento positivista y que, más que aprovechar la existencia de ese espacio entre sistema y mundo para analizarlo, el estructuralismo objetivó (cosificó, en algunas tendencias) la lengua para poder aplicar en ella los principios de objetividad, neutralidad y observación que las ciencias naturales positivas proclamaban en su época como ideales del pensamiento científico. Sin embargo, la reivindicación del estudio de la lengua por la lengua misma permitió pasar de la idea de que los signos son dependientes de los objetos que nombran, a la contraria, que destaca el papel primario del signo sobre el referente en lingüística y que, como se ve en la semántica de Ullmann, aún permitió sostener que el referente, en cuanto tal, no es objeto de consideración en la ciencia del lenguaje.

La necesidad de hacer una lingüística autónoma, el surgimiento en lingüística del concepto de "sistema" y la teoría de la arbitrariedad del signo son, por lo tanto, elementos axiomáticos de la lingüística saussureana que no pueden separarse si se quiere conocer su perfil epistemológico. Al mismo tiempo, en la medida en que la lingüística estructuralista se conciba como un desarrollo del pensamiento de Saussure o simplemente se adhiera a la mayor parte de sus postulados, en esa misma medida puede volverse objeto de una crítica que parta de un nuevo análisis de esos tres axiomas. En el caso presente se trata de

plantear el sentido en que o dan lugar al análisis semántico práctico o lo obstaculizan.

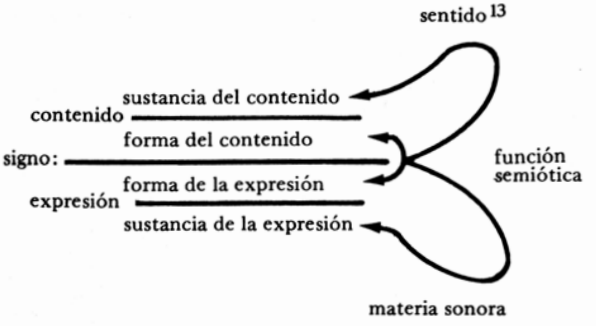
1.4.1. Significación, signo y referente.

Como se decía arriba el concepto de arbitrariedad tiene varios niveles que se deben considerar antes de aplicarlo al análisis semántico concreto. Para pasar del orden semiológico en que se ha instituido, es decir, de un nivel metalingüístico en que solamente se funda axiomáticamente la teoría de los signos, al orden lingüístico en que se considere un sistema lingüístico determinado, hace falta profundizar sobre la naturaleza de los signos en la lingüística estructural. Para ello me serviré del conocido estudio de Louis Hjelmslev sobre "La estratificación del lenguaje"[11], que considero la más rigurosa explicación del signo desde el punto de vista estructuralista.

Significado y *significante,* las dos caras del signo saussureano, se designan ahora como *contenido* y *expresión* respectivamente. Ambos planos en la concepción de Hjelmslev deben dividirse en cuatro estratos, con el objeto de reconocer con mayor claridad la distinción entre *forma* y *sustancia* que queda implicada en la lingüística saussureana[12]. Así, se puede representar el signo lingüístico con el siguiente esquema:

[11] L. Hjelmslev, "La stratification du langage" en *Essais linguistiques, Travaux du Cercle Linguistique de Copenhague,* IX, 1959, reeditados por Minuit, París, 1971. En versión española *Ensayos lingüísticos,* Trad. E. Bombín I. y F. Piñero Torre, Gredos, Madrid, 1972.

[12] Es un debate clásico del movimiento saussureano si la oposición entre *forma* y *sustancia* debe interprestarse, como Hjelmslev, hacia la *forma* como principal objeto de la linguística o, como el círculo de Praga, hacia el reconocimiento de la *sustancia* como determinante última del signo. Puesto que aquí sigo a Hjelmslev, la *forma* es la que se destaca.



De acuerdo con el postulado de autonomía del sistema lingüístico, la naturaleza del signo como sistema es formal; es la forma la que no varía y la que encauza, delimita y da valor a la sustancia fonética o semántica. La sustancia del contenido y de la expresión *depende* del juego de relaciones sistemáticas que se dan entre las formas que constituyen el sistema. Así por ejemplo, en cuanto a la expresión (que es en donde es más fácil explicar las distinciones), lo que identifica a los sonidos que pronunciamos es el fonema, una forma cuyo valor se obtiene de las relaciones que guarde con otros fonemas. En español el fonema /p/ vale por sus relaciones con /b/, con /t/, con /k/, con el resto de los fonemas y con las relaciones que tengan los otros fonemas

[13] En los *Prolegómenos* (cito de la traducción francesa, Cap. 13 p. 74) define el *sentido* diciendo: "Une expérience qui par contre, semble justifiée, consiste à comparer différentes langues et à en extraire ensuite ce qu'il y a de commun à toutes, et ce qui reste valable dans tous les cas, quel que soit le nombre de langues que l'on considère (. . .) on decouvre que ce facteur commun est une grandeur qui n'est définie que par la fonction qui la lie au principe de structure de la langue et à tous les facteurs qui font que les langues diffèrent les unes des autres. Ce facteur commun, nous l'appellerons le *sens.*" En la versión española es *sentido.*

entre sí. En cambio, la calidad sonora de los fonemas, el detalle articulatorio o acústico de nuestra pronunciación no indica nada acerca del valor estructural del fonema. Se puede pronunciar una /p/ más o menos tensa, más o menos sorda, sin que ello altere su valor fonológico, definido desde el interior del sistema. Es entonces la forma la que identifica los elementos del sistema.

Paralelamente, las cosas suceden igual en el plano del contenido (lo que no quiere decir que ambos planos sean isomorfos). La identidad formal de una preposición, de un tipo de oración o de los morfemas que se combinen en un vocablo, está definida por el sistema, mientras que su sustancia es variable y ha de entenderse siempre como determinada por la forma. En tratándose de la sustancia semántica, los significados de esos elementos han sido seleccionados por la forma.

Entre la forma del contenido y la forma de la expresión no hay una relación de dependencia como la que hay entre sustancia y forma. Entre las dos formas *no* hay sistema, sino que se relacionan *arbitrariamente* por la actividad significativa que Hjelmslev llama *función semiótica.* Esta función se ejerce, hay que recalcarlo, desde fuera del sistema lingüístico, más allá de la lengua, en el horizonte semiológico.

La función semiótica no puede explicarse desde el interior del sistema lingüístico porque, si así fuera, implicaría que entre contenido y expresión hubiera alguna relación permanente; dicho de otra manera, implicaría que o bien algo del significante determinara algo del significado, o bien la determinación fuera en sentido inverso. La consecuencia de una tal implicación sería que

entre los dos planos no habría diferencia, sino que formalmente fueran uno solo y continuo. Algunos lingüistas han intentado concebir así el sistema lingüístico, muchas veces por una razón práctica de la descripción, pero si se considera bien lo que implica, se verá que el paso del fonema (diacrítico) al morfema (significativo), lleva a la creencia de que también los fonemas participan de la significación; es decir, que /a/ o /s/, o cualquier diacrítico significa algo y ayuda a determinar el significado de un morfema. Es una creencia típica de algunos iluminados y subyace a todos los esfuerzos que se han hecho, por ejemplo, por desentrañar un sentido divino de las letras contenidas en la Biblia.

Entonces la función semiótica se tiene que concebir como otro axioma que se refiere a la actividad significativa de los humanos, a su capacidad de usar y hacer signos sobre la realidad.

Cuáles sean las vías por las que el ser humano logra reunir los dos planos del signo (por el momento sin tomar en cuenta la localización histórica de las lenguas) mediante la actividad de la función semiótica, es materia que corresponde investigar a una psicolingüística o quizá a una neurolingüística.

La función semiótica, según Hjelmslev, no actúa solamente una vez, sino que vuelve a operar al relacionarse entre sí los estratos de la forma con la sustancia de la expresión y los de la forma con la sustancia del contenido. Pero esta operación es nuevamente especial porque ahora no se trata de reunir dos formas sistemáticas como lo son la fonología y la sintaxis (¿o la semántica?), sino de pasar de la forma del sistema a una selección de la sustancia. Para Hjelmslev la

sustancia no tiene existencia teórica independiente, sino que se forma y transforma según los moldes que ofrece la forma. La forma se tiene que aplicar, en consecuencia, sobre alguna materia prima apenas maquilada por distinciones perceptuales y cognitivas que la aíslan de otras materias. A esa materia prima, amorfa según Saussure y según Hjelmslev, la considera este último por el lado del contenido, como *sentido*. Qué sea el *sentido* es ahora la interrogante fundamental de las ciencias del hombre. Para Hjelmslev el *sentido* podría interpretarse como lo concebible, lo pensable, lo imaginable y no como concepciones, pensamientos o imágenes ya dadas. En ese sentido el sentido es amorfo, es solamente una virtualidad de arreglo respecto del mundo que rodea al ser humano.

Visto así el sentido y entendida la lengua como lo hace Hjelmslev, la función semiótica se vuelve a concebir como actividad significativa inherente a la especie humana. La función semiótica parte, entonces, de las capacidades de aprehensión que tiene el hombre (capacidades sobre las cuales la psicología de las formas, los estudios sobre la percepción y la neurología tendrían que ofrecer indicaciones) y se aplica sobre una realidad que se proyecta en un *sentido*. Digamos que el *sentido* es una posibilidad de interpretación intelectual del mundo sensible.

Reconozco el enorme peligro de caer aquí en una metafísica de quinto patio. Más que deseo de sostener que así sean las cosas en realidad, más que verme enfrentado a la posible demanda de localizar el sentido en forma experimental, lo que busco es ampliar el problema de la relación signo-referente para eludir las simplificaciones a que, en este punto, nos tiene acostumbrados la

lingüística. Bajo las teorías convencionalistas el problema no está en esa relación, pues se la concibe lineal, biunívoca y asociativa[14], sino simplemente en la naturaleza del referente; en la cuestión de si el referente es un conjunto de reacciones conductuales en el interior del sistema nervioso —como creían los fisicalistas bloomfieldianos— o si se trata de una entidad "sobrenatural" como la mente o las ideas. En cambio, cuando se desecha la teoría convencionalista, el problema de la significación vuelve a aparecer con la totalidad de su riqueza (y de sus dificultades). Creo que este es uno de los principales efectos benéficos del cambio que implica la noción de arbitrariedad.

Para resumir, la función semiótica actúa implantando la forma del contenido sobre el sentido, al que transforma ahora en sustancia del contenido, en objeto de estudio para la lingüística. La sustancia del contenido, como diría Hjelmslev, se vuelve ahora la determinación positiva, real, de la red formal de relaciones que es la forma, el sistema lingüístico.

Una consecuencia terminológica se obtiene también de lo anterior: en lo que sigue, utilizaré indistintamente *función semiótica y significación* para hablar de la actividad, el proceso de prestar significado a un signo lingüístico relacionándolo arbitrariamente con el *sentido*. El *sentido* es una proyección que la significación realiza por medio de los mecanismos cognitivos propios de la especie humana, de los referentes, que no son

[14] Sobre el concepto de "asociación" y su papel negativo en la filosofía del lenguaje, cf. Bernard Harrison, *Meaning and structure. An essay in the philosophy of language*. Harper & Row, New York, 1972.

cosas en sí, previamente delimitadas y concebidas, sino objetos delimitables y concebibles.

1.4.2. Significado, significante y sentido.

Pasando ahora al siguiente aspecto del nivel metalingüístico correspondiente al sistema de signos, el signo lingüístico se concibe como ya constituido, como producto de la actividad significativa. Del signo se ven sus dos caras: *significado* y *significante,* que son valores estructurales, es decir, se identifican por las relaciones que establezcan con el resto de los elementos del sistema. La significación ha dejado de ser un proceso para quedar plasmada en los *significados* de los signos; los significados se reconocen aplicando sobre los signos la regla fundamental de la conmutación estructuralista.

La arbitrariedad radical ha quedado ahora convertida en una *arbitrariedad de la relación* significado/significante; se refiere al reconocimiento de que, entre las dos caras del signo, no hay dependencia de una a otra, sino solamente reciprocidad.

Este paso del *sentido* al *significado,* que es como se puede simbolizar este punto, requiere a su vez de un cuidadoso análisis. Corresponde a la objetivación del signo por una ciencia descriptiva y, en palabras de Bühler, del *producto lingüístico (ergon).* La actividad significativa se sitúa epistemológicamente en la esfera de las acciones y los actos verbales, es *energeia;* en tanto que el sistema lingüístico, que es en donde los signos valen por su estructura, corresponde a la esfera del producto, de lo ya realizado. Para matizar más aún, corresponde al dominio de la *forma lingüística,* que es un nivel abstracto.

Ha sido esta esfera del *ergon* en donde la lin-

güística estructuralista se ha desarrollado, como han mostrado tanto Karl Bühler como Eugenio Coseriu. Desde que se objetiva solamente esta parte de la lengua, las relaciones del sistema con el sentido se tienen que ver completamente distintas. La función semiótica no se ve en acto sino solamente su resultado; la relación del signo con el sentido se traduce solamente en la relación entre los dos planos del significado y del significante. De ellos se vuelve a decir que están arbitrariamente relacionados.

Pero esta relación arbitraria ya no es tan sencilla de cambiar como aparecía en el nivel semiológico, pues el sistema se aplica sobre el sentido tan claramente, que antes de la libertad del cambio, lo que se ofrece a la visión estructuralista es el orden de los valores estructurales complejos, múltiples, que someten significantes y significados a las exigencias opositivas de la sistematicidad. Quizá así es como deba entenderse la afirmación de Benveniste de que "the sign as a combination of *un concept* and *une image acoustique* is not arbitrary; ≪au contraire, il est *nécessaire*. Le concept ("signifié") 'boeuf' est forcément identique dans ma conscience à l'ensemble phonique ("signifiant") *böf*. . . Ce qui est arbitraire, c'est que tel signe, et non tel autre, soit appliqué à tel élément de la réalité, et non à tel autre. . . Le domaine de l'arbitraire est ainsi relégué hors de la compréhension du signe linguistique≫".[15]

Más abajo volveré a tratar la manera como se imbrican el sistema lingüístico y la significación.

[15] Cito a Benveniste a través de Hennings Spang-Hanssen, *Recent theories on the nature of the language sign*. Travaux du Cercle Linguistique de Copenhague, 1954, § 49, p. 97.

1.4.3. Signo, sentido e historia.

Para la lingüística estructuralista el análisis se detiene en el nivel del sistema; su objeto de estudio se colma con la pura descripción del aspecto sistemático de las lenguas. Hay, por supuesto, una amplia región fronteriza entre lo estrictamente formal o sistemático y lo que muestra características sustanciales, como demuestra el debate clásico entre las escuelas de Copenhague y de Praga; de cualquier manera para ambas, ya fuera con mayor preocupación por la sustancia o en total desapego de ella, lo único pertinente para la lingüística estructural es lo que contribuya a la definición de la estructura autónoma de las lenguas.

Muestra de ese interés exclusivo por lo sistemático pueden ser las otras disciplinas de la lingüística que, con el advenimiento del estructuralismo o en su pleno reinado, se han visto en dificultades para sostener sus intereses o decididamente se han tenido que concebir como disciplinas mixtas y marginales, como en los casos de la dialectología, de la geografía lingüística o de la misma sociolingüística.

La exploración de la manera como el sistema lingüístico se presenta en la realidad inmediata de los hablantes tenía, por tanto, que verse obstaculizada. El sistema como presencia real en una comunidad lingüística se concibe solamente como realización *(habla),* o como actualización *(performance)* y no hay teorías consecuentes para tratarlas. La lingüística del sistema permanece irremediablemente en la idealización del hablante-oyente perfecto, en unidad de registro, en unidad de capa social, en unidad dialectal y en unidad temporal.

Lo que se necesita es, entonces, un instrumen-

to teórico que recupere la realidad social e histórica y supere las reducciones que, como expliqué antes, sirvieron para constituir esta lingüística que todavía hoy hacemos. Contra lo que se supone generalmente, también Saussure dejó abierta esta puerta en los varios lugares de su obra en donde explica que la *lengua* sin sociedad y sin historia no puede entenderse cabalmente[16].

Según Engler este último nivel es el idiosincrónico, en donde la lengua aparece inmersa en la historia y sostenida por la tradición de una comunidad lingüística. Para un hablante la lengua solamente puede aparecer como una herencia de su sociedad; los signos no son fáciles de alterar o de concebir siquiera si no es como un sistema de comunicación previamente convenido y orientado por normas que poco a poco va aprendiendo en el interior de su grupo social. La lengua no es más algo que un hablante pueda cambiar a voluntad; es, en el peor de los casos, una imposición; en el mejor, una obligación compartida, como dice Coseriu. Además, para su visión de la lengua no hay un juego de reglas y un conjunto de elementos que pueda utilizar indiscriminadamente, sino que cada uno de ellos tiene diferentes dimensiones de validez, determinadas por la sociedad en que vive. Así es que la lengua se ofrece como "motivada" por todas esas condiciones. Los significantes y los significados no se cambian sin antes haber pasado por el tamiz de la sociedad, tamiz del cual el hablante individual

[16] Por ejemplo en el *Cours* (edición De Mauro), Cap. II. 1, p. 108: "Si la langue a un caractère de fixité, ce n'est pas seulement parce qu'elle est attaché au poids de la collectivité, c'est aussi qu'elle est située dans le temps. Ces deux faits sont inséparables. (...) C'est parce que le signe est arbitraire qu'il ne connait d'autre loi que celle de la tradition, et c'est parce qu'il se fonde sur la traditon qu'il peut être arbitraire."

no llega a percibir sus fronteras; para cada significante hay un significado que comprende y acepta la sociedad; hay un "origen" histórico de los signos y la mayor parte de las veces, a falta de un árbitro que juzgue la correcta aplicación de un signo a un sentido, se cede el paso a la tradición heredada. Así se buscan nuevas justificaciones para la relación entre los signos y los objetos. Esta es la verdadera *motivación* de los signos en los objetos, motivación aprendida, ilusión etimológica que no por falsa deja de ser real. Piénsese en los llamados casos de "etimología popular", en los que claramente se ve el esfuerzo de los hablantes por hallar una razón de los signos. *Vagabundo* se motiva en *vagamundo; cadáver* en *caláver* (por *calavera*); *ceviche* en *sea beach; gringo* en *green goat; a la malacanchoncha* en *mala* (por lo mal que se siente uno al girar muchas veces, que es *canchoncha,* según nuestro presidente de la Academia Mexicana de la Lengua), etc., etc.

