LIAMES 1 - pp. 11-18, Primavera 2001

Jon Landaburu*
(C.C.E.L.A./ CNRS)

Últimos Desarrollos de la Lingüística Amerindia en Colombia: El Programa de la Base de Datos del Centro Colombiano de Estudios de Lenguas Aborígenes

I. LA DIVERSIDAD LINGÜÍSTICA COLOMBIANA

Actualmente se habla en Colombia :

• **la lengua castellana**, llegada de Europa en el siglo XVI con la conquista española. Es la lengua oficial del Estado colombiano y es utilizada por la casi totalidad de la población nacional (casi cuarenta millones de personas). Entre los hispanoamericanos, los colombianos hacen frecuentemente alarde de hablar un castellano muy "puro", es decir más cercano al castellano clásico. Aunque este afirmación no deje de ser discutible y remita a intereses ideológicos particulares, es cierto que existe en Bogotá, la capital del país, una fuerte tradición de purismo castellano y de estudios gramáticos de calidad. En la práctica, el castellano real se manifiesta por un gran número de variantes regionales relativamente distantes entre sí, la mayor fractura siendo la que existe entre las variedades andinas y las variedades caribes. Como en otras partes, la expansión del mercado y de los medios de comunicación masivos está reduciendo la diferenciación de las hablas hispánicas locales. Esta misma expansión implica también la penetración cada vez más fuerte de la lengua española en las zonas de refugio donde hasta ahora las culturas y las lenguas indígenas y afroamericanas habían podido mantenerse y sobrevivir.

• **Dos lenguas criollas**, habladas en el Caribe por poblaciones de origen negroafricano: el criollo de San Basilio de Palenque cerca de Cartagena de Indias hablado por unas 3.000 personas; el criollo de las islas de San Andrés y Providencia (Old Providence) frente a las costas de Nicaragua hablado por unas 30.000 personas. Estas dos lenguas son recientes.

* Director científico del Centro Colombiano de Estudios de :Lenguas Aborígenes (C.C.E.L.A.) de la Universidad de los Andes de Bogotá (Colombia); Investigador del Centre dËtudes des langues Indiennes d'Amérique (C.E.L.I.A.) del Centre National de la Recherche Scientifique de Francia (C.N.R..S.). Para toda información, contactar el e-mail: ccela@unidades.edu.co

Fueron creadas por esclavos de origen etnolingüística africana diversa (más netamente bantú en el caso del criollo de Palenque), en la época de la trata y de la esclavitud impuesta por los europeos en los siglos de la colonia. El criollo de San Basilio o "palenquero" nace en un contexto hispánico y el mayor número de sus raíces léxicas proviene del castellano constituyéndose así aparentemente en el único criollo de base hispánica del continente americano. El criollo de San Andrés y Providencia nace en un contexto de habla inglesa (migraciones desde la Jamáica) y su fondo léxico es mayoritariamente inglés.

• **Sesenta y cinco lenguas indígenas amerindias** extremadamente diversas, habladas por unas quinientas mil personas en 22 de los 32 departamentos del territorio colombiano. Tanto estas lenguas como las dos criollas que acabamos de mencionar fueron reconocidas por la Constitución política de Colombia de 1991 como oficiales con el español en los territorios en las que se hablan.

II. LAS LENGUAS INDÍGENAS DE COLOMBIA

Se puede en este momento reagrupar las sesenta y cinco lenguas indoamericanas presentes en el territorio colombiano en trece familias lingüísticas diferentes, a las cuales hay que añadir ocho lenguas aisladas no reagrupadas en este momento con otras, lo cual nos da veintiún estirpes diferentes. Algunas de estas estirpes tienen una presencia continental importante como las grandes familias **Arawak** (8 lenguas en Colombia), **Caribe** (2 lenguas), **Tupi-Guaraní** (2 lenguas), **Quechua** (2 lenguas) o la gran familia **Chibcha** (7 lenguas) de probable procedencia centroamericana ; otras son de ámbito más regional como las familias **Chocó** (2 lenguas), **Guahibo** (4 lenguas), **Sáliba** (2 lenguas), **Macú** (3 lenguas), **Huitoto** (3 lenguas), **Bora** (2 lenguas),**Tucano** (18 lenguas) y **Barbacoa** (2 lenguas), solamente presentes en el noroeste de Suramérica, en Colombia y en sus vecinos. Las ocho estirpes de lenguas únicas son las siguientes : **andoque**, **cofan**, **kamëntsá**, **páez**, **tinigua**, **yagua**, **yaruro**, **ticuna** (esta última también presente en Brasil).

Más allá del número muy grande de estirpes, la diversidad tipológica de estas lenguas es notable ya que se encuentran entre ellas lenguas tonales, acentuales o tono-acentuales; lenguas que utilizan la harmonía nasal o la harmonía métrica; lenguas cuya estructura sintáctica dominante es holofrástica y por lo tanto sin distinción sujeto-predicado en el sentido clásico, frente a lenguas con una estructura claramente dominada por un sujeto o frente a lenguas de estructura predicativa existencial; lenguas de sistema argumental ergativo/absolutivo frente a lenguas de sistema acusativo/nominativo o de sistema activo/inactivo; lenguas de morfología aglutinante frente a lenguas muy flexionantes en el sentido clásico o frente a lenguas fuertemente aíslantes, etc.

El grado de divergencia entre lenguas de la misma familia es muy variable y va desde un compartir considerable de cognados (algunas lenguas de la familia Guahibo), hasta una diferenciación considerable que remite probablemente a desarrollos separados desde varios milenios (familia Chibcha). La dialectalización interna a una

misma lengua es también muy variable y se manifiesta bajo configuraciones topológicas diversas (continuos de comunicación con diferenciación cumulativa, áreas de núcleos homogéneos con zonas de transición, áreas homogéneas discontinuas de evolución divergente reciente, etc.).

Esta diversidad lingüística debe ser referida a una diversidad cultural muy grande que refuerza el interés de tal universo. Aquí no tenemos solamente gente que dice cosas de manera diferente sino ante todo gente que dice cosas muy distintas, lo que se refleja evidentemente en el léxico pero también en la gramática. Entre estas poblaciones indígenas demográficamente reducidas (solamente tres grupos etnolingüísticos tienen más de 50.000 personas), encontramos efectivamente cazadores recolectores de bosque, pastores nómadas de semidesiertos, horticultores de tumba y quema, pescadores costeros, agricultores andinos transhumantes, agricultores estables, etc. Por fin las circuntancias de multilingüismo y las prácticas de habla son también extremadamente variadas y ameritan la observación del lingüista. Según los casos se ven situaciones de multilingüismo entre lenguas indígenas o entre lenguas indígenas y el español u otras lenguas vehiculares, prácticas de exogamia lingüística, lenguas rituales, etc...

III. EL ESTUDIO DE LAS LENGUAS INDÍGENAS DE COLOMBIA

Hasta estos últimos años, este amplísimo campo de las lenguas de Colombia como, en general, el universo de las lenguas indias de Suramérica, era muy poco conocido. Algunos trabajos de misioneros durante la época colonial y las publicaciones de algunos viajeros europeos durante el siglo XIX habían dejado vislumbrar la dimensión sorprendente de esta realidad. Durante el siglo XX el interés creció y los trabajos con finalidad científica fueron aumentando. A partir de los años sesenta, los misioneros norteamericanos del Instituto Lingüístico de Verano (Summer Institute of Linguistics), instalados en Colombia desde 1962, empiezan a publicar además de sus cartillas y traducciones religiosas, algunos trabajos más cercanos a los cánones científicos modernos. Al principio de los años ochenta arranca en la Universidad de los Andes de Bogotá, con la ayuda del Centro nacional de investigaciones de Francia (C.N.R.S), una operación de formación de lingüistas colombianos, formación a la investigación por la investigación. El programa de formación de postgraduados desemboca en la creación de un centro de investigaciones, el C.C.E.L.A. (Centro Colombiano de Estudios de Lenguas Aborígenes). Para nuestras fechas de hoy y después de quince años, el C.C.E.L.A. ha logrado articular una red de unos cincuenta profesionales - de los cuales doce indígenas - que trabajan en la actualidad cuarenta lenguas amerindias. En lo que sigue vamos a presentar uno de los proyectos de esta red de investigadores. Consiste en el montaje de una base de datos computarizada multimedia que puede ser de interés para los investigadores brasileros dedicados a las lenguas indígenas de Brasil.