En todos esos casos la noción de *convención* se aplica difícilmente. Si uno hace algunos experimentos con los hablantes de una lengua verá que nunca quedan satisfechos cuando, a la pregunta de si un cierto objeto se puede nombrar en una cierta manera, se les responde que sí, que es lo mismo en tanto se haya convenido en ello previamente. La convención es un puro "porque sí" de los que se dicen a los niños.

Esto mismo se puede comprobar también en la terminología científica. Los teóricos de la definición científica (Hempel, Quine, Papp, etc.) dan como ideal de la definición el caso perfecto de la convención: la definición nominal. En esta, basta estipular claramente la relación entre una expresión y un objeto (o un concepto) para que

la definición sea buena; es el ideal del lenguaje lógico.

De ser posible realizar convenciones en esa forma, daría lo mismo que el lenguaje científico del inglés o del español se sustituyera por fórmulas algebráicas, o que el español adoptara sin ninguna dificultad la terminología científica proveniente de otra lengua. Sin embargo, los llamados *agujeros negros (black holes)*, esas masas densas de energía que no emiten radiación, han sido nombrados con un sugerente signo de la lengua natural; un multivibrador biestable, en la ingeniería de la computación electrónica, se llama *flip-flop* en inglés y, según parece, causa risa a los ingenieros mexicanos su equivalente español citado. Así se encuentran cientos de ejemplos en los que la ciencia prefiere signos de la lengua natural a pesar (o quizá por eso) del "peligro" de las metáforas y de la imprecisión que tienen respecto de sus descripciones científicas o matemáticas.

Esto se puede interpretar como parte del fenómeno de la motivación que aparece cuando una lengua se considera en su contexto histórico y social. Hay una necesidad de "transparencia" de los signos a los objetos representados. Cuando el signo se vuelve opaco es difícil de manejar, no se comprende en su totalidad a qué se refiere y los hablantes se refugian en un verbalismo confuso. En la lengua inglesa la transparencia de los términos científicos es más natural que en el español, para el que trescientos o cuatrocientos años de dependencia científica han venido a significar una especie de *diglosia* entre la lengua común y la lengua científica. Considérese el caso, por ejemplo, de un término usado en neurofisiología para designar un proceso cerebral

que consiste en la manera en que ciertos estímulos esporádicos y nuevos sobre las neuronas van desencadenando, prendiendo, reacciones cerebrales que más tarde se estabilizan como normales. En inglés es *kindling,* que quiere decir *ocote* en México, *encendeja,* esa madera de rápida combustión que se usa para encender chimeneas; Para un especialista anglohablante la relación del fenómeno con su experiencia cotidiana es inmediata; para un médico mexicano es simplemente un término, sin más sentido que una convención, pero desligado completamente de su vida diaria.

La *arbitrariedad* radical del nivel semiológico ha venido a presentarse, en el nivel idiosincrónico, como determinada por una historia de la sociedad. No es que el principio arbitrario cese de operar, pues como se decía antes, permanece siempre en el horizonte semiológico de la lengua y funda el cambio, la conservación y la posibilidad lingüística de la metáfora. Solamente se ve matizada, primero, por la estructuración del sistema y el juego de valores que determina; luego, por la mediación de las tradiciones lingüísticas que se dan en la historia de una comunidad.

Me parece que es un importante aporte el que ha dado Rudolf Engler no solamente a la lectura contemporánea de Saussure, sino en especial para eliminar algunas de las dificultades más serias de la teoría semántica moderna y para aproximarnos un poco más a la superación deseada de las concepciones estructuralistas que, en ese sentido, se extienden a las transformacionalistas.

Desde el punto de vista semiológico ya es posible, con la acción axiomática del principio de la arbitrariedad, integrar a la teoría un aparato más rico para comprender y tratar las relaciones

entre signos y objetos. Una vez que se abandona como espurio el criterio de biunivocidad de la relación, la función semiótica permite comprender en muchas maneras posibles las relaciones que establecen los seres humanos entre su lengua y su mundo: a partir de interpretaciones onomatopéyicas de los ruidos, a partir de formas lineales, superficiales o tridimensionales, a partir de colores o de texturas, a partir de relaciones condicionadas por experiencias anteriores de temperatura, de alegría, de ira, a partir de clasificaciones míticas, funcionales, cuantitativas o cualitativas, a partir de paralelismos, proporcionalidades o antagonismos entre procesos, etc., etc. La interpretación del sentido por la función semiótica puede aprovechar un número muy grande de percepciones humanas variables, aleatorias, que se ven apresadas por la forma del contenido. Pero puesto que una lengua natural es más que una forma y la sustancia se ha venido tejiendo en la historia, el tamiz último para transformar el sentido en signos es el de la memoria de la comunidad, como se encuentra contenida en su cultura lingüística.

1.5. Significación y código.

Ahora hay que retomar el tema del sistema lingüístico que se había dejado pendiente en el § 1.4.2 para acabar de hacer algunos matices importantes en la secuencia de razonamientos que se ha venido siguiendo. La cuestión que se plantea una vez explorados los alcances del axioma de la arbitrariedad es la de la interdependencia entre la actividad significativa, que ha venido a quedar fuertemente revalorada en los parágrafos anteriores, y el sistema lingüístico,

cuya posición relativa puede haber sufrido cierto empobrecimiento.

El hincapié hecho sobre la noción de la arbitrariedad tuvo como resultado una mayor comprensión de lo que implica para la teoría lingüística la significación, así como permitió explicar con mayor verosimilitud los cambios, tanto de significado, como los metafóricos. Pero eso no quiere decir que se tenga que optar por una especie de teoría de la significación pura, de la libertad total para significar (un exceso ya documentado en alguna corriente de la lingüística de este siglo), sino que hay que considerar en todo su valor la enseñanza de la lingüística del sistema: que es posible significar algo en lengua porque hay un hecho sistemático lógicamente anterior a la significación.

Visto así, el sistema lingüístico no es algo de lo que pueda uno prescindir teóricamente. Se trata solamente de no caer en la posición de muchos estructuralistas para los cuales el concepto de sistema se volvió *real* —es decir, lengua en sí— y en tal forma todopoderoso que, en un translado constante de la especulación teórica hacia la realidad, recubrió aparentemente todo el ámbito de la significación con una estructuración a ultranza del significado, hasta hacerlo aparecer como la simple realización de una virtualidad desde siempre prevista por el sistema. Por el contrario, lo que interesa es conservar los aportes de la lingüística del sistema como objetos teóricos y no como enunciados sobre el ser de la lengua (lo que no quiere decir, tampoco, que la lengua no tenga alguna sistematicidad).

Si la significación se muestra tan importante para la teoría como se ha defendido más arriba y si el sistema ha permitido conocer mejor las caracte-

rísticas de las lenguas, las relaciones entre sistema y significación no se deben ver como antitéticas o contradictorias, sino que es mejor asumirlas como una tensión constitutiva de lo que entendemos por lengua natural. En otras palabras, es preferible pensar que las lenguas sirven para comunicarse y sirven para "decir cosas nuevas" porque tienen un sistema que las organiza –pero no las agota– y porque hay una facultad significativa de un orden semiológico que las relaciona con el conocimiento y la percepción. Todo estudio unilateral de estos dos principios tendrá, por lo tanto, que conducir a los callejones sin salida que hemos experimentado a todo lo largo de la historia moderna de la lingüística.

La ya tratada posición central del sistema en la lingüística estructural ha producido varias versiones diferentes de la distinción entre estructura o sistema y significación. Así por ejemplo, en la lingüística que niega la existencia teórica del significado de los signos, la descripción de una lengua se termina cuando se han clasificado todas las combinaciones pertinentes de unidades lingüísticas desde el morfema hasta la oración. En otras palabras, la lengua aparece como la pura combinatoria de segmentos materialmente identificables y, por lo tanto, la significación se convierte en algo ajeno y externo a la lengua, que el lingüista reconoce "heurísticamente", pero que no puede tratar científicamente.

Cuando, a partir de procedimientos de clasificación de los signos similares a los que utiliza la lingüística descriptiva norteamericana (sea de Bloomfield o de Harris), se logra enumerar los miembros de una clase de morfemas o de signos de cualquier magnitud, la armazón sintáctica que relaciona los elementos pertenecientes a distin-

tas clases se distingue radicalmente y permite establecer las conocidas diferencias entre "palabras llenas" y "vacías" o "gramaticales" y "léxicas", es decir, se opera la dicotomía tradicional entre lo que corresponde al léxico y lo que corresponde a la gramática.

En esos casos lo que predomina es la caracterización del sistema, mientras que la significación —o su producto: el significado— se abandona sea a las intuiciones acientíficas, sea a unos "inventarios abiertos" correspondientes al léxico, que se posponen o se envían al cajón de sastre llamado lexicología o lexicografía.

Se han intentado varias justificaciones de la distinción léxico/gramática; las más simples son las anteriores, que no fallan por lo que afirman —la sistematicidad de los elementos— sino por su incapacidad aceptada para tratar fenómenos "léxicos". A pesar de esa limitación teórica, en la medida en que permanecen en el marco de las técnicas de análisis sus resultados no solamente convencen, sino que llegan a ser muy buenos desde el punto de vista del sistema.

Otras justificaciones más ambiciosas se vuelven al mismo tiempo más discutibles, como lo son las de M. Mathiot —que en realidad representan a muchos otros lingüistas— para la cual "the lexicon is the cultural system of reference, i.e. the system of reference to the various basic types of phenomena (such as entities or actions) distinguished by the culture. The grammar of a given language is the structure of discourse in that language." En seguida agrega: "Note that the distinction is based on ontological definition of both grammar and lexicon".[17]

[17] Madeleine Mathiot, "The place of the dictionary in linguistic description", *Lan*, 43 (1967), 703-724.

Con una dicotomía tan radical entre léxico (externo a la estructura del discurso y determinado por la cultura) y gramática (la estructura misma y ¿no determinada por la cultura?) la naturaleza de la lengua se vuelve un agregado de dos tipos de entidades cualitativamente distintas ("ontológicamente") y ahora el problema que habría que resolver sería cómo justificar la comprobación cotidiana de que segmentos mayores que el vocablo funcionen como si fueran de ese mismo tipo, por ejemplo *ojo de agua, turrón de almendra, dar chicharrón, colgar los tenis,* etc., por una parte, y por la otra cómo sostener que la adquisición de una lengua en los niños se logre por el reconocimiento de unidades necesariamente discretas, como lo son las "palabras" de los adultos.

La misma M. Mathiot sostiene, en el trabajo citado, otra razón para hacer la distinción léxico/gramática: "A distinction is regarded to be theoretically essential if it fits either one of two cases: (1) the distinction is suggested by intuitive observation and proves to be crucial for carrying out the analysis; (2) the distinction imposes itself as a result of the analysis. The distinction between lexicon and grammar fits the first case." (p. 708) Es indudable que ambas condiciones no provienen de una observación ingenua de los "hechos", sino que precluyen una toma de posición teórica anterior, consciente o inconsciente respecto de la distinción.

En épocas más recientes la diferencia se ha sostenido por otros motivos; la entrada al panorama de la lingüística de las formalizaciones de inspiración lógica que significó la corriente transformacionalista, parte de la definición de un monoide generativo compuesto por un conjunto

de elementos (vocabulario) y un conjunto de reglas que se aplican a los elementos (gramática). No se toca el problema teórico de la justificación del vocabulario sino que, en los inicios del movimiento transformacionalista, se dejó a la mera consulta del diccionario *Webster* (Katz). Posteriormente, en las corrientes de la semántica interpretativa y la semántica generativa, el problema del léxico se ha vuelto el núcleo de todas las discusiones en torno a las relaciones entre el significado y la estructura formal de la oración.

En el importante trabajo de Bernard Harrison arriba citado está la que, a mi juicio, es la mejor explicación de las dificultades que ha tenido la lingüística del sistema para integrar léxico y gramática en una forma coherente; según él, las "teorías empiricistas del lenguaje" (ETL), heredadas desde la época de Locke y Condillac hasta ahora con Russell, Quine, Bloomfield, etc., sostienen que "since the semantics of a language has to do only with associations between words and things, or between sentences and things, or between sentences and sentences[18], whereas what syntax has to do with is, roughly, the ways in which the members of certain structurally defined classes of words or morphemes can and cannot stand in relationship to one another in sentential contexts, syntactic description can proceed without ever mentioning any point of semantics, and viceversa." Es decir, el problema del léxico y la gramática no es un problema de dos tipos de elementos (palabras y combinaciones) sino de semántica, de significado.

Con lo que uno se enfrenta no es, por lo tanto,

[18] Cf. B. Harrison, Op. cit. (n. 14), p. 24.

con la discusión en torno a la mejor o peor definición de clases extensionales de palabras en gramática, sino con la tensión que existe entre la significación y el sistema, tal como se indicó arriba.

Me parece que se podrá comprender mejor esa relación si, en vez de *sistema,* usamos un término equivalente para estos fines, que es el de *código.* Se ha estudiado suficientemente y en campos no estrictamente lingüísticos, sino psicológicos y matemáticos, las características generales de los códigos para poder servir de vehículos de información. Todo código requiere una estructura ordenada de los elementos que maneje y una medida de repetición de algunos de ellos (redundancia), que tiene la función de asegurar la transmisión y comprensión del mensaje. La capacidad de crear códigos no es accidental en muchos animales, sino que se comprueba en hormigas, abejas, delfines y algunos monos; lo mismo se comprueba en los seres humanos y, de acuerdo con Lenneberg[19], hay que considerar esa capacidad como biológica, es decir, no imputada al hombre, sino constitutiva suya.

La estructura de una lengua es un tipo especial de código; facultad universal, sostiene Chomsky, que adquiere su forma específica en el interior de cada comunidad lingüística. En cuanto universal, la función del código es servir de vehículo a la significación, prestarle un elemento sensorial para que se comuniquen los seres humanos. Pero precisamente porque es una capacidad, el código no está estructurado, sino que se estructura permanentemente; es decir, el código no se impone a la significación en forma total, sino que la

[19] Eric Lenneberg, *Fundamentos biológicos del lenguaje.* Trad. N. Sánchez S. T. y A. Montesinos, Alianza, Madrid, 1975.

forma y recibe a su vez transformaciones originadas en la significación. La investigación en gramática transformacional nos da una buena muestra de la complejidad de la estructuración del código.

Cuando el código se especifica, o sea, cuando se aprende a hablar una lengua, su actividad estructurante se restringe a las condiciones étnicas de una comunidad lingüística. Como en el caso de los diferentes niveles del principio de la arbitrariedad, en el nivel particular de una sola lengua el código se ve como algo heredado y es la historia de esa lengua la que delimita las posibilidades reales de transformación del sistema.

Lo que hay que destacar, entonces, es que el sistema de una lengua es igualmente importante para comprender su naturaleza como lo fue comprender su actividad significativa. Sin un código, la significación no es concebible bajo las condiciones de la especie humana, pero al mismo tiempo, sin significación el código no pasa de ser, o un primitivo mecanismo diacrítico como el de la danza de las abejas, o un simple juego lingüístico.

Como se deja ver en la cita de Harrison, es la misma concepción de los signos como etiquetas de los objetos la que ha conducido a hacer la distinción de dos constituyentes de la lengua no en términos de significación y código, sino de semántica y sintaxis o su variante más tradicional de léxico y gramática. La ventaja de comprender la dualidad en estos otros términos aquí propuestos es que la semántica, los significados, se extienden sobre todos los elementos codificados y no sobre una parte de ellos; también se gana una mayor flexibilidad para explicar por qué el léxico varía constantemente en sus relaciones sintácticas y por qué hay relacio-

nes sintácticas que se lexicalizan. Hay que comenzar a considerar, por lo tanto, la semántica como un campo con diferentes rangos de organización formal, como los propone Klaus Heger[20].

Bajo una concepción parecida no es necesario suponer que las palabras tengan una existencia separada del resto del sistema lingüístico, determinada por sus relaciones hacia el exterior, hacia el mundo y la cultura, ni que la gramática es un juego vacío de combinaciones sin sentido. Gramática y léxico, por el contrario, son dos variaciones formales del código que se ajustan a las necesidades de la significación. No hay una naturaleza dicotómica de la lengua, sino la dualidad del sistema y la significación. De tal manera lo distinguible para fines metodológicos es el sistema formal –la forma del contenido– y el significado –sustancia del contenido– que selecciona la función semiótica aplicada al sentido. Así tampoco hay que caer en el debate tendiente a la unilateralidad de la sintaxis frente a la semántica; se trata de dos estratos inseparables, que se distinguen con puros fines teóricos y metodológicos.

La cuestión que se plantea en seguida es ¿por qué entonces la distinción léxico/gramática se ha presentado en la historia de la humanidad desde hace tantos siglos? El problema no es de fácil solución y requiere una investigación particular. Sin embargo, me parece explicable a partir de las características del sistema como código, es decir, a partir del orden y de la redundancia que requiere la transmisión de la información. Las palabras del código ofrecen la característica peculiar de ser lo suficientemen-

[20] Cf. Heger 76a.

te compactas como para poderse repetir y guardar en la memoria, pero al mismo tiempo son parte de la capacidad del código para generar salidas ilimitadas desde su estructura general, al igual que con los otros rangos superiores al de la palabra: composiciones, frases, oraciones, combinaciones de oraciones, etc. Porque el código admite tanta variación y porque la significación es la que determina las necesidades de comunicación, las palabras se cristalizan entre ambos principios, como dejó entrever Saussure[21] y como creo quisieron decir los lingüistas praguenses en 1929.[22]

Pero lo que orienta la cristalización de las palabras no es, en consecuencia, el código, para el cual es lo mismo producir un signo de la magnitud de la palabra que de la perífrasis, la oración o, en general, cualquier paráfrasis; es la tradición de la comunidad lingüística, la lengua histórica la que sedimenta tipos de signos y fija los márgenes de variación de sus magnitudes. De ahí las dificultades que enfrenta la lexicografía, para la cual restringirse a la inclusión de signos delimitados por pausas o blancos de la escritura significa, o bien la aceptación de un caprichoso criterio de la tradición (por eso la dificultad en la lexicografía francesa para hacer una entrada *pomme de terre*), o bien el convencimiento de

[21] Cf. *Cours*, Caps. 4 y 5 de la 2a. parte donde, a base de sus explicaciones sobre la diferencia entre *valor* y *significación* y sobre las relaciones sintagmáticas, puede inferirse lo anterior.

[22] En las "Thèses présentées au Premier Congrès des Philologues Slaves", *Travaux du Cercle Linguistique de Prague*, 1(1929), 5-29 se propone la dualidad entre "actividad lingüística denominadora" y "actividad sintagmática" como la que explica la formación de las palabras y su encadenamiento sintagmático. En mi interpretación, lo que llamo *significación* corresponde a lo primero y lo que llamo *código* a lo segundo.

que, en tratándose de unidades significativas, los diccionarios debieran listar no sólo palabras sino derivaciones, composiciones, locuciones, proverbios, etc.[23]

Esto quiere decir que la palabra no es una unidad delimitable por criterios exclusivamente sistemáticos, sino que su último perfil lo adquiere en la historia. También quiere decir que, en el aprendizaje de una lengua, lo que se adquiere no son signos previamente delimitados, sino bloques expresivos que la sociedad va analizando según su tradición lingüística. Por eso a los niños les cuesta tiempo llegar a la conciencia de la palabra que tienen los adultos. Recuerdo que cuando yo era niño asistía a un jardín de niños llamado Brígida Alfaro; durante mucho tiempo este nombre me fue indescifrable: lo segmentaba como Brígid-Alfaro o como Brigidal-Faro; lo mismo le sucede a todos los niños: en vez de *el-elevador* dicen *lelevador;* en vez de *el-hilo, el-lilo,* etc. Las palabras son entonces signos cuyas fronteras se aprenden con la educación; el léxico es producto de una cultura lingüística y no reflejo de una dicotomía léxico/gramática.

1.6. Recapitulación.

No es gratuita la intención de revisar las enseñanzas de la lingüística moderna respecto al sistema de signos lingüísticos que es una lengua. Obedece a una necesidad apremiante de la práctica lexicográfica para la cual el modelo semántico disponible hasta ahora es casi inaprovechable, por la cantidad de reducciones con que es

[23] Cf. el fundamental estudio de Josette Rey-Debove, *Etude linguistique et sémiotique des dictionnaires francais contemporains* Mouton, La Haya, 1971, pp. 86 y 112.

necesario operar y por lo insatisfactorio de sus concepciones para el tratamiento de los datos lingüísticos de una lengua no idealizada sino cotidianamente concreta. Se ha querido reconocer el valor de los postulados estructuralistas sobre todo para la fundamentación de un enfoque autónomo de la lengua, que permite un modelo del signo más adecuado a los fenómenos lingüísticos corrientes. Pero al mismo tiempo se ha atribuido a ese movimiento intelectual que postula la autocontención del sistema lingüístico la imposibilidad teórica y práctica de recuperar la naturaleza histórica de las lenguas.

A cambio de eso se ha relatado la importancia del concepto saussureano de la arbitrariedad radical de los signos —a partir de la exégesis moderna de su obra— para comprender la variabilidad permanente de la lengua y para oponerse, en especial, a la concepción convencionalista de las relaciones signo-referente. Se sostiene, por lo tanto, que la hipótesis de la naturaleza arbitraria del signo lingüístico permite superar los marcos estrechos del estructuralismo y buscar métodos más acordes con la experiencia lexicográfica, métodos que, a la vez, no desmerezcan en cuanto a rigor teórico y a coherencia intelectual.

Es evidente que el enfoque que se propone está más orientado a lo que la tradición lingüística alemana llama *semasiología,* es decir, estudio de la semántica propia de una sola lengua; es de eso de lo que trata precisamente una lexicografía. Pero ello no implica ni el desconocimiento del valor analítico de la *onomasiología* —estudio de las características semánticas universales a las lenguas o de *la* lengua—, ni mucho menos la propuesta de una concepción teórica incompatible con ella.

En lo que sigue se trata de adecuar un método de análisis semántico a las hipótesis presentadas en esta sección.

2. La teoría del campo léxico.

Experimentado el análisis de los significados de las palabras por la lexicografía durante toda su historia, y realizado en forma implícita a partir de las diferencias que aparecen entre dos significados distintos de dos o de una misma palabra[24], la llamada teoría del campo léxico o del campo semántico representa una explicitación de aquel procedimiento tradicional de la lexicografía desde el enfoque particular del estructuralismo. Puesto que, por lo tanto, se trata de una concepción estructuralista de los significados de las palabras, hay en ella varios elementos teóricos y metodológicos que es necesario analizar para poder establecer la medida en que es aplicable al análisis semántico lexicográfico bajo las hipótesis desarrolladas en la sección anterior de este trabajo. Aquí no intentaré ni repetir la tarea de documentar históricamente el desarrollo de esta teoría, sobre el cual se han escrito ya varios estudios[25], ni elaborar una crítica exhaustivamente documentada de todas sus características. Reconozco que este procedimiento implica el riesgo de manipular equívocamente o de uniformar prematuramente las ideas de cada autor de tra-

[24] Salvo, por supuesto, la discusión teórica entre las versiones polisémica y homonímica del significado de los vocablos. Cf. al respecto Heger 76a.

[25] Por ejemplo el de K. Baldinger, *La semasiología. Ensayo de un cuadro de conjunto.* Universidad del Litoral, Argentina, 1954, y el de Horst Geckeler, *Semántica estructural y teoría del campo léxico.* Gredos, Madrid, 1976, en los que aparece además una abundante bibliografía.

bajos sobre el campo léxico; sin embargo confío en que la crítica sea válida en el plano de la generalidad en que se sitúa.

2.1. La estructura semántica.

Lógicamente el postulado de partida para la teoría del campo léxico es que, siendo el léxico una parte de la lengua y siendo la lengua un conjunto de relaciones estructuradas, el léxico tiene a su vez una estructura propia. Esta idea no era, en sus orígenes, declaradamente estructuralista: partía de comprobaciones hechas en épocas muy tempranas de la lingüística del siglo XX, cuando aún no era por todos reconocido el valor de Saussure o de las escuelas que se formaron en la década de 1920. Se basaba generalmente en los estudios hechos sobre la terminología de los colores o de los grados militares. En efecto, sobre la gama cromática, los límites del *rojo* se encuentran al pasar al *anaranjado* y los de este al pasar al *amarillo,* por lo que el "significado" de *anaranjado* se puede concebir como "lo que no es ni rojo ni amarillo". En el caso de la jerarquía del ejército se puede decir, también, que el "significado" de *capitán* solamente se reconoce si se sabe que es un grado inferior al de *mayor* y un grado superior al de *teniente.* El léxico se presenta, en consecuencia, como una lista paradigmática de relaciones opositivas y negativas, en que cada término se define por su posición en la estructura y se presenta al análisis como un sugestivo paralelismo con las estructuras fonológicas, que tanto éxito tuvieron y de las que dependió fundamentalmente el desarrollo del estructuralismo lingüístico.

2.1.1. Antecedentes europeos de la teoría del campo léxico.

Hubo varias concepciones distintas del campo léxico: la asociativa, expuesta en el curso de Saussure y desarrollada por Bally, que proponía la existencia de constelaciones de palabras en que, por formas similares del significante o del significado, por contigüidad física de los objetos representados, como en el caso de *buey* y *carreta*[26], o por proximidad cultural, como *buey* y *trabajo,* se establecía una esfera de relaciones múltiples. Pero no es difícil imaginar en qué consistió la crítica de esta idea: relaciones de órdenes tan diferentes, que no solamente dependen de alguna organización del léxico, sino que sobre todo se producen a partir de experiencias subjetivas imposibles de generalizar para todos los miembros de una comunidad, se prestan solamente a un manejo intuitivo, aleatorio e inverificable. (Lo que es francamente extraño es, sin embargo, que todavía existan métodos de estudio psicológico que partan de una concepción similar para analizar no un comportamiento psíquico, sino un comportamiento lingüístico).

Otra concepción del campo léxico fue la de Matoré[27], quien vio en el estudio de las palabras documentadas en textos de una época determinada la posibilidad de discernir una organización de conceptos característica de una sociedad en una época de su historia; esfuerzo este emparentado con el todavía actual análisis de conte-

[26] A partir del concepto saussereano de las relaciones asociativas en la 2a. parte del curso, V. 3 y expuesto en varias obras de Ch. Bally.

[27] En *La méthode en lixicologie,* París, 1953.

nido que se hace en sociología. Los estudios de Matoré, que aparentemente se orientan hacia lo social pasando sobre lo lingüístico, no dejan de ser sugerentes para el análisis semántico actual.

Los antecedentes directos de la teoría del campo léxico se encuentran —según los estudios de Osswald y Geckeler[28]— en las obras del lingüista alemán Jost Trier. Fue él quien por primera vez propuso en forma explícita la idea de que el vocabulario de una lengua está organizado en un sistema de relaciones comparables con los vidrios que componen un vitral: en una de esas construcciones cada vidrio ocupa un lugar preciso y perfectamente ensamblado con el resto de los vidrios que la componen. Cada palabra queda entonces determinada, en su significado, por los significados que forman parte de la estructura total. Posteriores aclaraciones, matices y correcciones a la teoría, en especial debidos a Leo Weisgerber, pulieron la noción del campo léxico hasta que se convirtió en un concepto estructuralista y se extendió por la lingüística europea.

2.1.2. Antecedente norteamericano: el análisis componencial.

Paralelamente al desarrollo del campo léxico en Europa, en los Estados Unidos de Norteamérica se elaboró un método de análisis muy parecido, del cual, por su desconocimiento general en la vertiente europea, vale la pena resaltar algunas de sus características. Se trata del llamado análisis componencial, resultado de la experiencia de los lingüistas de campo en la organización del

[28] Geckeler, op. cit. y P. Osswald, *Frz. "campagne" und seine Nachbarwörter im Vergleich mit dem deutschen, englischen, italienischen und spanischen; ein Beitrag zum Wortfeldtheorie.* Tübingen 1970.

léxico de lenguas desconocidas, especialmente americanas. Su punto de partida, aunque englobado en los cánones de la lingüística descriptiva (estructuralista) es tan diferente como puede ser una semántica que no ha delimitado claramente la diferencia entre signo y cosa (Cf. supra § 1.2). Para los propulsores del análisis componencial las palabras son, una vez más en este trabajo, etiquetas de los objetos pertenecientes a una cultura. No se delimitan dentro de un campo propio de los signos, sino que su significado está inseparablemente ligado a las distinciones que hace una cultura sobre su realidad circundante.

Si el campo de la semántica se enfoca de esa manera, es natural que el instrumento de trabajo del análisis componencial sea aquel que mejor se aplique a la clasificación y denominación de los objetos de una cultura: la taxonomía, concebida según los principios usuales en las ciencias de la naturaleza.

El procedimiento del análisis componencial consiste, por lo tanto, en definir los elementos distintivos de la taxonomía (*taxones*)[29] respecto de un *conjunto léxico* previamente aislado, del que se supone el taxón es una unidad distintiva. En esa forma se elabora una *taxonomía folk*, que consiste de un sistema de palabras aisladas *(segregados monolexemáticos)*, relacionadas entre sí por sucesivas inclusiones jerarquizadas.

Como en toda taxonomía científica (por ejemplo la botánica), la estructura es piramidal:

[29] En H. C. Conklin, "Lexicographical treatment of folk taxonomies" *Problems in lexicography*, F. W. Householder y S. Saporta (eds.) anejo al *IJAL*, v. 28, No. 2 Indiana Univ. Press, Bloomington, 1962.

en el vértice superior hay un solo taxón que incluye a todos los inferiores (cf. el *archilexema* del campo léxico); a partir de ese nivel la sucesión jerárquica de inclusiones se construye de arriba hacia abajo. Entre taxones de un mismo nivel hay, necesariamente, oposiciones binarias, que tienen el efecto de impedir cualquier traslape entre ellos. Esta estructura tiene por consecuencias, a) que cada taxón pueda pertenecer solamente a un nivel, y b) que no sea posible la existencia de sobreposiciones o intercambios entre taxones. El modelo de estructura del análisis componencial es, por lo tanto, uno solo y muy rígido.

Si se toman en consideración las condiciones epistemológicas que quedan implicadas en esta concepción del campo léxico, se pueden sacar conclusiones importantes: ante todo, que la cualidad o el perfil lingüístico de una taxonomía *no* es relevante para la clasificación de los objetos de la cultura; es decir, puesto que lo que interesa es descubrir la forma como una cultura concibe el mundo que la rodea, las relaciones entre los signos que prestan cuerpo a la nomenclatura no son pertinentes al estudio; los signos son puros *nombres* de cosas.

Por otra parte la limitación a objetos materiales de una cultura y, posiblemente, el tradicional pragmatismo del pensamiento norteamericano, obliga a concebir la taxonomía resultante como algo del ámbito de lo "directamente observable"; dicho en otra forma, el análisis componencial no se plantea la relación entre objetos de la realidad y su conceptualización como algo problematizable, sino que supone que, una vez consultados sus informantes sobre los elementos clasificatorios que utilizan, estos elementos son "observa-

bles" y no media entre ellos y su aplicación ningún proceso de abstracción como el de la ciencia occidental. La taxonomía folk viene a ser, por definición, cualitativamente distinta de la taxonomía científica. Así por ejemplo, Conklin asienta que la taxonomía hanunoo de las plantas tiene 1800 términos frente a la científica que solo llega a 1300, sin añadir más comentarios sobre el sentido de tal comparación.

En forma consecuente con las dos características mencionadas, los practicantes del análisis componencial llegan a la conclusión de que los taxones son inmanentes a cada comunidad lingüística particular y por ello la comparación entre lenguas viene a ser si no imposible, por lo menos muy difícil.

Por último, también conviene destacar que los taxones del análisis componencial son —según se deduce de los textos en cuestión— palabras pertenecientes a la lengua en estudio, por lo que se reafirma la inmanencia del análisis a la lengua en consideración (Cf. infra § 3.3.2.).

Todos estos antecedentes aportan datos importantes para comprender el horizonte en que se desarrolla el análisis estructuralista del campo léxico. En lo que sigue intentaré sistematizarlos en un esquema general de la teoría.

2.2. Los elementos significativos.

En forma paralela al modelo fonológico de la estructura lingüística, la concepción estructuralista del léxico también supone una estructura relacional y opositiva de los significados de las palabras. En esa estructura los elementos de análisis son *rasgos significativos, pertinentes* o *semas.*

La determinación de los rasgos es uno de los

problemas fundamentales del análisis del campo léxico. En un famoso trabajo de B. Pottier[30] los rasgos pertinentes se obtienen mediante una encuesta entre hablantes de la lengua en estudio, encuesta que consiste en inquirir sobre aquellos elementos que le permiten al hablante clasificar un objeto respecto de un vocablo de su lengua. Es decir, el análisis de Pottier se orienta, en forma muy similar al análisis componencial, hacia la relación entre palabra y cosa. Los rasgos así encontrados se vuelven pertinentes cuando resultan comunes a la clasificación que hace la mayor parte de los hablantes de una lengua[31].

De este trabajo de Pottier no es posible deducir cómo se plantea la estructura taxonómica implícita en su concepción. Aparentemente la conmutación binaria no se aplica desde los primeros momentos, sino ya que se obtuvieron los semas correspondientes a cada palabra. Como el análisis se hace sobre la clasificación de los objetos creo que no se puede pensar que en ese momento haya una estructura, sino apenas haces de características determinadas singularmente para cada objeto. Suponer, en cambio, que sí hay una estructura al inicio del análisis, implica que se trataría de estructuras de los objetos, lo cual o es lo suficientemente trivial como para que no importe, o es una toma de posición muy fuerte desde el punto de vista de la ontología y de la teoría del conocimiento.

Una vez reunidos los semas correspondientes

[30] B. Pottier, *Recherches sur l'analyse sémantique en linguistique et en traduction automatique*, II. Univ. de Nancy, 1963.

[31] Una versión interesante del método de Pottier, en que se opera con comunes denominadores, es el trabajo de Raúl Avila, "El campo semántico 'aparatos eléctricos para iluminación' ", *NRFH*, 21,2 (1972), 273-300.

a la aplicación de la palabra (o de la *lexía*, en la peculiar terminología de Pottier) a un objeto, estos se articulan en un *semema*, que es el conjunto de semas definidores de la palabra.

Si se conviene en representar con *S* al semema, con *s* al sema y con los subíndices *1, 2, i, x, y, n* semas o sememas distintos, la fórmula del significado de una palabra sería:

$$S_i = \{ s_1, s_2, s_3, \ldots, s_n \}$$

Una vez aislados los sememas en cuestión, el análisis estructural comienza en forma similar a como se plantea el análisis fonológico, es decir, aplicando la conmutación binaria a la comparación entre dos sememas, hasta lograr oposiciones que distingan un semema de otro. También en ese proceso la verificación de los semas se realiza consultando a los informantes.

A manera de ilustración, sin que haya mediado una aplicación completa del método de Pottier, se presenta el siguiente cuadro de los *asientos* en español:

	s_1	s_2	s_3	s_4	s_5	s_6	s_7
silla	+	+	−	−	+	+	−
banca	+	+	−	−	−	−	−
banco	+	+	−	−	−	+	−
sillón	+	+	+	−	+	+	−
sofá	+	+	+	−	+	−	−
butaca	+	+	−	+	+	+	+

en que:

s_1 = para sentarse
s_2 = elevado respecto al suelo
s_3 = de material duro
s_4 = con brazos
s_5 = con respaldo
s_6 = para una persona
s_7 = fijo al suelo

Según Pottier, a falta de un método que permita agrupar esos vocablos en un "campo" de una manera segura y verificable, lo que habría que hacer sería proceder a una larga serie de pruebas de intersección entre los semas del cuadro, para que fuera ese resultado empírico y cuantitativo el que determinara los sememas agrupables en un campo semántico. El problema de la delimitación del campo semántico es todavía hoy uno de los más espinosos de la teoría en cuestión (cf. infra § 2.3).

Pero por el momento lo que interesa es suponer que se ha logrado constituir un campo semántico de los asientos y que se ha aplicado a su construcción los principios básicos del estructuralismo. Aunque otros teóricos del campo léxico no reconozcan a Pottier –y supongo que ni él mismo lo haría– como coautor de la teoría, el cuadro de las *sièges* en francés se ha convertido en ejemplo clásico de lo que es un campo semántico.