IV. EL PROYECTO DE LA BASE DE DATOS INFORMATIZADOS Y CARTOGRAFIADOS

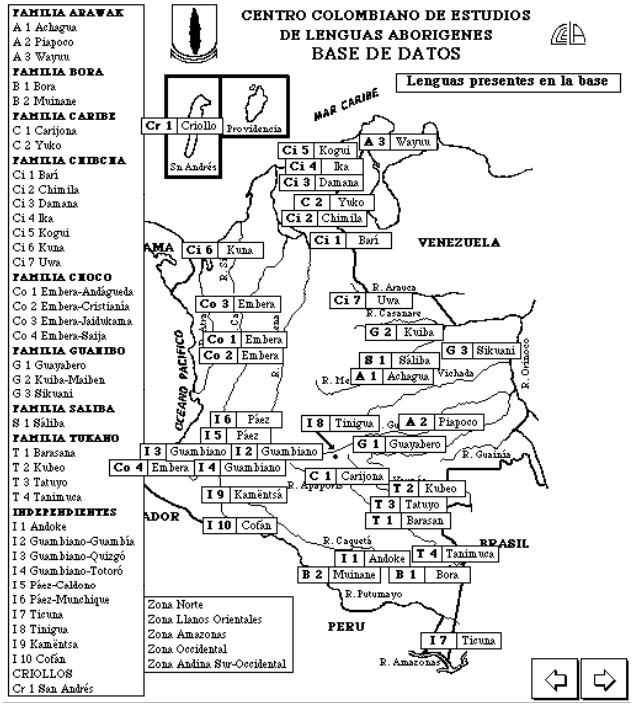
En 1991, después de siete años de formación de postgraduados, el nivel de competencias adquiridas por los investigadores y la precisión de los datos conseguidos por los estudiantes y los docentes, hicieron surgir la idea de crear una herramienta de consulta electrónica, capaz de poner a disposición de un círculo mayor de investigadores y de un público más amplio los conocimientos disponibles. Los objetivos finales de una tal herramienta serían de permitir el acceso rápido a los datos conseguidos y elaborados en el C.C.E.L.A. a sabiendas que este tiene la pretensión de cubrir a mediano plazo todas las lenguas indígenas y criollas de Colombia. Una de las grandes ventajas de este banco de datos estaría en que los recolectores de datos serían los mismos lingüistas, formados en la misma escuela y comprometidos en discutir en la institución todos los problemas epistemológicos y prácticos planteados por esta empresa. Los datos serían pues de primera mano, controlables científicamente por todo el grupo y relativamente homogéneos en cuanto a las opciones teóricas. Bien se sabe que buena parte de los datos de bases de este tipo como las informaciones sobre estructuras fonológicas, gramaticales, semánticas, etc., no son brutos sino mediados por una elaboración científica dependiente de elecciones teóricas.

En este diseño original se previó que los datos de la base serían multimedia : cartografiados, textuales y sonoros. La presentación de los datos sobre mapas fue considerada como primordial en el proyecto por lo que facilitaría el estudio comparativo, tipológico e histórico de los fenómenos. Se procedió entonces a escoger cuidadosamente puntos de encuesta en el territorio colombiano, en función de los investigadores disponibles y en función de una dispersión genética y tipológica óptima. En cada punto de encuesta se buscaría recoger una masa cuantitativamente importante de datos lingüísticos obtenidos gracias a las respuestas a cuestionarios fonético-fonológicos, gramaticales, léxicos y sociolingüísticos. El sistema de gestión de estos datos debería permitir representar las respuestas dadas en diferentes puntos a la misma pregunta en mapas regionales, de familias lingüísticas o areales. Ningun tratamiento o asociación “inteligente”por la maquina sería buscado en ese momento. Se buscaría en cambio optimizar la “navegación” dentro de la base de datos para permitir al investigador exterior consultante no solamente tener un acceso facil a los resultados sino también poner a prueba sus hipótesis.