Del ejemplo aquí introducido se pueden hacer algunas observaciones de importancia sobre los rasgos significativos.

2.2.1. Interpretación de la oposición binaria.

En este análisis la oposición binaria se ha con-

cebido como de presencia o ausencia de un rasgo. Así en el caso de *silla,* puesto que el rasgo s_3 ('de material duro') no está presente, es decir, puede haber sillas de material blando, su ausencia no marca nada; el sema resulta indeterminado para *silla.* Esta concepción implica por lo tanto pérdida de especificidad en el análisis, pues no se puede decir si el semema de *silla* nunca debe adoptar ese sema o si el sema puede o no puede aparecer en su interior aleatoriamente.

Como señala Greimas[32] la interpretación de la oposición también puede ser como de positividad/negatividad. En ese caso el cuadro debería reformularse:

	s_1	s_2	s_3	s_4	s_5	s_6	s_7
silla	+	+	0	0	+	+	−
banca	+	+	+	−	−	−	−
banco	+	+	0	0	0	+	−
sillón	+	+	−	+	+	+	−
sofá	+	+	−	+	+	−	−
butaca	+	+	−	+	+	+	+

Los semas marcados con 0 evidencian, al hacer su interpretación con la relación positivo/negativo, la necesidad de analizar más cuidadosamente su importancia para el campo semántico. Una solución es desdoblarlos en, por ejemplo s_3' 'de material duro' y s_3" 'de material blando' (igual mente en los casos de s_4 y s_5), dada la dificultad que representaría conservar un sema con una disyunción en su interior (s_3' v s_3").

La interpretación de la oposición binaria es,

[32] A. J. Greimas, *Sémantique structurale; recherche de méthode.* Larousse, París, 1966, p. 8.

por lo tanto, un factor muy importante del método del campo léxico pues, de no hacer explícita la interpretación seleccionada, da lugar a constantes contradicciones en el análisis.

2.2.2 La pertinencia de los semas.

El término *pertinente* aplicado a los rasgos es muy afortunado en cuanto tiene el objeto de llamar la atención sobre la posibilidad de que haya una multitud de rasgos sin interés para el análisis. De acuerdo con Pottier la pertinencia se explora a partir de pruebas realizadas con los informantes; son los hablantes mismos los que deben seleccionar los rasgos adecuados para clasificar un objeto con la ayuda de un vocablo específico. Sin embargo, en la práctica, el método resulta muy complicado. Como lo habrá experimentado el lector que haya tratado de elaborar un cuadro como el anterior con la ayuda de un grupo de personas, cada hablante puede proponer otros rasgos distintos o interpretar un rasgo en forma diferente y, además, hacer surgir la evidencia de que el proceso de reconocimiento de los objetos no es sencillo ni es igual para todos los miembros de una comunidad (posiblemente, entre las causas fundamentales, por el carácter arbitrario del signo, cf. 1.4.3.). En el trabajo citado de R. Avila se puede ver el tipo de dificultades que se suscitan.

Esta situación es la que da origen a la aplicación de otro de los aparatos metodológicos del estructuralismo: si entre informantes o entre enunciados de los informantes aparece una heterogeneidad en la selección de rasgos, la manera de solucionarla es con la ayuda de instrumentos que tienen la facultad de aislar lo variable y someterlo a las necesidades de homogeneidad del

estructuralismo: los conceptos de *subsistema* o de *diasistema* sirven para, primero, reunir por separado aquellos semas que hayan sido propuestos por un solo informante (idiolecto) o por un grupo homogéneo de informantes (dialecto) y así reorganizar el cuadro de los semas en varios subsistemas uniformes; segundo, volver a juntar los subsistemas construidos en un orden más general en que se obvien las diferencias (diasistema). Es entonces la necesidad epistemológica del estructuralismo de tratar con datos homo géneos la que genera los conceptos de *subsistema* y de *diasistema* y la que da origen al concepto de *lengua funcional* propuesto por Coseriu como unidad temporal, geográfica, social y de estilo. De manera que la *pertinencia* de un rasgo muestra dos de sus condicionantes: a) la variabilidad de los procesos de conocimiento de los informantes, y b) su dependencia del modelo estructuralista que la explica.

Pero hay un elemento más en la selección de los rasgos: si es más o menos sencillo suponer que la clasificación de objetos como los asientos se basa en características físicas fácilmente observables, cuando pasa uno a trabajar con taxonomías botánicas o zoológicas comprueba que, contra lo que suponen los practicantes del análisis componencial y contra lo que parece suponer el mismo Pottier, no solamente hay características que no son físicas, sino que aun supuestas características físicas están altamente elaboradas por abstracciones que realiza la sociedad durante años de experiencia cultural. Cuando trabajaba con la taxonomía de los peces en la comunidad veracruzana de Tlacotalpan, pude comprobar que la condición de que los rasgos taxonómicos sean físicos es más una con-

dición de la ictiología occidental que de los hablantes, en este caso pescadores con una tradición cultural de cerca de cuatrocientos años. Hay muchas características físicas de los peces que no se enfocan en la taxonomía "folk" y, en cambio, muchos de los rasgos pertinentes para los pescadores no son describibles sino a partir de concepciones ligadas al comercio del pescado (o sea, al *interés*), a las costumbres alimenticias, a la época del año y fundamentalmente a la experiencia tradicional de los pescadores más viejos. De manera que la condición de que los rasgos sean "directamente observables" no pasa de ser una ingenuidad del análisis componencial y la condición de pertinencia ligada a los hablantes algo muy difícil de lograr sin que aparezcan los obstáculos indicados.

2.2.3. El sema como unidad mínima.

Otra de las características del análisis en campo léxico que se revela en el ejemplo propuesto es la de la magnitud del sema. Los rasgos que proporcionan los hablantes o los que elabora el lingüista se tienen que someter a las necesidades de la oposición binaria. Esta exigencia —cuya justificación va más allá del campo de la lingüística y corresponde a una de las más generales concepciones de la teoría de las ciencias— se puede formular en la forma siguiente: la oposición *sí/no* es la única que garantiza que los elementos a los que se aplica no sean complejos y por lo tanto constituyan el límite necesario del proceso analítico.

Como se ve en el ejemplo hay semas que podrían continuar descomponiéndose: el s_2 podría descomponerse en un análisis de 'elevado', de 'respecto', de 'suelo', y de 'elevado respecto

al'; el s_4, el s_5, el s_1 y el s_6 sugieren también la necesidad de analizar el tipo de relación implícita en 'para' y 'con'. Lo mismo se podría hacer con cada uno de los semas propuestos. Aún más, cada uno de los conceptos utilizados para establecer los semas podría ser elemento de otros campos semánticos a su vez. 'Brazos' podría incluirse en un campo que contuviera 'patas' y 'respaldo'; 'duro' en otro donde se consideraran los 'materiales', etc.

El concepto de pertinencia también funciona en este aspecto, pues indica que el análisis debe detenerse cuando los semas que uno comience a proponer ya no se reconozcan como constitutivos del campo considerado, sino de otras estructuras más generales que engloban 'persona', 'animal', 'objeto', 'altura', 'dimensión', etc. Desde ese punto de vista la solución de Pottier es más prudente, pues da valor a la dependencia del análisis respecto de la competencia de los hablantes.

La alternativa estructuralista a la definición de la pertinencia de los rasgos significativos propone desechar la concepción de Pottier por pertenecer al mundo de los objetos y no al de los signos y, a cambio de ello, definir la magnitud de los rasgos pertinentes desde el interior de la estructura del campo. Esto quiere decir que el principio del binarismo se aplica radicalmente y que, por lo tanto, al hablar de semas se debe llegar a demostrar que se trata de los más simples posibles, o sea, que los semas son *rasgos significativos mínimos.* Más adelante volveré al tema, para sacar las conclusiones necesarias.

2.3. La estructura del campo.

De la discusión sobre la pertinencia de los semas se pueden sacar dos conclusiones: una

sobre la forma del semema y otra sobre la estructuración del campo.

El semema se obtiene con dos pasos sucesivos: en un primer momento, se hacen comparaciones entre vocablos reunidos sea a partir de los objetos representados (Pottier), sea a partir de las definiciones de un diccionario previo (Mounin)[33], sea a partir de una selección intuitiva de las palabras que entran en el campo. La comparación en ese nivel es previa al análisis estructural y consiste en una serie de tanteos sobre la pertinencia de rasgos seleccionados al azar. Una vez delimitado el grueso de los posibles semas incluidos en los sememas de los vocablos, el segundo paso consiste en aplicar sobre él los métodos de la oposición estructural con todas las variantes necesarias[34]; la medida de la pertinencia de los semas que forman el semema se basa en la necesidad de llegar a oposiciones binarias simples suficientes para que cada semema se distinga de los demás por la presencia de *un* sema diferenciador.

El conjunto de los sememas que se forma constituye el campo semántico. La fórmula de la página siguiente lo representa.

El campo semántico (CS) tiene a su vez una fórmula: $CS_1 = S_1 \cap S_2 \cap S_3 \cap S_4 \cap \ldots \cap S_n$ esto es, se trata de los semas comunes a todos los sememas del campo (intersección).

Este resultado forma lo que se llama *archise-*

[33] Cf. G. Mounin, "Essai sur la estructuration du lexique de l'habitation", *CLex*, 6, 1 (1965), 9-24.

[34] Propuestas especialmente por E. Coseriu en "Pour une sémantique diachronique structurale", *TLL* 2, 1 (1964), 139-186 y "Structure lexicale et enseignement du vocabulaire", *Actes du 1er. Colloque international de linguistique appliquée.* Nancy 1966. Cf. también Geckeler, op. cit.

$$CS_l = \{ s_1, s_2, s_3, s_l \} \left\{ \begin{array}{l} S_1 \{ s_1, s_2, s_3, s_4, s_l \} \\ S_2 \{ s_1, s_2, s_3, s_5, s_l \} \\ S_3 \{ s_1, s_2, s_3, s_5, s_l, s_j \} \\ S_4 \{ s_1, s_2, s_3, s_6, s_l \} \\ \cdot \\ \cdot \\ \cdot \\ S_n \{ s_1, s_2, s_3, s_n, s_l \} \end{array} \right\}$$

mema. El archisemema puede tener un significante que lo convierta en palabra; por ejemplo *asiento* tiene por significado el archisemema del campo ejemplificado y, consecuentemente, se convierte en el *archilexema* del campo. Cuando no hay una palabra que cumpla con esa función, el archisemema puede representarse con cualquier perífrasis.

El prefijo *archi–*, usual entre los teóricos del campo semántico se presta a confusión. Siguiendo el modelo fonológico que inspira a la semántica estructural, conciben el archisemema o el archilexema como el producto de una neutralización entre dos sememas. Me parece que hay una diferencia importante entre el plano de la expresión y el del contenido, diferencia de la cual depende la concepción de la archiunidad: mientras que las neutralizaciones en fonología aparecen determinadas por el contexto fonológico, es decir, el entorno sonoro del fonema que se neutralizará, en el plano del léxico difícilmente se puede pensar que haya una neutralización entre, por ejemplo, *silla* y

sillón determinada por el contexto; lo que determina que a veces se diga *asiento* es la necesidad de utilizar un conjunto de semas o un vocablo de orden más general que los dos específicos. Igualmente se puede ir descendiendo en una jerarquía taxonómica diciendo *animal, animal doméstico, perro o perrito chihuahueño* sin que haya contextos que determinen obligatoriamente su selección. La estructura fonológica no está compuesta de diferentes rangos y no creo que un fonólogo hable de sistemas de archifonemas; el archifonema "emerge" del contexto sonoro; el "archilexema" no "emerge", es un vocablo distinto. De otra manera, cualquier perífrasis de un semema que no se exprese con un vocablo, se acercaría a la "archiunidad".

El problema, una vez que se ha definido un campo semántico o un campo léxico, es que se puede demostrar que los sememas no se agrupan siempre en los mismo campos, ni que los campos se vuelven a agrupar en campos más amplios ordenadamente. Como se comprueba al intentar estructurar campos, hay sememas que pueden pertenecer a varios campos a la vez, según las características de los semas que los compongan y aun según el orden en que aparecen en el interior de la fórmula.

El problema de la posible pertenencia de un semema a varios campos continúa siendo uno de los insolubles de la teoría del campo léxico.

2.3.1. Semas y clasemas.

Los rasgos significativos muestran comportamientos distintos; mientras que algunos de ellos aparecen como independientes y constantes fuera de contexto —constituyen el núcleo sémico de Greimas— otros dependen del contexto y cons-

tituyen las articulaciones de los signos en el sintagma: los semas contextuales o *clasemas*. La fórmula del semema se convierte ahora en:

$$S_i = \left\{ (s_1, s_2, \ldots s_n) \bullet (c_1, c_2, \ldots, c_n) \right\}$$

Los ejemplos comunes de clasemas son los rasgos 'humano', 'animado', 'femenino', etc. de los que se puede ver que pertenecen a una gran cantidad de sememas muy distintos entre sí. En las versiones clásicas del campo léxico el clasema tiene una procedencia diferente a la del sema: en Pottier (y en parte en Coseriu) se obtiene del análisis de las distribuciones de un vocablo en sus contextos; en Greimas el clasema ya se declara totalmente procedente del discurso[35] y, lo que resulta mas interesante, corresponde a un plano de significado diferente de aquel correspondiente al sema: mientras que este último remite "à des systèmes sémiques d'une nature particulierè, dont l'ensemble constitue le *niveau sémiologique* de l'univers signifiant" (Op. cit., p. 50), un nivel del *sentido* (tal como se propuso entenderlo en 1.4.1), el clasema se localiza en la sustancia y la forma del contenido de una lengua particular, es decir, es significado formalizado por la lengua– lo que Greimas llama *nivel semántico*[36]. El clasema ofrece lo que se podría llamar la "sintaxis del significado".

Lo que resulta de esta ampliación de la fórmula

[35] Op. Cit., p. 53.

[36] La división que hace Greimas entre los niveles que llama semiológico y semántico presenta matices interesantes para la semántica que desearía haber podido explorar en este trabajo; en sí, es comparable a las subdivisiones que hace Hjelmslev en el interior de la sustancia del contenido, a pesar de la opinión de Greimas mismo.

del semema es la necesidad de jerarquizar o de ordenar la aparición de rasgos en su interior. El semema se convierte en un conjunto ordenado de semas y clasemas, por lo que ahora se vuelve necesario dar una regla de su ordenación y del tipo de relaciones que aparecen entre ellos. Me parece que esta consecuencia –bien mostrada por Greimas, pero casi intratada por los demás teóricos del campo léxico– merecería un estudio aparte, pues la fórmula deja de representar un simple haz de rasgos para pasar a cuestionar si se trata de relaciones de conjunción o de implicación, o si debiera incluirse en su problemática relaciones negativas, de adyunción, modales, etc.

Se puede ver, como conclusión de todo lo anterior, que los problemas del campo semántico o léxico consisten, fundamentalmente, en las dificultades inherentes a la determinación o cálculo de los semas y, en un nivel superior, en las dificultades que plantea la estructuración mutua de los sememas.

También espero que se haya hecho claro el tránsito de la determinación externa del campo en los hablantes y en los objetos (análisis componencial y Pottier), a su determinación interna, basada en la aplicación del razonamiento estructuralista. En consecuencia lo que viene a definir la moderna teoría del campo léxico es el postulado de la autocontención del sistema, discutido anteriormente (1.1).

2.4. La autocontención del campo léxico.

Si el sistema léxico es autocontenido, la pertinencia de los rasgos significativos no depende de la heterogeneidad propia de las relaciones entre signo y referente, sino del valor que tenga cada sema en el interior de una estructura

total. Es decir, la pertinencia de los rasgos es resultado de un cálculo que realiza el lingüista abstrayendo rasgos de la realidad e infiriendo oposiciones binarias simples.

Pero llegar a manejar los semas en esa forma implica una preparación previa de los materiales que elimine todas las dificultades anteriores. Esa preparación se encuentra en la teoría del campo léxico propuesta por E. Coseriu y resumida en la obra de H. Geckeler[37]:

a) Distinción entre la "realidad extralingüística" (las "cosas") y el "lenguaje" (las "palabras"). Según esta distinción, que es precisamente la condición que, en el plano epistemológico, dio lugar al postulado de la autocontención del sistema, los semas no se determinan a partir del conocimiento de los objetos —como propone Pottier y como lo hace el análisis componencial— sino en el interior de las estructuras de significados de las palabras. Por lo tanto, todo vocablo que pertenezca a una terminología no forma parte de un campo semántico: "El vocabulario técnico corresponde simplemente a una nomenclatura y como tal no está estructurado a partir de la lengua, sino sobre la base de la realidad extralingüística, de los objetos de la disciplina correspondiente. La terminología representa, pues, una clasificación objetiva, estructurada sobre distinciones lógicas." (Op. cit. IV. 1, p. 215). De acuerdo con ella se debe establecer una nítida diferencia entre léxico de la lengua y terminología a que la misma lengua le presta significantes.

b) Distinción entre "lenguaje" (lenguaje primario) y "metalenguaje". Consecuentemente con el punto de vista anterior, también la lengua

[37] Geckeler, op. cit., capítulo cuarto.

natural cuando se usa como metalengua –por ejemplo el discurso de un lingüista– debe excluirse de los datos que se quieran analizar en campos léxicos. La metalengua es también una terminología (Op. cit. p. 220).

c) Distinción entre "sincronía" y "diacronía". La dicotomía saussureana clásica constituye una consideración fundamental para hacer posible el análisis estructural. Dice Coseriu[38]: "Cada estructura debe establecerse en su 'sincronía' propia, es decir, en su funcionamiento y no en el estado de lengua completo, porque eso significaría confundir o identificar arbitrariamente estructuras diferentes, modalidades funcionales autónomas." (Op. cit. p. 223).

d) Distinción entre "técnica del discurso" y "discurso repetido". Para Geckeler "en esta distinción, efectuada dentro de la sincronía, se entiende por *técnica del discurso* los elementos y procedimientos de una lengua libremente disponibles, mientras que el término *discurso repetido* abarca todo aquello que aparece en una tradición lingüística sólo en forma fijada, es decir, expresiones y frases hechas, modismos, proverbios, refranes, wellerismos, citas (también de otras lenguas), etc." (Op. cit. p. 224). El motivo para elaborar esta distinción es que, según Coseriu, en el discurso repetido los elementos léxicos se dan en bloque, como inserciones casi automáticas en el flujo creativo del habla; no implican, según su punto de vista, un uso de la técnica del discurso sino una repetición pura, cuyo valor no se puede obtener haciendo conmutaciones entre esas expresiones y las que provienen de la creatividad.

[38] En "Structure lexicale. . .", citada por Geckeler.

e) Distinción entre "arquitectura de la lengua" y "estructura de la lengua" o entre "lengua histórica" y "lengua funcional". Esta condición previa del trabajo, a la que ya hice alusión párrafos arriba, propone que "sólo dentro de la estructura de la lengua se pueden determinar las *oposiciones;* en la arquitectura de la lengua no domina el principio de la oposición sino el de la *diversidad.*" (Op. cit. p. 224). La arquitectura corresponde a la lengua histórica –es decir, a la lengua concreta que conocen los hablantes– mientras que la estructura es una construcción que se obtiene de la aplicación de las cuatro homogeneidades que forman la *lengua funcional:* la sincrónica (que se resuelve en el punto (c)), la sintópica (geográfica), la sinstrática (social) y la sinfásica (estilo o modalidad expresiva).

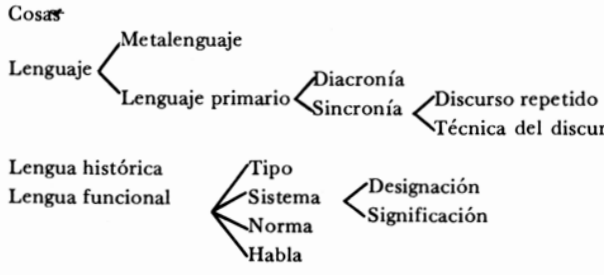
f) Distinción entre "tipo", "sistema", "norma" y "habla". Dice Coseriu: "La distinción que nos parece esencial para la lexicología estructural es la distinción entre *sistema* y *norma* de la lengua. La *norma* comprende todo lo que en la "técnica del discurso" no es necesariamente funcional (distintivo), pero que no obstante está fijado tradicionalmente (socialmente), que es de uso común y corriente en la comunidad lingüística. El *sistema,* por el contrario, comprende todo lo que es objetivamente funcional (distintivo)." (Op. cit. p. 227). Y Agrega Geckeler: "Se deduce claramente que para la semántica estructural importa, en primer lugar, el plano del sistema (sistema entendido como sistema de lo ya realizado y como sistema de posibles realizaciones) como lugar de las oposiciones funcionales." (Op. cit. p. 228).

En otro trabajo he discutido exhaustivamente

el concepto que tiene Coseriu de la *norma*[39]. Lo que vale la pena recalcar en este punto es la importancia primaria del sistema –de la lengua funcional– y, consecuentemente, el lugar secundario en que se toma en cuenta lo social, lo común y corriente en la comunidad lingüística.

g) Por último se introduce la distinción entre 'significación' y 'designación'. La teoría coseriana sostiene que "las estructuras lexemáticas afectan a los contenidos lingüísticos, no a la realidad extralingüística." (Op. cit. p. 228). Es *significación* lo correspondiente a las relaciones estructurales de los vocablos y *designación* lo que relaciona los vocablos con los objetos. El interés de la teoría del campo léxico excluye, por lo tanto, cualquier fenómeno de designación. El signo vale por sí mismo; es autónomo.

Vale la pena aprovechar el esquema (Op. cit. p. 229) que ofrece Geckeler de la sucesión y mutua dependencia de todas estas condiciones, para así comprender mejor su significado para el análisis semántico práctico. Se ofrece una serie de oposiciones dicotómicas de la que habrá que suponer un orden de importancia:



[39] L. F. Lara, *El concepto de norma en lingüística*. El Colegio de México, México, 1976, Cap. 3.

Este planteamiento de la teoría del campo léxico se puede considerar desde tres puntos de vista: epistemológico, metodológico y práctico.

Espero que sea claro que, desde el punto de vista epistemológico, la teoría de Coseriu del campo léxico cumple rigurosamente con cada uno de los pronunciamientos del estructuralismo lingüístico. La distinción radical entre cosas y lengua, que obedece a la definición de la naturaleza del signo lingüístico desde Aristóteles hasta la actualidad, se ve teorizada en la dicotomía del léxico y la terminología, que en el esquema queda borrosa, pero que implica la exclusión total de los términos técnicos como simples nomenclaturas a las que la lengua natural les presta solamente un apoyo material: les ofrece significantes para *designaciones,* ya que el *significado* es un valor estructural del signo completo. (La distinción siguiente entre metalengua y lengua primaria contiene a su vez dos aspectos: en un sentido inmediato, define la metalengua como terminología y por lo tanto la excluye del marco de las preocupaciones del campo léxico; en un sentido mediato y más amplio, reconoce uno de los más importantes elementos de la epistemología moderna, que consiste en la delimitación teórica de la metalengua).

La autonomía de la lengua, tal como se postula en esta teoría, es una comprobación del principio de la autocontención del sistema previamente discutido.

La dicotomía de sincronía y diacronía, a cuya discusión Coseriu ha hecho aportes considerables, es en primera instancia una cuestión metodológica que obedece a la necesidad de abstraer un sistema de los hechos diarios del habla. En ese sentido depende de una dicotomía anterior no in-

cluida en el esquema, pero que campea en este punto y en todos los siguientes: el valor teórico de la distinción entre *lengua* y *habla.* Evidentemente el objetivo de una semántica estructural es construir o descubrir el aspecto sistemático de los hechos del habla. Pero precisamente porque los alcances de las dos dicotomías de *lengua* y *habla* y de *sincronía* y *diacronía* son los que definen el objeto de estudio de la lingüística, su valor epistemológico priva sobre su aspecto limitadamente metodológico.

La exclusión de los hechos que se puedan considerar "discurso repetido", que parece necesaria con el objeto de tener la seguridad de que los términos bajo comparación son verdaderamente comparables, también está de acuerdo con el respeto al principio epistemológico de la sincronía. En efecto, sólo se puede hablar de "discurso repetido" si antes –en el tiempo– ha habido otros discursos y sobre ellos ha habido un juicio normativo por parte de la comunidad. Cuando la comunidad valora segmentos del discurso como dignos de repetición el aspecto tradicional y cultural de la lengua queda en relieve. Así por ejemplo, decir que alguien tiene su "noche triste" implica, en México, comparar a esa persona con Hernán Cortés o lo que le haya sucedido con la derrota de los conquistadores españoles en Popotla. Por supuesto la conmutación con *noche alegre* o *noche feliz* no sirve para establecer una estructura en ese caso. Este ejemplo cae, por lo tanto, entre lo no estructurable y ni siquiera se le podría considerar ejemplo de un rango superior en que se compararan formas de *noche + adjetivo,* pues *¿noche buena, noche vieja* y *noche triste* pertenecerían al mismo paradigma? ¿Se-

rían hechos lingüísticos o más bien reflejos de una tradición cultural en México? Lo llamado "discurso repetido", por lo tanto, es un hecho de diacronía en el uso lingüístico.

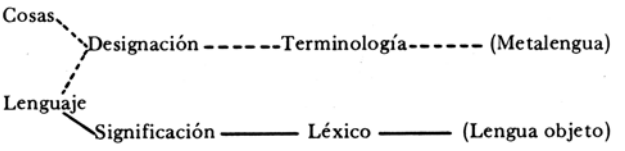
Pero hay también la posibilidad de que la distinción entre el "discurso repetido" y la "técnica del discurso" refleje uno de los elementos fundamentales de la teoría del lenguaje de Coseriu: el aspecto creativo, de *energeia,* que ha defendido en tantos lugares en especial para combatir la concepción estructuralista del *ergon,* el producto lingüístico. Sólo su "técnica del discurso" puede representar a la creatividad. Y aquí surge la cuestión práctica: ¿qué tanto se puede decir de la lengua real, de la que usan los hablantes diariamente si se eliminan locuciones, frases hechas, muletillas, refranes, etc.?

A base de la serie sincronía → técnica del discurso → lengua funcional → sistema se tiene todo el perfil epistemológico de la teoría coseriana del campo léxico. Su objeto de estudio es el sistema estructural, no la lengua histórica ni la lengua efectivamente utilizada por los hablantes. Lo que busca la lexicografía es precisamente lo contrario: los vocablos registrados en el interior de una sociedad, con una cultura y con una tradición: hechos de 'arquitectura', de uso (para no utilizar el término ambiguo de *norma* en este contexto).

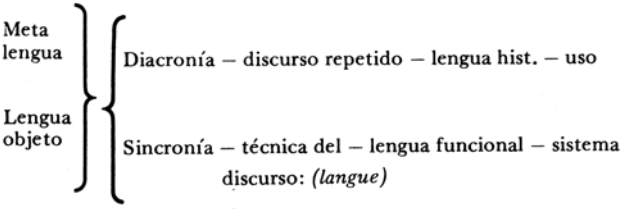
El cuadro de las condiciones previas para el análisis estructural del campo léxico puede, por lo tanto, rehacerse para hacer más clara la combinación de cuestiones epistemológicas y teóricas que contiene:

La teoría de Coseriu conlleva una concepción de la naturaleza del signo lingüístico, basada en la distinción total entre cosas y lenguaje, junto

con su equivalencia de designación y significación:



Al mismo tiempo se hace una diferencia sobre el vocabulario que proviene, ya no de la concepción de la naturaleza del signo, sino de la teoría lingüística del estructuralismo:



Así se podrá ver por qué las condiciones para el análisis del campo léxico no son limitadamente metodológicas, sino que conllevan tomas de posición propias de la doctrina estructuralista en su conjunto. Cada una de ellas es un mecanismo de reducción del objeto de estudio de la lingüística y tiene, como único fin, la elaboración de un *sistema,* como dice Geckeler, "de lo ya realizado. . . y de posibles realizaciones." (Op. cit., p. 228).

2.5. La estructura del vocabulario.

Esta idea del sistema como conjunto de posibilidades (unas realizadas y otras capaces de realizarse) implica la existencia de una estructura del vocabulario tal como la imagina y la requiere la lingüística estructural. La serie de inclusiones

sucesivas de semas y clasemas en sememas y de sememas en archisememas y campos semánticos conduce necesariamente a la sistematización cerrada del vocabulario, sistematización que está postulada en la base misma del estructuralismo.

Bajo una concepción rigurosa del estructuralismo no solamente debe haber una estructura cerrada, sino que las mismas unidades léxicas son reconocibles *si* entran en oposición binaria. Es decir, el *objeto* de conocimiento del estructuralismo es la estructura.

Por esas razones no era extraña la pretensión —que se adivina en múltiples obras de semántica— de llegar a proponer una estructura total de los semas, arreglados de los más generales o "primitivos" a los más particulares. Se ha querido buscar una pirámide conceptual de cuyas combinaciones se obtengan todos los semas y sememas posibles en una lengua. El sistema es así "de lo dado y de lo que se puede dar".

De allí se deriva una consecuencia importante para la fórmula del semema: si los semas están en la virtualidad del sistema, el semema solamente organiza semas preexistentes; toda la tarea de la investigación consiste en probar tantas veces como sea necesario la organización más conveniente de los clasemas, por un lado, de los semas por el otro, y del juego de inclusiones sucesivas de unos campos en otros más amplios hasta llegar a la cúspide de la pirámide.

El léxico de una lengua llega a convertirse en una gran virtualidad, de la cual solamente algunos elementos (o muchos elementos) se *realizan* en la lengua histórica. El vocabulario que recoje el lexicógrafo no pasa de ser, por lo tanto, más que un conjunto de realizaciones del sistema léxico. La lexicografía sólo registra realizaciones,

en tanto que es la semántica o la lexicología estructurales la que legítimamente puede describir el sistema léxico.

2.6 Los problemas prácticos.

Pero posiblemente todas las consecuencias que se han mostrado respecto de la teoría estructuralista del léxico no serían tan importantes si la práctica del análisis en campo semántico fuera sencilla, ampliamente aplicable y produjera resultados de la magnitud requerida por la lexicografía (que es decir el hablante común y corriente, si es que la lingüística alguna vez ha de tomarlo en cuenta).

El hecho es que me parece que la práctica resulta difícil y las más de las veces insatisfactoria.

Hasta hoy hay brillantes análisis "de gabinete" sobre campos semánticos privilegiados; han servido sobre todo para ilustrar aspectos teóricos más que para producir una descripción general del léxico. Pero ¿qué implica practicar el análisis semántico estructuralista? Implica homogeneizar inmediatamente datos variables y diversos bajo los cuatro puntos de la *lengua funcional.* El continuo del discurso natural de los hablantes se fragmenta en varias sincronías coexistentes, en varias sintopías que se entremezclan en la cadena, en referencias alternantes a distintos estratos lingüístico-sociales, y en un constante ir y venir entre los varios niveles de lengua o estilos de que dispone un hablante. Diez minutos de habla natural ofrecen dificultades de homogeneización casi insuperables. El recurso de ampliar la muestra de datos para contar con corpora suficientemente grandes como para deslindar "arquitecturas" en el habla que interfieran con las estructuras, solamente vuelve las dificultades más dila-

tadas pero no permite ver soluciones practicables.

En el habla todo se entremezcla; se repiten elementos del discurso, se hacen referencias a la realidad extralingüística, se utilizan terminologías, dejan de cubrirse elementos pertinentes para un campo determinado, se recuerdan estados anteriores de la lengua histórica. Que lo que busca el estructuralismo es solamente la sistematicidad invariante del aparato que permite el habla común; que la reducción científica se legitima desde el momento en que se reconoce la multiplicidad de los fenómenos lingüísticos y la necesidad de abstraer unos a costa de otros para alcanzar a constituir un objeto de conocimiento, no es algo que pueda aceptarse absolutamente, sino sólo como momentos en la historia de la ciencia. Nada de lo que hoy sabemos en lingüística podría haberse alcanzado sin la teoría estructuralista; pero al mismo tiempo lo que hoy conocemos de ella nos permite intentar resolver la anomalía —como dice Thomas Kuhn— que se produce por una definición demasiado limitada de los fenómenos léxicos.

3. Condiciones para el análisis semántico en lexicografía.

La concepción de la naturaleza del signo lingüístico que se discutió en la primera parte de este trabajo es fundamental para el desarrollo de la práctica lexicográfica. A partir de ella se intentará resolver las dificultades más importantes de la teoría del campo léxico con el objeto de aprovechar la enseñanza del estructuralismo sin caer nuevamente en sus fallas.

3.1. Los materiales para el análisis.

Para el lexicógrafo los fenómenos lingüísticos se prestan en toda su diversidad y complejidad. Textos hablados o escritos, procedentes de fuen-

tes no siempre controladas y verificables como lo desearía un lingüista quisquilloso; hablantes de cuya procedencia geográfica y social, o sobre cuya edad no se conocen los datos; discursos en los que, sea por una voluntad de estilo o sea porque, a fin de cuentas, así son siempre, se ofrecen constantes citas de discursos anteriores, creaciones metafóricas, significados plurales, son el material de trabajo cotidiano en lexicografía.

Prácticamente es imposible contar con un corpus previamente homogeneizado, sobre el cual se puedan hacer oposiciones estructurales rigurosas y compatibles; prácticamente también, el arreglo alfabético de los vocablos del diccionario orienta el trabajo de análisis y obliga a partir de datos singulares sobre los vocablos, sin que que sea fácil contar con toda la documentación de un campo léxico antes de iniciar el trabajo.

Teóricamente la postulación del sistema como algo virtual, como algo cuya realización es secundaria respecto de su valor teórico, tiene una consecuencia inquietante para el lexicógrafo: dada la fragmentariedad natural de sus datos, o dada la accidental historia del vocabulario de una lengua, a falta de un vocablo que, en el cálculo, debiera completar un campo o debiera cubrirlo como archilexema, ¿toca al lexicógrafo "realizarlo" (inventarlo)? ¿Es parte de la lexicografía desarrollar "posibilidades del sistema" que no se hayan realizado? ¿De qué habla la lexicografía: de la estructura lingüística o de una lengua histórica?

Frente a todas estas cuestiones se vuelve a plantear la pregunta sobre la pertinencia de la teoría estructuralista para una disciplina como la lexicografía. O al revés, se puede plantear la cuestión de si lo que constituye el objeto de tra-

bajo del lexicógrafo es ajeno a lo que ocupa al lingüista y, por lo tanto, el pensamiento teórico que genera la lexicografía da lugar a una versión opuesta y contradictoria con la teoría lingüística.

Lo que intento es reunir en un esbozo de teoría un objeto común para la lingüística y la lexicografía. No solamente sacar provecho de la enseñanza estructuralista y de la cotidianeidad de la lexicografía, sino ofrecer una salida teórica y prácticamente aceptable a los callejones sin salida expuestos páginas antes.

3.2. La metalengua lexicográfica.

Hagamos la convención de que por "discurso lexicográfico" se entienda, fundamentalmente, el texto de cada uno de los artículos que componen un diccionario[40] (también habría de considerarse el arreglo de los artículos en una macroestructura, el texto de explicación de la factura y el uso del diccionario y las tablas en que, en algunas obras, se exponen patrones gramaticales, de conjugación, de ortografía, etc. Para los fines que persigo me restrinjo a lo primero).

El discurso lexicográfico es, a primera vista, un texto que trata de la lengua natural; la lengua natural es la lengua objeto de una metalengua; el plano del contenido de la metalengua lexicográfica es la lengua natural. Esta metalengua cumple, *en principio*, con la condición de Hjelmslev para los lenguajes científicos: que se defina por una operación acorde con el principio de empirismo: no contradicción, exhaustividad, sencillez[41].

[40] Por *artículo* entiendo la *entrada*, las marcas gramaticales y de uso, la estructura de acepciones, de ejemplos, de construcciones, etc. que componen un todo.

[41] L. Hjelmslev, *Prolegómenos a una teoría del lenguaje*. Trad. J. L. Díaz de Llaño, Gredos, Madrid, 1971. Cap. 3.