V. EL ESTADO ACTUAL DE LA BASE DE DATOS

Ocho años después de su iniciación, la base sigue en construcción pero ya puede presentar resultados muy apreciables. Hasta la fecha, el programa ha logrado recopilar datos relativos a 32 lenguas diferentes, recogidos por 32 investigadores en 38 puntos de encuesta. Las lenguas que tienen cabida en la base son : de la familia Arawak, el wayuu o guajiro, el achagua y el piapoco; de la familia Chibcha, el uwa o tunebo, el kogui, el damana o wiwa, el arhuaco o ika, el chimila, el barí y el cuna ; de la familia Caribe, el carijona y el yuko ; de la familia Chocó, el embera-chamí, el embera-catío y el embera-saija; de la familia Guahibo, el cuiba, el sikuani y el guayabero; de la familia Sáliba, el sáliba; de la familia

Macú, el puinave; de la familia Bora, el bora y el muinane; de la familia Tucano, el cubeo, el tanimuca, el tatuyo y el barasana; de la familia Barbacoa, el guambiano y el totoró; de estirpes de lenguas únicas, las lenguas andoque, cofán, kamëntsá, páez, ticuna, tinigua. O sea que de las 21 estirpes de lenguas indias presentes en Colombia, 16 tienen en este momento representación en la base de datos. El criollo de San Andrés también está presente y esperamos la integración de los datos de nuestros investigadores del criollo de san Basilio de Palenque. A continuación, mostramos el mapa que indica la localización de estas lenguas en el territorio colombiano :



A) El componente fonético-fonológico de la base

Su montaje está terminado, lo cual significa que el cuestionario ya fue concebido, puesto a prueba y aplicado; los datos recolectados en trabajos de campo, digitalizados y revisados; el sistema de gestión y los mecanismos de consulta automática, implementados. Este dispositivo permitirá ahora integrar datos relativos a tantas lenguas o variedades como se quiera. La elaboración del cuestionario se había completado desde julio de 1992. Su redacción y su puesta a prueba habían necesitado varias reuniones colectivas durante las cuales los investigadores venían a explicar las características o los problemas particulares planteados por su lengua con vista a su integración en el cuestionario definitivo. La especificidad de las estructuras de estas lenguas o su poca consideración en la elaboración de cuestionarios existentes por cuenta de linguistas generalistas nos obligaba a esta adaptación. La orientación teórica del cuestionario fonético-fonológica sigue los estandares tradicionales (estructuralismo pragués y/o distribucionalismo americano). El

cuestionario comprende 1809 preguntas-fichas distribuídas en 18 secciones : fonos, fonemas, reglas de realización, procesos fonológicos, neutralizaciones, frecuencias y rendimientos, etc. De los 39 programados, 27 cuestionarios están ya asequibles por el sistema de gestion computarizado. El sistema de notación es el del Alfabeto Fonético Internacional (AFI). Las 1809 fichas-preguntas se proyectan sobre el fondo de mapas generales, regionales y por familias. El componente fonético-fonológico de la base tiene en la actualidad 7332 mapas.

La base permite contestar preguntas sencillas del tipo : ¿Dónde se dan consonantes prenasalizadas? o ¿Qué lenguas presentan un patrón de lexemas bisilábico? estas preguntas pudiendo ser efectuadas en un ámbito nacional, areal o por familias. Se pueden cruzar preguntas. Por ejemplo, si se pregunta : ¿Qué lenguas tonales tienen una estructura silábica CV y morfemas gramaticales de tamaño infrasilábico? se puede luego examinar los rasgos pertinentes del sistema consonántico de las lenguas que contestan afirmativamente. La proyección geográfica de las respuestas permite asociar a los interrogantes estructurales y tipológicos internos a una lengua, una dimensión comparativa, areal o genética. La recolección de datos sonoros para el léxico no ha sido aún iniciada. Se piensa que aportará no solamente un soporte documental importante para el léxico sino también ilustraciones sonoras computarizadas para los fenómenos fonético-fonológicos.