Cabe discutir si realmente el discurso lexicográfico es o puede ser un lenguaje científico. Se puede suponer que la tradición lexicográfica, atenida siempre a fuertes obligaciones del orden práctico y determinada por la necesidad de hacerse inteligible a los hablantes, no se haya preocupado por hacer de su propio discurso una red descriptiva coherente, en que cada uno de los componentes del artículo se ajuste al resto sin producir contradicciones. También el hecho de que lo que presenta un diccionario es un universo abierto de palabras, puede hacernos dudar de la posible exhaustividad del discurso lexicográfico. Por último, el criterio de sencillez es, como siempre, algo difícil de medir entre diccionarios: ¿es el arreglo de acepciones y de entradas y subentradas más sencillo o más complicado si se le dan soluciones "distribucionalistas" —como en el *Dictionnaire du Français Contemporain*[42]— o no, si por cada vocablo se hace una entrada o no, si se opta por la homonimia o por la polisemia, si se hace una prolija lista de marcas de uso o se deja al criterio del lector la calificación de los usos a partir de contextos *hic et nunc,* etc.? Sin embargo difícilmente se encontrará un diccionario —excepto, claro, los pastiches comerciales que inundan los supermercados— en que sus autores no hayan intentado establecer guías sistemáticas tanto para la definición lexicográfica como para la inclusión de ejemplos, locuciones, recciones verbales, usos gramaticales, abreviaturas y marcas de uso. La exhaustividad de las descripciones, la necesidad de que todos los vocablos recogidos sean com-

[42] De Jean Dubois et al. Larousse, París, 1971. Su justificación aparece en J. y C. Dubois, *Introduction a la lexicographie: le dictionnaire,* Larousse, París, 1971.

prensibles desde la metalengua, indican que el discurso lexicográfico es muchas veces más exhaustivo que las descripciones lingüísticas, para las cuales el ámbito especializado en que circulan impone menos requerimientos o permite más libertades provisionales. Por eso sostengo que, al menos en principio y sin querer cerrar o agotar esta discusión, la metalengua lexicográfica es un lenguaje científico.

Más acá de ese tema, son las precisiones que se pueden hacer sobre los problemas de la metalengua lexicográfica las que más interesan en este trabajo. Como sucede en la lingüística, el principal de ellos es el fenómeno peculiar de que la metalengua es, por sobre su definición en un nivel epistemológico diferente, la lengua natural. Es decir, la lengua natural es una metalengua para el estudio y la descripción de la propia lengua natural. Esa característica semiológica de la lengua natural, que la caracteriza frente a todos los otros lenguajes como —en palabras de Hjelmslev— "passe par tout", impone una dificultad teórica y metodológica muy grave al trabajo lingüístico: se presenta como una total falta de espacio entre el objeto de estudio y el instrumento de análisis. Para el lingüista y el lexicógrafo como para el hablante común la inmediatez de su lengua a su conciencia, la tautología aparente que se crea entre su lengua como objeto y su lengua como instrumento científico, implica la creación, a lo largo de su experiencia profesional, de un método de extrañamiento que consiste en alejarse, paso a paso, de su conocimiento lingüístico inmediato hacia la objetivación necesaria de la lengua como materia de estudio. La principal cuestión de la formación lingüística es precisamente lograr el espacio necesa-

rio entre ambos niveles de la lengua natural; hacer que la lengua como capacidad del hablante y la lengua materna de cada lingüista se extrañen y permitan la tematización de lo que, en palabras de Coseriu, ya está en el "saber original" de los hablantes en cuanto humanos.

De no mediar ese espacio el lingüista se encontrará en muy pobres condiciones para tratar objetivamente los fenómenos lingüísticos. El problema no es solamente de aquellos que estudian su propia lengua materna; también aparece entre los que se dedican a lenguas "exóticas" a su cultura, como lo demuestra el papel que juega la traducción en la recolección de materiales en el campo, en el establecimiento de paradigmas entre clases de palabras y hasta en la capacidad descriptiva de un fonetista para el que, provenir de una lengua con vocales nasalizadas —por ejemplo— le hace más sencilla la ocupación con nasalizaciones en una lengua diferente de la suya. Un buen ejemplo del papel de la traducción en la descripción de lenguas desconocidas lo puede ofrecer la obra de B. L. Whorf, en la cual más de una vez lo que se muestra como extraño en la lengua estudiada se destaca solamente por la intervención de una traducción previa al inglés, no siempre claramente reconocida. En trabajos como ese se percibe claramente la posibilidad de que se atribuya a una lengua en estudio propiedades que no le son inmanentes sino de la lengua utilizada como metalengua.

En lexicografía el problema se presenta desde que comienza uno a hacer el análisis de los significados de un vocablo, y por eso buena parte del tiempo de preparación y de reflexión del lexicógrafo se emplea en el ejercicio de extrañarse de su lengua objeto, en la crítica permanente de su

actividad metalingüística. Este ejercicio de extrañamiento implica, por una parte, la revisión a cada instante del idiolecto del lexicógrafo, con el objeto de no hacer pasar a su análisis preconcepciones o ideas particulares suyas; por la otra, cuidarse de no llevar el extrañamiento hasta el punto en que las palabras pierden todo valor significativo y empiezan a presentarse como emisiones irracionales de sonidos. Hay un juego infantil que ilustra muy bien este extrañamiento extremo: si uno pronuncia muchas veces la palabra *tigre,* al cabo de un rato ya no se sabe si la palabra es *tigre* o *trigue;* la palabra deja de tener una identidad y se convierte en una caprichosa y extraña secuencia de sonidos, respecto de la cual el significado 'tigre' se ha desligado totalmente. El lexicógrafo sufre estas dos experiencias con cada una de las palabras que tiene que analizar; es un proceso ineludible que además conviene experimentar a diario si uno no quiere perder toda posibilidad de hacer lexicografía.

Me parece que en este problema de la metalengua está una de las características fundamentales de lo que Alain Rey llama justamente "le fait dictionnaire", pero también creo que se trata de una cuestión más general que toca a la lingüística en su conjunto. A pesar de ello, es algo a lo que se ha concedido muy poca atención.

En este sentido son muy comprensibles los esfuerzos realizados a lo largo de la historia de la lingüística por construir una metalengua no solamente controlada, sino neutral respecto de cualquier lengua natural. Los intentos de la glosemática por elaborar una metalengua simbólica y los actuales en lingüística computacional, en lingüística matemática y en inteligencia artificial forman parte de lo mismo; de lo que no puedo

estar seguro es que tengamos hoy en día mucha conciencia de sus implicaciones epistemológicas, teóricas y prácticas.

El problema de la metalengua es también característico del estructuralismo; por su misma clausura del sistema respecto del mundo externo, por sus aspiraciones a convertirse en ciencia positiva, el asunto de la metalengua se trató de resolver o bien desde una posición comparable a la ciencia natural –Bloomfield–, o bien ignorando (y por lo tanto mezclando) la necesaria diferencia entre la estructura construida por el lingüista y la estructura que muchos han creído *descubrir* realmente en una lengua natural. El criterio de objetividad en lingüística no puede ser el mismo que en la ciencia natural desde el momento en que, como hablantes, *conocemos* la lengua; el extrañamiento metódico respecto de ella no consiste en negar nuestro conocimiento –nuestra *competencia,* dirá Chomsky– originario, implícito si se quiere, y colocarnos delante de ella como si fueran cadenas de sonidos indistinguibles de otros que hay en la naturaleza. De lo que se trata es de liberarnos de la inmediatez de nuestro conocimiento originario, damos cuenta de que la introspección, por inmediata a la conciencia, no puede ofrecer puntos de referencia segura al análisis de la lengua, pero a la vez de que en toda recolección "objetiva" de los datos hay una interpretación previa de lo que es la lengua, presente aún en la "heurística" con que la lingüística blomfieldiana trató de eludir este problema constitutivo de nuestra ciencia.

Si el lingüista que se ocupa del sistema lingüístico se enfrenta más tardíamente a la cuestión de la metalengua, el lexicógrafo no solamente lo hace de inmediato, sino que además descu-

bre que su misma finalidad: la producción de un diccionario, lo obliga a conservarse en el interior de la crisis; el lexicógrafo tiene, por ello, que asumir la totalidad del problema como una cuestión de teoría y de método; tiene que situarse enmedio de la tensión creada entre el extrañamiento y la enajenación de y por su conocimiento originario de la lengua. Cabe preguntar al lingüista ¿no es verdad que, en este punto, la lexicografía destaca una cuestión de principio para la aclaración de la posición de la lingüística entre las ciencias?

Por todas esas razones se puede postular como cuestión primaria para la teoría y la práctica lexicográfica que, puesto que los objetivos de la lexicografía trascienden al ámbito de la ciencia hacia la sociedad, la metalengua lexicográfica *es* la lengua natural y, en consecuencia, debe someterse a una tematización crítica que abra el espacio entre lengua objeto y metalengua sin salir de él.

Un corolario de lo anterior es el papel que juega el lexicógrafo en cuanto sujeto del análisis lexicográfico y de la práctica de la definición en los diccionarios; si el ideal positivo de la lingüística es llegar a desdeñar totalmente al lingüista como sujeto, la realidad lexicográfica es incluirlo entre sus problemas y colocar los temas de la introspección, la descripción objetiva, la interpretación y la crítica en el centro de su constitución científica.

3.3. Elementos de la metalengua lexicográfica.

A partir de las reflexiones anteriores se puede pasar ahora a la presentación de los elementos más necesarios de la metalengua lexicográfica: es decir, a una serie de postulados que explican el

comportamiento y las características del análisis semántico en lexicografía.

3.3.1. La noción de igualdad.

Es un tópico clásico de la semántica lingüística el debate sobre la existencia o inexistencia de la sinonimia. Esa posibilidad de igualar dos signos lingüísticos en el dicurso común y corriente, que se avala en los diccionarios de sinónimos, ha sido ampliamente debatida por los lingüistas. El ejemplo bien conocido de la designación del número '70' en francés ilustra el problema y sus respuestas. Para '70' hay los signos *soixante-dix* y *septante,* que resultan referencialmente sinónimos; es decir, ambos designan el número '70'. No obstante, *soixante-dix* es un signo del francés de Francia y *septante* del francés de Bélgica. Esta importante diferencia, no en el orden referencial sino en el sintomático (marca a un hablante como francés o como belga) basta para sostener que en cierta medida no son sinónimos. Igualmente se podría decir que *pavo* y *guajolote* en México son referencialmente sinónimos; sin embargo cuando se habla del *pavo de Navidad* muchos hablantes mexicanos considerarían equivocado sustituirlo por *guajolote de Navidad* y, a la vez, cuando se come *mole de guajolote,* preferirían no sustituirlo por *mole de pavo* (¡perdería el sabor!). Se puede también concluir que no son del todo sinónimos, pues aparecen o bien en dos niveles de lengua diferentes o bien en distribución complementaria.

La respuesta del lingüista de que en ningún caso hay verdadera sinonimia, pues siempre aparecen matices que distinguen un signo de otro, con ser válida, también deja de reconocer que para los hablantes la sinonimia es algo muy natural en su habla diaria y que, en muchos casos,

ellos mismos utilizan dos signos como perfectamente sinónimos.

El diccionario plantea el problema en otra forma: sólo porque es posible pensar que dos signos sean sinónimos se puede establecer la ecuación sémica entre el vocablo de entrada y la definición lexicográfica en un artículo. En otras palabras, el diccionario se basa en alguna concepción de la sinonimia.

Se puede distinguir entre una idea de la sinonimia fuera de contexto (sinonimia estructural, por llamarla de alguna manera) y en contexto (sinonimia en el discurso). Igualmente se puede atribuir la explicación de la sinonimia a la lingüística (como han hecho quienes sostienen que la paráfrasis explica a la sinonimía (¿!)) o a una instancia diferente. Indudablemente la sinonimia estructural es resultado de la atomización estructuralista discutida en la segunda parte de este trabajo y, ni explica la posibilidad de la ecuación sémica en los diccionarios, ni permite tratar la sinonimia en contexto. Por otra parte, la sinonimia en contexto no es causa sino efecto de una noción de la sinonimia, que puede o no puede explicar la lingüística. Si pudiera, como creen algunos autores de la llamada "lingüística del texto", ser la paráfrasis la que la explicara, entonces alguien tendría que explicar qué hace que un signo sea paráfrasis de otro. Por eso prefiero proponer que lo que funda la sinonimia está fuera del ámbito lingüístico, en algún lugar de la percepción, lógica y genéticamente anterior al lenguaje. En la percepción debe estar la posibilidad humana de establecer relaciones de *igualdad*[43] entre dos objetos presentes a nuestros sen-

[43] No puedo formular de una manera más clara lo que todavía

tidos. Es esa igualdad la que explica la sinonimia y es esa igualdad la que explica la paráfrasis; hay paráfrasis –para decirlo sintéticamente– porque hay sinonimia y no al revés.

El establecimiento perceptual de la igualdad también es anterior a la *función semiótica.* Esta aprovecha las posibilidades que le ofrece aquélla.

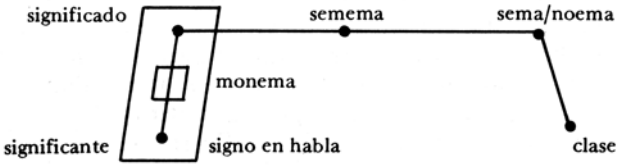
Si se acepta lo anterior, es posible que la función semiótica, dirigida por las necesidades de significación de un discurso, establezca relaciones sinonímicas entre dos signos en forma arbitraria.

3.3.2. La verbalización de los semas.

Por las razones explicadas en 2.2, el estructuralismo intentó descomponer los significados de los signos –en este caso *palabras*– en rasgos mínimos significativos que aseguraran haber llevado el análisis hasta sus elementos más simples. El sema que resulta del análisis estructuralista es, por lo tanto, resultado de una serie de oposiciones binarias que, a cada momento del análisis, van despegándose más y más del significado de partida. Un haz de semas y, como se vio posteriormente (2.3.1) un conjunto ordenado de semas y clasemas se convierten en sememas del signo analizado, sememas que abandonan su ligazón con un signo particular para convertirse en unidades de análisis en el sistema de la lengua. Así, dice Greimas: "le concept de 'sémème construit' libère ... la description du contenu des dernières attaches que ce dernier pouvait avoir avec la manifestation discursive: le sémème ainsi conçu est une unité de contenu, indépendante de sa couverture

me resulta confuso. Por "igualdad" me refiero a similitudes, analogías, simetrías, identidades, entre objetos del sentido en órdenes perceptuales diferentes: visión, oído, tacto, etc.

lexématique et de son entourage contextuel."[44] El modelo de trapecio de K. Heger indica claramente esa posición sistemática del semema:[45]



El semema es, en consecuencia, una unidad lingüística construida en la metalengua. Ahora bien, los semas que lo componen –en el caso de los noemas, que no he discutido aquí porque este es un trabajo semasiológico, el problema no se plantea más que en último término –o bien pertenecen a una metalengua simbólica, como los operadores de la lógica, o bien se verbalizan en términos como 'dimensionalidad', 'proceso', 'causa', etc. De esta verbalización se puede decir, también, que es con puros fines ilustrativos, pero que es neutral a cualquier lengua; es decir, la verbalización no desempeña ningún papel en la metalengua del análisis. Pero cuando uno observa, como en el ejemplo citado en 2.2 y 2.2.1, que elementos como 'para', 'con', etc. también podrían analizarse y que, pese a todo, las nociones de 'causa' o de 'proceso', etc. no son entidades puramente simbólicas sino que conllevan una interpretación al menos científica del mundo, la

[44] Op. cit., p. 85.

[45] Cf. Heger 76a. donde se explica claramente la diferencia entre *significado*, ligado a un *significante* por el postulado de la "consustancialidad cuantitativa", obtenido por Heger de la famosa metáfora saussureana de la hoja, y el *semema*, unidad de lengua.

verbalización deja de ser indiferente y accidental para plantearse como el último plano del análisis semiológico, como un horizonte de sentido en donde, agotado el lenguaje formal-simbólico, reaparece el sentido y reaparece la lengua natural. La conclusión de Greimas de que "toute recherche portant sur les significations inhérents à une langue naturelle reste enfermée dans ce cadre linguistique et ne peut aboutir qu'à des expressions, formulations ou définitions présentées dans une langue naturelle" (Op. cit., p. 13) se puede extrapolar a la metalengua de la lingüística y, en concordancia con Hjelmslev, sostener que sea la "lengua de todos los días" la que traduce y explica al resto de los lenguajes, especialmente a la metalengua de la lingüística.

La lexicografía es la que permite observar desde un lugar privilegiado la forma como se cierra el círculo entre la lengua natural y la metalengua. Semas y sememas no son construcciones metalingüísticas ajenas y neutrales respecto de la lengua objeto; son a su vez interpretaciones científicas verbalizadas. No se trata, con esta afirmación, de proponer soluciones simplistas como la del análisis componencial ni como la de la semántica generativa, en que la verbalización no es objeto de estudio ni implica una investigación previa sobre la naturaleza de los signos lingüísticos; se trata de no olvidar que más allá de nuestra metalengua lexicográfica está una interpretación basada en el *sentido* y que, puesto que la metalengua es la lengua natural, el objeto del análisis lingüístico y lexicográfico es abrir el espacio entre esos dos niveles, con la conciencia de que hay una tensión constitutiva del trabajo lingüístico.

Bajo estas consideraciones se puede proponer

que el análisis semántico se realiza con verbalizaciones de los rasgos significativos y que, por lo tanto, en todo análisis hay una interpretación guiada por los objetivos que se plantea el trabajo.

3.3.3. La ecuación sémica.

Por lo tanto, si bien es posible que la ecuación entre el vocablo de entrada y su definición lexicográfica –llamada por J. Rey Debove "ecuación sémica"[46]– se sostenga por un concepto de la sinonimia basado en la percepción humana de relaciones de igualdad (3.3.1), también lo es que cada vez que se sustituye el vocablo de entrada por su definición, aunque la lectura de un contexto cualquiera se conserve "la misma", en la verbalización de la definición lexicográfica aparece un *sentido,* un "lo mismo" necesariamente interpretado en forma distinta. Digamos que el diccionario presenta el mismo tipo de paradoja que presenta la traducción: entre dos textos en lenguas distintas, la traducción es la misma porque hay un *sentido* igualmente recuperable; no obstante, ninguna traducción puede reproducir un texto en forma idéntica; en ambas lenguas hay una interpretación. La tensión entre el "lo mismo" que asegura el *sentido* y lo distinto que se crea en la verbalización es un punto de crisis constante de la lexicografía. En torno de él se elabora todo el aparato metodológico que guía la práctica diaria.

Los diccionarios, en consecuencia, se desarrollan como textos a su vez, en los que las definiciones, aunque operan sobre el principio de la ecuación sémica, agregan significación y carac-

[46] Op. cit., especialmente el capítulo 6.

terizan a un diccionario particular frente a los otros. Greimas, nuevamente, ha hecho notar esta peculiaridad, no respecto de la lexicografía, sino del análisis semiológico en general: "Le décalage qui se produit entre l'ensemble signifiant premier et sa traduction intéresse non seulement la sémantique, mais toute deiscipline de signification: la distance qui les sépare peut être interprétée comme créatrice d'alienations et de valorisations" (*op. cit.*, p. 13).

Lo anterior se puede interpretar con la ayuda de los postulados de arbitrariedad y función semiótica discutidos en la primera parte. La arbitrariedad es la que permite que, respecto de un sentido, se produzcan signos diferentes; la arbitrariedad del signo es la que permite explicar dentro de la teoría la posibilidad ilimitada de construir signos diferentes para sentidos "iguales" —aunque en rigor aun los sentidos cambian, porque cambia la cultura y la tradición desde donde se los significa—; la función semiótica es la que produce el establecimiento de relaciones entre el signo y el sentido. Ambas están en la base de una teoría semántica y semiológica.

Pero puesto que las cosas se ven así y ahora las relaciones entre signo y referente se ofrecen como aleatorias, como susceptibles, por definición, de cambiar en cualquier momento, la idea de que los signos designan biunívocamente los objetos no puede sostenerse. Tampoco, en consecuencia, la posibilidad de que los signos conduzcan directamente a lo que designan. Entre referente y signo hay un amplio espacio.

Según lo expuesto aquí ese espacio se llena de interpretación; tiene un *sentido*. Que la lingüística descriptiva haya podido creer que su

ocupación con los signos era neutral respecto de la interpretación y la valorización, sólo conduce al error. El lingüista no puede quedarse al margen del fenómeno que estudia; no es un científico comparable al entomólogo frente a los insectos que estudia. Cada paso de su trabajo interpreta y valora. De lo que se trata ahora es de asumir esa posición y hacer de ella objeto de crítica y de teoría.

Pero la misma jerarquía expuesta en 1.4 sobre la arbitrariedad indica que, en el nivel de la lengua concreta, en donde se mueve el hablante, las relaciones entre signo y referente se fijan por la acción de la cultura y la tradición. Por ello la crítica de la interpretación de la ecuación sémica tiene que comenzar considerando un espacio y un tiempo sociales, culturales e históricos que determinan, tanto al signo en estudio, como al que lo estudia. El análisis semántico en lexicografía resulta necesariamente basado en la situación que lo rodea. El diccionario deja de ser una obra de referencia atemporal y neutral; es, como lo demuestra la historia de los diccionarios, un documento de su tiempo y de sus autores. En los mejores casos, cuando los lexicógrafos cobran conciencia de ese hecho, la definición lexicográfica abre su estructura y resalta la interpretación que la funda; es decir, permite la crítica por parte del lector; no neutraliza, sino que ofrece la posibilidad de varias interpretaciones críticas. En los peores casos oculta su interpretación y adopta una postura coercitiva frente a sus lectores; deja de ofrecer la posibilidad de recrear los significados y, por lo tanto, de que el hablante individual ejerza su libertad y más bien legisla, 'autoriza' y 'desautoriza' sentidos: cae una vez más en el nomenclaturismo.

3.4. La estructura semántica.

Si por todas las razones expuestas el trabajo del lingüista no puede dejar de reconocer la presencia de una interpretación que conduce su análisis, resultará claro que no se puede hablar de estructuras del significado en la misma forma en que lo hacía el estructuralismo. Para este, el sistema de una lengua constituye una red de relaciones que se contiene a ella misma. El hablante y el lingüista solamente pueden *realizar* —en el caso del hablante— las posibilidades virtuales que ya ha definido previamente el sistema o *descubrir* —en el caso del lingüista— unas estructuras preexistentes, sincrónicamente inmutables y diacrónicamente inexplicables. El sistema lingüístico es tan poderoso, tiene tal capacidad de predecir lo realizable, que la lengua histórica no puede pasar de ser un accidente del sistema. Los órdenes tradicional y cultural no son capaces de alterar directamente al sistema autocontenido, pertenecen inevitablemente a la "lingüística externa".

Si, en cambio, se parte de la idea de que lo sistemático es un nivel intermedio de la lengua natural, interpretado eso mismo por la metalengua de la lingüística y ésta, a su vez, por el *sentido* que proyecta la vida histórica, sentido que descubrimos en una última verbalización en lengua natural, se puede llegar a la conclusión de que las estructuras existen, es verdad, pero que dependen del arreglo que genere la función semiótica. Dicho en otra forma, las estructuras lingüísticas son resultados de las necesidades de significación de los humanos y solamente por un artificio —legítimo, pero artificio al fin— de método podemos aislarlas y conferirles una cierta capacidad de predicción. Esta capacidad no es, en

cambio, característica de la estructura, sino que la estructura es característica de la tensión entre código de información y significación que se discutieron en 1.5.

La estructura semántica no se puede entonces concebir como algo dado y ordenado de una vez por todas, sino como algo que se produce cada vez que se significa. Su relativa fijeza no revela un carácter estructural, sino su pertenencia al orden tradicional e histórico de la arbitrariedad saussureana. Lo que muestra un análisis en campo semántico son estadios históricos de una lengua y la posibilidad de predicción que ofrece la estructura está en relación directa con la documentación real y posterior que se encuentre. Así por ejemplo, no tiene sentido hacer (como escuché en un congreso hace tres años) un cálculo de cómo "debiera" haber evolucionado el francés desde el latín vulgar porque de hecho no evolucionó así y esa virtualidad no tiene ningún sentido (ni siquiera pedagógico, como alegaba su autora).

El análisis en campo semántico por eso no puede resolver los problemas de la delimitación de los campos, de los huecos o fallas de su estructura, de su superposición con otros, del arreglo jerárquico de sus elementos, etc. Se debiera poder resolverlos si las estructuras lingüísticas antecedieran eternamente a sus realizaciones, pero en ese caso la sombra del nomenclaturismo —de la clasificación ordenada de los objetos del mundo o de los significados— volvería a cernirse sobre la naturaleza misma del objeto de estudio de la lingüística.

Porque el léxico no es así, porque las estructuras léxicas que uno obtiene en el análisis son construcciones abstraídas de los datos reales y no son

sino modelos que se crean sobre ellos a partir de la epistemología estructuralista, es por lo que el análisis en campo semántico se presenta a nuestros ojos tan cuestionable.

En cambio, si con el concepto saussureano de la arbitrariedad se sostiene la posibilidad de que las relaciones signo-objeto cambien en cualquier momento y, a la vez, con el mismo concepto como pivote se espera la conservación de las relaciones entre signos como resultado de la tradición, y el cambio de las relaciones como efecto de la significación, las estructuras léxicas quedan en la posibilidad de variar. El análisis estructuralista del campo semántico puede redefinirse en estos nuevos términos: la estructura sémica obtenida no es algo desligado de la historia ni parte de un sistema de posibilidades, sino resultado de la aplicación de un cierto tipo de análisis sobre discursos concretos en el tiempo. Por eso cada vez que se hace un análisis con nuevos datos o con distintos objetivos, los resultados son diferentes: la función semiótica orienta la estructuración a partir de los intereses de significación de los hablantes en momentos determinados. Para hacer una analogía con algunas ideas de la física, se puede decir que el universo léxico no es estructurado sino difuso y que es el discurso el que ordena los elementos del universo en una microestructura.

La idea de que lo que permite obtener el anásis en campo semántico es una microestructura tiene, por lo tanto, dos ventajas: a) pueden crearse tantas microestructuras como necesidades de significación diferentes; b) no hay una sola estructura del léxico y por lo tanto no hay que buscar una jerarquización total de los campos semánticos; traslapes, mutuas inclusiones, indefiniciones, son síntomas de ese fenómeno.

Se pueden formular efectivamente los campos léxicos porque hay una sistematicidad del código lingüístico en y por la historia; no pueden formularse de una vez para siempre porque la sistematicidad es sólo una condición de la lengua (necesaria pero insuficiente); la otra condición es la significación, pero es ésta la que supera constantemente a la sistematicidad.

3.5. Para redefinir el campo léxico.

Ahora sí se puede intentar reunir todo lo expuesto anteriormente en una nueva definición del campo léxico.

3.5.1. El postulado de la arbitrariedad radical de los signos ha servido para construir un marco general de variabilidad en las relaciones entre signos y objetos. Nada hay que obligue a la teoría a regresar a las posiciones convencionalistas en que inevitablemente la lengua se convierte en una nomenclatura y un juego de reglas de aplicación. Con el principio saussureano de la arbitrariedad se rompen las ligas biunívocas entre signos y referentes y, correlativamente, se gana en capacidad para comprender la autonomía del signo.

Al mismo tiempo los varios niveles de interpretación de la arbitrariedad propuestos por R. Engler permiten insertar en la teoría los planos de la estructura y del uso social e histórico de los signos. Entre otras cosas gracias a esos niveles se puede comprender por qué la lengua natural se presta a la convención —en el caso de las comunidades científicas que definen sus terminologías— y a la naturalización de los signos —en el caso de la llamada etimología popular.

3.5.2. Lo anterior deja las estructuras lingüísticas en un plano en que la importancia del

uso real de la lengua crece y la rigidez del sistema se abre hacia la variación social. Se requiere por lo tanto un concepto del campo léxico en que ambas características se vean representadas.

3.5.3. En contra del postulado de autocontención del sistema lingüístico, que tiene el efecto de volver inexplicable (o de sacar fuera de la teoría lingüística) la relación de los signos con los objetos del mundo sensible, se propone el trasiego constante de la significación entre el sistema lingüístico y el *sentido*. El sentido se concibe como un horizonte en que se proyectan las cosas del mundo sensible en cuanto concebibles, no en cuanto previamente delimitadas y estructuradas. El sentido está fuera de la teoría en cuanto se postula axiomáticamente (o sea, en cuanto la funda), pero también desde ella se puede tematizar a través de la reflexión sobre las relaciones entre la lengua objeto y la metalengua.

Bajo este punto de vista la cuestión ya no es dividir entre un mundo de objetos preexistentes y preestructurados y otro de signos asociados convencionalmente con el primero, sino pensar que el mundo de los objetos se concibe y adquiere sus características solamente mediante la intervención de una verbalización histórica y social que corona el trabajo de una inteligencia inicialmente del orden biológico. Así las ciencias de los objetos y la lengua común y corriente no son sino dos momentos de la significación. Aún más, hay un discurso científico porque hay una lengua natural que le ofrece instrumentos de significación.

Se concluye de este punto que la definición del campo léxico no tenga que separar "estructuras de los objetos" de "estructuras lingüísticas",

sino que se trata de un continuo entre la lengua natural y la terminología científica.

3.5.4. Por la misma variabilidad y actividad de la significación se ha sostenido que en el caso del discurso lexicográfico (pero como se vio es el caso de todos los dicursos), cada vez que se significa algo hay una interpretación distinta del sentido. Si esto es así difícilmente se puede sostener la existencia de estructuras léxicas previamente construidas y arregladas en una jerarquía bien establecida y armónica. Más bien hay que estar preparado a que el universo léxico no esté estructurado y sea la función semiótica la que origine la formación de microestructuras; es decir, el campo léxico vale en cuanto microestructura, pero no puede ser un modelo total, y ajeno a la significación, del léxico de una lengua.

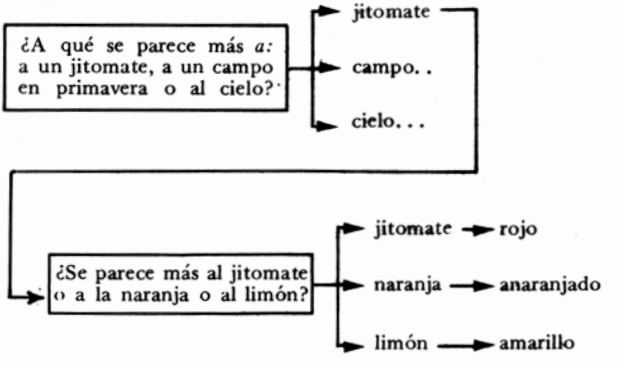
3.5.5. Por último, se ha propuesto que la percepción —anterior a la lengua— se puede basar en distintos tipos de relaciones entre los objetos que se propone conocer y que, en consecuencia, cabe esperar arreglos muy distintos del léxico según lo que haya destacado la percepción.

3.5.6. Aunque con una posición fundamentalmente distinta de la de B. Harrison (Cf. n. 14), pues en este trabajo la noción de *signo* es la más importante, mientras que él aparentemente la ataca (cap. 2), me parece que el campo léxico puede redefinirse con la ayuda de sus *esquemas taxonómicos.*

Una taxonomía en el sentido en que él la entiende es un conjunto de descripciones de objetos junto con una serie de reglas que le permiten a uno reconocer especies nuevas de la misma clase. Cuando se construye una taxonomía sobre cada uno de los objetos puestos bajo consideración

se aplica una serie de preguntas sobre las características de ese objeto. Estas preguntas se responden en forma binaria con *sí* o *no*. Cada vez que la respuesta sea *no*, la taxonomía ofrece o bien una solución clasificada o bien un nuevo camino de interrogaciones hasta que el objeto queda bien definido. En este sentido la taxonomía es generativa, pues con la aplicación de sus reglas y en un tiempo limitado se llega siempre a clasificar un objeto. A la vez uno no tiene que conocer previamente toda la estructura, pues ésta es algo que uno crea aplicando sus reglas. Es decir —extrapolando al tema del campo léxico— el sistema léxico no existe de por sí, sino que se genera en el momento en que es necesario.

Veamos el funcionamiento de un esquema taxonómico para el campo de los colores, adaptado del ejemplo que da Harrison:



En el ejemplo, *a* constituye un problema, pues no puede tratarse de un objeto (en este caso *el* color rojo) previamente deslindado en el mundo de los objetos, ya que implicaría reconocer esa

división entre la "estructura" de los objetos y la estructura lingüística que se ha negado. Por eso *a* no es todavía un objeto clasificado . Como dice Harrison, el esquema taxonómico no es "a device for recording the empirical discovery of nameables" (p. 75), sino para *crearlos,* es decir, significarlos. Esta propuesta concuerda con el postulado del sentido y de la significación que he discutido antes. Pero a la vez se plantea ahora el problema de qué clase de "objeto" es *a* que nos permite incluirlo en la taxonomía. La pregunta se puede hacer en la misma forma en que se la hace la teoría del campo léxico: ¿cómo sabemos que *a* pertenece o puede pertenecer a una taxonomía particular (la de los colores) y no a otra?

Según Harrison el problema se resuelve si *a* mismo se define "in terms of the taxonomic schema for assigning names to colors itself: or rather, in terms of the peculiar character of the structure of rules which the schema comprises. We shall say that anything is a color which can be assigned a name by the operation of a taxonomic schema of that type." (p. 70) Esto quiere decir que no es necesario saber previamente que hay algo *a* que se puede introducir en la taxonomía de los colores, sino que es la propia taxonomía la que decide si *a* se puede clasificar como un color. Aparentemente esto no constituye una respuesta, pues no se ve cómo se selecciona *a* ni cómo se selecciona la taxonomía necesaria. Para ello es imprescindible tomar como punto de partida la posibilidad que tienen los humanos de percibir y de establecer relaciones comparativas dentro de lo que perciben. Es decir, se presupone que parte de nuestra capacidad

como humanos es la percepción del color y que, a partir de comparaciones como las que se hacen en el ejemplo, se puede deslindar un campo en que hay características similares. Hacerlo, según Harrison, implica un *modo de atención perceptual* que simplemente ofrece el punto de partida para que se genere una taxonomía. Que puede uno equivocarse y atender a otras características que se ofrecen es algo que se comprueba en el aprendizaje de la lengua. Un niño puede, a partir de los elementos que se ofrecen, llegar a la conclusión falsa de que *azul* pertenece a la taxonomía en que se incluye el aire, en vez del color; no obstante, la sociedad dirige la aplicación de taxonomías durante la enseñanza. Pero dirigirla y establecer modos de atención perceptual no es lo mismo que asociar objetos preconocidos y signos biunívocamente. Entre objetos y signos hay una adecuación que se obtiene en el interior de la comunidad lingüística. No hay una relación convenida entre ellas, sino que la acción de la arbitrariedad y la función semiótica permiten aplicarse a lo conocible de muchas maneras diferentes.[47]

En estos términos el problema de la delimitación del campo semántico queda resuelto: por un lado se trata de una microestructura que se crea y no de la realización de una parte de una estructura total; es decir, no hay una macroestructura léxica que obligue a establecer fronteras para un campo determinado. Por el otro, no hay

[47] Soy yo quien introduce en la propuesta de Harrison las nociones de arbitrariedad y función semiótica. El objetivo de Harrison es atacar las concepciones empiricistas que proponen *categoremata* mostrados ostensivamente como orígenes del desarrollo del lenguaje; especialmente se opone a las ideas de B. Russell y W.v.O. Quine. Yo estoy de acuerdo con él.

vocablos o significados que *se incluyan* en un campo, sino que es el campo mismo —el esquema taxonómico— el que define los elementos que lo componen.

Históricamente los campos semánticos se fijan y, nuevamente, la significación deja el paso a la codificación; pero esa codificación no es absoluta ni rígida: se recrea en cada discurso pivoteando sobre la arbitrariedad y varía permanentemente, aunque en una medida bastante lenta. Desde el punto de vista metodológico, por lo tanto, habrá que partir de los discursos concretos, que implican una función semiótica orientada por la significación. En esos discursos el material que se analiza es fragmentario y heterogéneo, pero la respuesta, como se sostuvo en la segunda parte de este trabajo, no está en la atomización de estructuras, sino en una aplicación adecuada del concepto de esquema taxonómico: si puede elaborarse un modelo de este tipo en que las variantes encontradas en los documentos se integren en una generación de nuevos elementos de la taxonomía (y no necesariamente al final del esquema sino enmedio o antes, pues la ventaja de esta concepción es su enorme flexibilidad) y a la vez, puesto que la verbalización del análisis mismo produce sentido, se espera que un nuevo lector —un nuevo hablante— lo interprete con la ayuda de una nueva taxonomía, no se habrá perdido rigor científico aunque sí (y de eso se trata) la pretensión estructuralista de la totalidad.