B) El componente léxico

En 1993 elaboramos un cuestionario léxico de 3360 entradas para el cual tenemos ya respuestas relativas a 38 puntos de encuesta. Contrariamente al área fonética, el área léxica no dispone de una categorización preestablecida, universalmente admitida. Se necesitan varias etapas antes de poder presentar listas léxicas relativamente cercanas al uso efectivo de los locutores, tanto a nivel de las palabras en tanto que unidades memorizadas como al de los lexemas. La primera etapa consistió en producir un cuestionario de caracter enciclopédico o ideológico, una especie de primera red destinada a recoger palabras y expresiones lexicalizadas, para la cual se trató que ningun sector importante de la experiencia humana faltase. Se estableció una larga lista estructurada en campos semánticos, inspirándonos del “Dictionary of Selected Synonyms in the Principal Indo-European Languages” de Carl Darling Buck, del “Notes and Queries on Anthropology” (versión española de 1966). del proyecto “International Dictionary Series Wordlist” de la Universidad de California en Irvine, del Thesaurus de Roget, de la Enciclopedia Británica, de los cuestionarios de la obra “Encuesta y descripción de lenguas de tradición oral” del laboratorio de “lenguas de tradición oral” (Lacito) del CNRS, etc. Como siempre, se rectificaron y completaron preguntas en función del conocimiento de los universos propios a cada grupo etnolingüístico aportado por los investigadores del CCELA.

Este cuestionario tiene tres grandes partes : 1- el conocimiento de la naturaleza (mundo físico, mundo vegetal, mundo animal, mundo sobrenatural); 2- el mundo humano (el cuerpo, la siquis, la sociedad); 3- la acción del hombre sobre la naturaleza (sobre el mundo físico (desplazamientos, transformaciones y acciones, instrumentos), sobre el mundo vegetal, sobre el mundo animal, sobre el mundo sobrenatural). Cada ítem de la lengua aparece en una ficha compleja que comprende las informaciones siguientes : numeración decimal de cinco dígitos que permite ubicar el ítem en las secciones y

subsecciones ideológicas del cuestionario; transcripción en AFI del ítem; informaciones codificadas relativas a las circunstancias de la recolección de la respuesta; nivel de complejidad sintáctica de la respuesta; clase sintáctica de la respuesta (por ejemplo nombre, verbo, calificativo, etc., para las palabras); segmentación de la respuesta en unidades mínimas con la glosa yuxtalinear; identificación más precisa del referente cuando es el caso (e.g. especies vegetales o animales); designación del ítem en castellano regional; informaciones sobre el registro de uso (vulgar, obsceno, infantil, sagrado, etc.).

Un primer balance de la recolección de datos nos ha mostrado, lo que desde luego era previsible, que un número importante de las preguntas solicitadas (más de la tercera parte, pero no siempre las mismas) quedaron sin respuesta pero que también campos semánticos enteros revelaron una complejidad subtratada en el cuestionario. Es claro que no se puede sino partir de una categorización occidental de la experiencia (punto de vista "etic") y que se trata de acercarse progresivamente a una categorización indígena (punto de vista "emic") sin perder de vista una tabla comparativa suficientemente poderosa. La necesidad de contrastar datos nos impone desde luego una referencia numérica común ordenada. Hemos organizado sesiones de trabajo (sobre la anatomía humana) para revisar la lista primera por subsecciones tratando de adaptarla a los universos encontrados.