Hay otras ventajas en la noción del esquema taxonómico: en primer lugar, como se habrá ya visto, el esquema no es solamente un modelo de los signos desligados de sus referentes, sino un modelo de los signos *con* los referentes que se proyectan en el sentido; así, no es cuestión de

separar para la lingüística y para la lexicografía dos esferas inconciliables: la de las cosas y la de los signos, sino que el paso de una a otra es constante y los objetos de otras ciencias no se separan sino que solamente constituyen otros escalones de la formulación de un esquema taxonómico. Así por ejemplo, como dice Harrison, la taxonomía de los colores no se tiene que dividir en una de la lengua ordinaria y otra, distinta, del cromatógrafo, sino que la de este último es la misma que la del hombre común y corriente, sólo que con mayor número de características agregadas por un conocimiento más profundo de los colores. Entre los vocablos de uso común y la terminología científica y técnica hay, en consecuencia, una diferencia sólo de grado y no cualitativa. Este punto, que tradicionalmente ha puesto en dificultades a la lexicografía, se puede resolver ahora armónicamente.

En segundo lugar tampoco hay dificultad para concebir la polisemia y las figuras de sentido desde la lexicografía y la lingüística. Si la relación de biunivocidad entre signo y objeto no existe y, en cambio, la arbitrariedad permite los cambios de significado, la distinción lógica entre denotación y connotación no opera en una teoría de este tipo; siempre se ha definido la connotación en forma dependiente del "sentido recto" de la denotación. Por lo tanto la noción de *denotación* implica aceptar la concepción que tiene la lógica de lo que debe ser un lenguaje; el fantasma de la nomenclatura convencional reaparece. Pero si los esquemas taxonómicos se pueden generar a partir de muy diferentes modos de atención perceptual es posible entonces que todo lo que se considera metáfora o

connotación no sean sino estructuraciones taxonómicas iniciadas en puntos diferentes y por discursos concretos distintos. En el ejemplo de los colores se decía que un niño podría "equivocarse" y concluir que *azul* significa algo como el aire. No es difícil imaginar que ya no se tratara de un niño que no conoce la historia de su lengua, sino de un poeta que metaforizara el vocablo *azul* como precisamente algo que significa el aire. La medida de la metaforización de un signo no sería entonces inmanente a la estructura (como lo propone la pareja denotación/connotación), sino que dependería de la tradición y la historia de una comunidad lingüística. De la poesía española de los siglos de oro hay seguramente muchos signos que hoy pasan como corrientes pero que en su época se reconocían como metáforas —lo que generalmente se designa como 'lexicalización" de una metáfora. Con esto no se trata de disolver el concepto de metáfora sino de determinar el campo en que puede significar algo. Puesto que la posibilidad de hacer metáforas y otras figuras de sentido no es ajena o anómala para la lengua natural, según lo propone este trabajo, su explicación no corresponde a las estructuras lingüísticas sino al momento histórico y cultural de una comunidad.

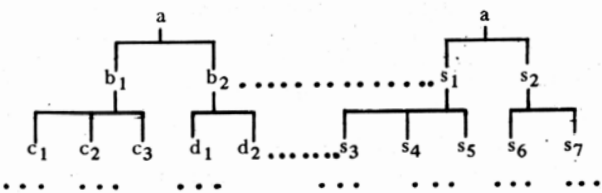
3.5.6.1. Semas en el esquema taxonómico.

Los esquemas taxonómicos eliminan también el problema del carácter mínimo de los rasgos significativos. Digamos que cada nodo de la arborescencia taxonómica puede considerarse como un rasgo significativo o sema. El carácter generativo del esquema permite en todo momento que los nodos terminales puedan seguirse analizando según las necesidades de significación

de un hablante. Los nodos, en consecuencia, pueden considerarse como simples si son terminales, pero complejos si es necesario continuar el esquema hacia un análisis más profundo. Pero la ventaja de esta concepción radica en el hecho de que es la misma función semiótica la que delimita el grado de pertinencia de un nodo, pues siempre necesariamente establece las fronteras hasta donde llega el análisis.

Cuando se codifica el esquema taxonómico en un momento histórico determinado y dentro de una tradición lingüística (que es lo que hacen los diccionarios) la dinamicidad del esquema se traduce en una clasificación de sus nodos y de sus reglas de generación. Desde el punto de vista metodológico este paso es peligroso, pues puede uno caer en una codificación parcial o en una rigidización de las reglas de producción. Con esta salvedad, el esquema se puede traducir al lenguaje de los conjuntos.

Cada capa del esquema taxonómico en el mismo nivel de análisis puede representarse como un conjunto de nodos y cada nodo se puede considerar un sema.



Bajo ese punto de vista, la traducción del esquema taxonómico (que pertenece a la significación) en una estructura concreta en un tiempo y una cultura dada (que pertenece a la codifica-

ción), origina *sememas:* sea uno que comprenda todo el esquema, o varios que correspondan a las distintas combinaciones de caminos en la generación del esquema; es decir, el semema tiene un carácter de matriz. Cuando hay un discurso determinado y la función semiótica orienta la generación del esquema, la traducción a una codificación produce *significados.* El significado es un arreglo jerarquizado de semas.

Se pueden representar ambos conceptos como sigue:

$$\text{Semema}_i = \begin{pmatrix} s_1 & s_2 & s_3 & s_4 & s_5 \\ s_6 & s_7 & s_8 & s_9 & s_{10} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ s_{n-4} & s_{n-3} & s_{n-2} & s_{n-1} & s_n \end{pmatrix}$$

$$\text{Significado}_x = \{ s_1 \ , \ s_3 \ , \ s_8 \ , s_k \ , \ldots , s_n \}$$

3.5.6.2. La práctica lexicográfica.

Antes de pasar a mostrar con un ejemplo concreto el funcionamiento del método que se ha obtenido de las primicias de teoría que se han venido elaborando, conviene recapitular una vez más los problemas que se han querido resolver:

a) Se niega la separación estructuralista entre lengua y objetos a través de la posición del *sentido* como el plano en que se proyecta la realidad con la actuación de la función semiótica. Los objetos lo son en cuanto inteligibles en el sentido.

b) Por lo tanto se elimina la dicotomía entre designación o denotación y significación, para pasar a considerarlas dos modalidades de la significación; la denotación es el modo selecciona-

do por el conocimiento científico para poder someter sus enunciados a la demostración lógica.

c) en forma concomitante con lo anterior, se elimina la dicotomía entre léxico y terminología, cuando la noción de *esquema taxonómico* permite concebir ambos procesos de significación bajo el mismo punto de vista y se los hace diferir en grado solamente.

d) La tematización de la metalengua lingüística permite proponer un campo de tensión entre la lengua objeto y el discurso referente a ella; el cruce constante entre ambas —especialmente en lexicografía— introduce los problemas de la valorización y la interpretación en lingüística. El método lexicográfico, por lo tanto, se tiene que convertir en un método crítico.

e) Se ha propuesto que los sistemas lingüísticos no son entidades fuera del tiempo y de la sociedad, sino que obedecen a codificaciones informativas para las que la biología humana ha determinado su posición. Los sistemas, por lo tanto, dependen de la significación y, cuando se los analiza, se les liga a una historia y una cultura determinadas.

f) Bajo estas condiciones las necesidades estructuralistas de contar con *lenguas funcionales* para poder hacer el análisis no tienen ninguna importancia, en la medida en que la creación de un esquema taxonómico tome en cuenta las reglas de producción necesarias para que, fenómenos diferentes, muestren su diferencia en el interior del esquema.

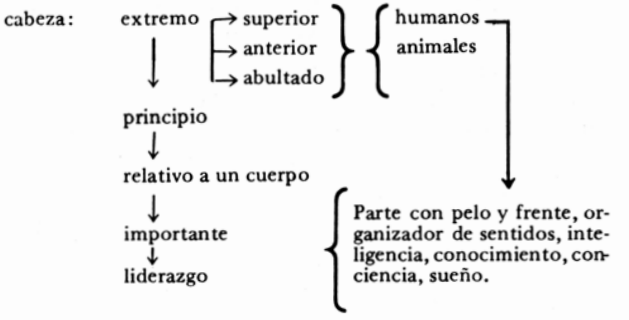
3.5.6.3. Un ejemplo de la práctica lexicográfica.

Del Corpus del Español Mexicano Contemporáneo (CEMC), he tomado un vocablo que pueda servir como ejemplo del tipo de análisis que propongo.

Sobre aproximadamente dos millones de ocurrencias de palabras, el vocablo *cabeza* tiene 170 ocurrencias. De ellas, 118 se refieren a la *cabeza* como parte del cuerpo humano y como señal del estado de ánimo de una persona (p. ej.: "Con la cabeza gacha, arrastrando sus pies, ridículo como un títere..." de F. Rojas González, *El diosero,* p. 131); 26 a la cabeza como el lugar donde está el pensamiento, la inteligencia, la invención, la sabiduría y el sueño; 13 a las cabezas de los animales y otros seres vivos (p. ej. el espermatozoide); seis a la medida de la unidad (p; ej: cuatro cabezas de ganado); cuatro al liderazgo (p. 3j: la cabeza de un movimiento sindical), dos a cierto tipo de extremos de objetos y una a un extremo superior de algo. El vocablo *cabeza* es ampliamente usado en cualquier nivel de lengua y en todo México.

El CEMC es un conjunto heterogéneo de muestras del español mexicano; difícilmente se podría someter a una preparación que homogeneizara niveles de lengua.

De un análisis en semas verbalizados se obtiene un cuadro esquemático como el siguiente:



Evidentemente los semas obtenidos ni son mínimos, ni son resultado de un análisis para el que se haya contado con todo el campo semántico de los 'extremos' o de la 'inteligencia', sino que provienen del análisis de las concordancias que ofrece el CEMC *con* la ayuda del conocimiento del lingüista. La primera verbalización, por lo tanto, se puede someter a análisis sucesivos que ofrezcan más información sobre los signos que se están usando como metalengua.

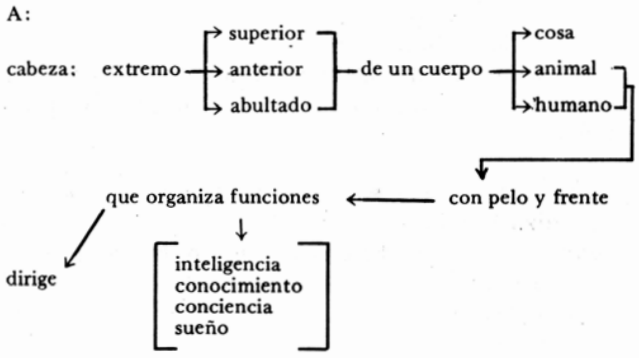
Las relaciones marcadas con flechas son orientaciones que ofrece el corpus y el conocimiento del analista para hacer inteligible la amplia polisemia de *cabeza*. Por lo tanto no son hechos formales ni verificables con un procedimiento específico, aunque sí con la repetición del análisis por otro investigador. En el caso presente el primer análisis fue hecho por Carmen Boullosa, del equipo de redacción del DEM y el segundo, sin conocer al anterior, fue hecho por mí.

Aunque el análisis no sea formalizable, no deja de llamar la atención que sea muy aproximado al que haría un hablante común y corriente, y al que hizo Greimas con el vocablo francés *tête* (Op. cit., pp. 42-50).

El problema es someter el modelo de esquema taxonómico a una prueba práctica.

Hay varios modos posibles de atención perceptual (Cf. 3.5.6). Cada uno haría que se generara un esquema taxonómico diferente, según las necesidades de significación de un hablante. El cuadro anterior es, en consecuencia (pues los analistas son también hablantes y se enfrentan a los contextos en estudio con una necesidad específica de significación), un esquema taxonómico más.

Otros esquemas taxonómicos podrían ser:



Bajo este modo A de atención perceptual se podrían comprender las siguientes clasificaciones de signos:

Extremo • superior • de un cuerpo• cosa:
cabeza 'parte más alta de los arrecifes, bajíos, escollos'
'extremidad del mástil de un instrumento de cuerda'
'encabezado de un periódico'
'parte superior del corte de un libro'

Extremo • anterior • de un cuerpo • cosa:
cabeza 'parte anterior de un convoy de trenes'
'parte delantera del carro de tren que se engancha a la locomotora'
cabeza de puente
cabeza de playa avanzadas de un ejército en el terreno.

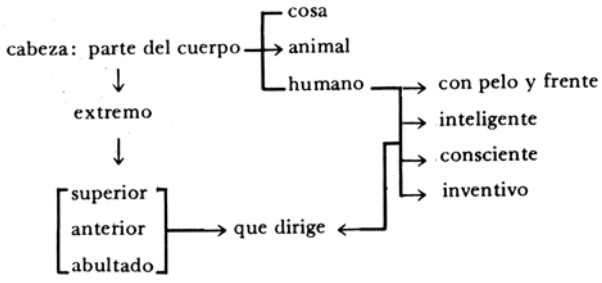
Extremo • abultado • de un cuerpo • cosa:
cabeza de alfiler, de clavo, de tornillo

de un broche, de un pasador
de penca
de silla (charra)
de émbolo

Para no explotar este esquema taxonómico exhaustivamente, veamos unos cuanto ejemplos más:

Extremo • (?) • de un cuerpo • cosa • que organiza funciones:
cabeza articular 'Extremo de un hueso por el cual se articula con otro'
de barrena
de émbolo
de cilindro
de grabadora
de computadora

Un esquema como el anterior puede representar la manera como la designación técnica de los objetos procede a partir de signos previamente codificados por la tradición lingüística de la comunidad. Pero en cambio aparenta ser un procedimiento bastante inverosímil para la designación de la 'cabeza' de los animales y los humanos. Por eso es posible un esquema alternativo B:



Así se podrían explicar:

Parte del cuerpo • humano • inteligente:

cabeza tener cabeza 'ser inteligente'
de cabeza 'de memoria'
venírsele algo a uno a la cabeza 'idear, recordar'
de su cabeza 'de su invención'
quebrarse la cabeza 'pensar un problema difícil'

Parte del cuerpo • humano • consciente:
subírsele algo a uno a la cabeza 'engreirse'
calentarle a alguien la cabeza 'hacerlo pensar algo'
sentar cabeza 'ser responsable'
dar con la cabeza en las paredes 'desesperarse'
andar de cabeza 'estar fuera de sí'
perder uno la cabeza 'enloquecerse, enojarse'
tener la cabeza en los pies

Se ha dicho que no se puede sostener la existencia de campos semánticos desligados de la significación y de la codificación que se hace en la tradición y en la historia. Por eso mismo estos esquemas no son estructuras del mismo tipo que las criticadas en la segunda parte de este trabajo, sino modelos interpretativos del significado de los signos, expuestos a la crítica desde dos ángulos: el metalingüístico, en que se califica el valor de las verbalizaciones usadas, y el lingüístico, en que el hablante juzga su capacidad de explicación.

Un diccionario es un discurso codificador, en el sentido de que solamente registra lo que *usualmente* se dice. Esta codificación corresponde a una tradición histórica de la comunidad a la que pertenece y se expresa —en forma limitada y cuestionable— en un *uso* cuantitativo, pero que encierra los valores pertinentes a la

sociedad en una época histórica dada. La aplicación de un modelo taxonómico al análisis semántico está ligada, por lo tanto, a la orientación del diccionario y a la interpretación de los documentos en los que se basa. Lo que no puede hacer un diccionario es hacer pasar por total y única una estructura interpretativa. En cambio sí puede ofrecer, en la sucesión de acepciones y agrupaciones de significados, una analogía con los esquemas taxonómicos pensables dentro de la sociedad, que permita a sus lectores tanto reconocer elementos del esquema, como generar, por su cuenta, otros nuevos.

El modelo de esquema taxonómico aquí propuesto necesita todavía someterse a mejores pruebas y comprobarse con otro tipo de datos, tanto de las teorías de la percepción, como de aquellos instrumentos de la lingüística contemporánea que, al matematizar, abren la posibilidad de efectuar verificaciones sobre amplios corpus. En este momento se me ofrece como una solución teóricamente sostenible y prácticamente aceptable, a esa parte de la semántica que trata con las palabras.

Investigaciones lingüísticas en lexicografía se terminó de imprimir en el mes de enero de 1980 en los talleres de Offset Setenta, S. A., Víctor Hugo 99, México 13, D. F. Se tiraron 2 000 ejemplares más sobrantes para reposición. Cuidó de la Edición el Departamento de Publicaciones de El Colegio de México.

En el panorama mundial de la lexicografía se ha escrito poco y se conoce menos acerca de las teorías y los métodos que dan forma a los diccionarios. Ha sido como si los lexicógrafos, entretenidos cotidianamente con los múltiples detalles que se entrecruzan en las palabras, se hubieran olvidado de exponer la estructura de sus andamios.

A la vez la lingüística moderna, en su cientificidad, se ha ocupado poco –y muchas veces a sabiendas– de lo que pasa con los datos y las concepciones que se reflejan en la *palabra*, ese extraño concepto que tienen los hablantes y que desafía persistentemente a las teorías del lenguaje.

Esta colección de trabajos pertenece a esa doble vertiente de la lexicografía y la lingüística. Explica una parte fundamental de lo que arma al *Diccionario del español de México*, una investigación que, con la ayuda del gobierno mexicano, se viene realizando en El Colegio de México desde hace seis años.

Como lo señala el título del libro, cada artículo desarrolla una investigación metódica para la lexicografía, pero al mismo tiempo integra una propuesta teórica para la lingüística contemporánea; especialmente para aquella que, superado el estructuralismo, busca asumir sus avances formales bajo la lente de una crítica constante, que resulta de las necesidades prácticas de la lexicografía.

Si bien el libro se dirige a lectores especializados, no dejará por eso de aportar algunas reflexiones interesantes a los que se preocupan por el lenguaje como fenómeno esencialmente humano.

EL COLEGIO DE MÉXICO