El ejercicio es dificil. No pretende substituirse a la recolección minuciosa hecha por el especialista de una lengua que acopia y clasifica todas las palabras de un gran corpus de textos y de audiciones asociándoles todas las informaciones pertinentes. Dicho de otra manera no se trata de la elaboración de un diccionario aunque sí puede ayudar a esta tarea. Lo que puede entre otras ventajas aportar es un listado sistemático de los lexemas de una lengua, asociado a una tabla de concordancias dentro de la misma base. Esta concordancia debería permitir la construcción más fina del significado de cada unidad, cada vez menos en función de la traducción y cada vez más en función de la combinatoria de signos. Es importante resaltar aquí que estamos en presencia de lenguas en mayoría aglutinante para las cuales la composición y la derivación son muy productivas y por lo tanto donde la combinatoria semántica es frecuentemente de gran ayuda para restituir el significado de las unidades complejas.

Otra ventaja de una base de datos léxicos como esta concierne la comparación entre lenguas y no solamente dentro de una lengua . En ella, la comparación genética puede más facilmente establecer entre morfemas y no entre palabras como es todavía demasiado frecuente en la lingüística indígena suramericana. La comparación areal puede también beneficiarse de estos datos al permitir manifestar patrones semánticos comunes a varias lenguas. Si por ejemplo descubro en una lengua que un mismo lexema es utilizado en expresiones que designan la mandioca, el canto, la duración, el hueso y que esta misma constelación está presente en otras lenguas de la región, he avanzado en el descubrimiento de campos semánticos por lo menos locales.

Mientras se avanza en estas exploraciones, la base puede ya rapidamente contestar a preguntas informativas sencillas del tipo : ¿ Cómo se dice "sal" o "tener hambre" en las lenguas de tal región o de tal familia ? Más allá de la estricta comparación lingüística, la proyección cartográfica de las respuestas ofrece también pistas interesantes para el etno-historiador en relación a préstamos y contactos.

C) El componente gramatical

A nivel gramatical, se hicieron ensayos de cuestionarios originales sobre estructuras predicativas, estructuras actanciales y tipos de frases. Aunque se consiguieron resultados interesantes en seminarios entre los lingüistas, no se fue mucho más allá de estos campos sintácticos. Las dificultades para organizar una tabla conceptual común son bien grandes y ameritan un esfuerzo más amplio. Dejando momentaneamente lo onomasiológico por lo semasiológico, solicitamos a cada investigación la recolección codificada de todos los morfemas gramaticales de su lengua. Cada unidad entra en una ficha muy detallada donde se precisan sus propiedades formales, distribucionales, funcionales y semánticas. La elaboración de la ficha dió también pie a intentos de estandarización de la morfología que hay que proseguir. Hasta el momento, se tienen 13 conjuntos de fichas de morfemas gramaticales relativos a las 13 lenguas siguientes : embera-chamí y embera-saija (fam. Chocó), uwa y wiwa (fam. Chibcha), tanimuca (fam. Tucano), guambiano (2) (fam. Barbacoa), sáliba, kamëntsa, andoke, ticuna, paez, cofan (lenguas aisladas).

D) Cuestionario sociolingüístico

Con el objetivo de medir la vitalidad actual de las lenguas indígenas y con la idea de aprovechar las posibilidades de los actuales 38 puntos de encuesta, se ha ido elaborando un cuestionario sobre actitudes frente a las lenguas, situaciones de contacto lingüístico, circunstancias de uso de la lengua vernácula según sexo, clases de edad, ámbitos. Una primera versión del cuestionario fue puesta a prueba entre comunidades de Caldono en el Cauca (hablantes de la lengua paez) en 1998.

E) La informática

Del lado informático, el ambiente es bastante sencillo y no hemos tenido necesidad hasta ahora de salir del nivel de los microcomputadores. La unidad central es un Macintosh 7200/90 asociado a dos discos duros de un Gygabyte y a un monitor de 16". Para el cuestionario fonético se utilizó el programa Hypercard que permite interfaces necesarios con el material cartográfico y el sonido. Para el cuestionario léxico y el de morfemas gramaticales se utilizaron versiones recientes de File Maker. Ya se realizó un CDROM del cuestionario fonético-fonológico. Esperamos superar unas dificultades institucionales momentáneas y poder colocar en un futuro cercano todos estos datos en un portal internet.

Estas son las principales características de un programa que merece ser continuado, perfeccionado, y, porque no, imitado.

Rio de Janeiro, noviembre 1999