Coracon: *corpus* de referencia del guaraní contemporáneo

José Maria Rodrigues Rodrigues

(Universidad Católica de Asunción)

Dora Bobadilla de Cazal

(Universidad Católica de Asunción)

Teresa de Jesús González de Benítez

(Universidad Católica de Asunción)

Introducción

Con el resurgimiento de las investigaciones basadas en *corpus*, o de corte empírico, que tuvo su auge en los años cincuenta de la mano de investigadores de la talla de Firth, Boas, Harris y Hill (LEECH, 1991, 1992) (CHURCH; MERCER, 1993), y que puede considerarse la lingüística prechomskyana o "Early corpus linguistics" (MCENERY; WILSON, 1996), se ha notado un creciente interés por parte de la comunidad científica en retomar el paradigma estructuralista y los métodos de investigación de análisis lingüístico típicos de la década de los cincuenta, debido en gran parte a los avances informáticos y a los recursos cada vez más sofisticados de almacenaje de datos en formato magnético, que permite a los lingüistas manipular cantidades masivas de información dispuestas en los *corpora* informatizados.

Con el advenimiento de la era informática, principalmente en las dos últimas décadas, algunas de las críticas realizadas por Chomsky hacia los métodos empíricos e inductivos han tenido que ser revisadas y reformuladas (SINCLAIR, 1991, *passim*), ya que el lingüista de nuestros tiempos puede contar no sólo con sus intuiciones lingüísticas o con las competencias interiorizadas de los hablantes nativos,

sino que se puede basar en el estudio pormenorizado del uso lingüístico. Hoy en día, a través del análisis de textos y ejemplos reales disponibles en los *corpora*, se puede demostrar con más precisión algunos aspectos lingüísticos que nos permiten extrapolar los límites de la introspección y los datos intuitivos propuestos por Chomsky para la generación de principios o teorías lingüísticas.

El carácter mentalista que condujo a algunos lingüistas estructuralistas norte-americanos — era post-Bloomfieldian — y a algunas figuras eminentes de la tradición lingüística británica, como J. R. Firth, a un segundo plano, ha tenido que dar paso al renacimiento de los estudios basados en *corpus*, que aprovechando el carro de la revolución tecnológica de los últimos 30 años, resurge con tal fuerza y con unos presupuestos teóricos tan bien asentados, que muchos lingüistas insisten en clasificarla como Lingüística de *Corpus*, confiriéndole el *status* de disciplina independiente (LAGER, 1995) (KENNEDY, 1998) (BIBER *et al*., 1998) (TOGNINI-BONELLI, 1996). Por otro lado, los hay que ven la lingüística del *corpus* como una base metodológica o como herramienta de apoyo que puede usarse en todas las ramas de la lingüística (LEECH, 1992, p. 105).

1. La moderna lingüística de *corpus* y los estudios lingüísticos

La revolución informática de las últimas décadas ha modificado profundamente la lingüística de *corpus*, y debido al uso creciente y sistemático de los ordenadores y de los programas informáticos, ha posibilitado el procesamiento de cantidades ingentes de texto; lo que ha dado paso a una investigación basada en la elaboración y el análisis de *corpus* de referencia del lenguaje de tamaños cada vez mayores. Los objetivos y los logros de los lingüistas se han ampliado considerablemente, y el ordenador y sus tecnologías se han convertido en herramientas o metodologías indispensables en cualquier estudio lingüístico

(MCENERY; WILSON, 2001, p. 114-115). Además, los avances tecnológicos y la creciente disponibilidad de los medios para procesar grandes cantidades de textos han resultado en el desarrollo de programas que permiten poner el lenguaje escrito bajo el microscopio y transforman al lingüista en científico informático del lenguaje. Hoy día mediante algunos programas se puede buscar palabras o concordancias, sacar datos estadísticos, etiquetar palabras, construcciones, frases, lemas etc., hacer listas y comparar textos, comprobar teorías o hipótesis (LEECH, 1991, p. 9) y mitigar las limitaciones de la intuición de la lengua y del lenguaje.

2. Proyecto *AVAKOTEPA*: justificación teórica

La motivación principal que llevó a nuestro grupo de investigación a acometer la tarea de crear una plataforma que albergará todos los recursos existentes sobre la lengua Guaraní, además de la recopilación de *corpus* textuales y orales en dicha lengua, fue la constatación de que, pese a que en Paraguay hay muchas personas y/o instituciones que se dedican al estudio de la lengua Guaraní, éstos mayoritariamente lo hacen de forma aislada y sin la coordinación de un "Centro de Referencia sobre el tema". En *internet*, por ejemplo, son escasas las páginas que abordan el tema de la enseñanza-aprendizaje del idioma Guaraní o que dispongan de recursos — *corpora* textuales, gramáticas, muestras de audio etc. — que permiten la realización de estudios empíricos sobre el legado cultural más valioso del Paraguay, su lengua. La firma del Acta de Asunción, el 2 de agosto de 1995, suscripta por los Ministros de Educación y Cultura de los países miembros del Mercosur, declaró al Guaraní como lengua oficial del Mercosur. La octava resolución del Acta propone:

> Declarar al Guaraní Lengua Histórica del Mercosur y revalorizar su legado cultural a través de la elaboración de un inventario de su patrimonio, la promoción de la investigación académica y la enseñanza de la Lengua, conjuntamente con el

> estudio y la preservación de las culturas de la región[1] (MERCOSUR, 1995, acta octava).

En este sentido, nuestro proyecto no sólo pretende convertirse en un marco de referencia, al revalorizar su legado cultural a través de la elaboración de un inventario de su patrimonio para que esté disponible en *internet*, sino que también contribuirá a la mejor comprensión de los estudios existentes a través de la promoción de la investigación académica y la enseñanza de la Lengua. Ahora bien, si pretendemos fomentar la investigación académica y la enseñanza de la lengua Guaraní con rigor científico, es imprescindible invertir en la recopilación de *corpora* lingüísticos — tanto orales como textuales —, ya que hoy en día, a la luz de la moderna lingüística de *corpus*, éstos se han convertido en la única fuente empírica fiable para el estudio de cualquier aspecto de una determinada lengua — ya sea con fines lingüísticos, lexicológicos o terminológicos.

3. Los elementos vehiculizadores de la Cultura guaranítica y el proceso de normalización del Guaraní paraguayo contemporáneo

El castellano y el Guaraní, que han estado en contacto por más de quinientos años y que han trasmitido valores y tradiciones que representan sendas culturas, han ido evolucionando a lo largo de los siglos y han convergido en un sistema lingüístico *sui generis* e interrelacionado que — debido a la permeabilidad cultural — representa y expresa los valores y la cultura del Paraguay a través de dos lenguas que conforman el conocido bilingüismo mestizo, que son, a saber, el “Guaraní paraguayo” y el “castellano paraguayo”. Es evidente que hoy en día hay demasiados puntos de contacto e interferencias de códigos para que no se tenga en cuenta a una de esas lenguas cuando se estudia algún fenómeno relacionado con la otra. Ahora bien, cuando se trata de los saberes y valores propios del universo de la comunidad guaraní

hablante — originaria de la cultura ancestral Tupí-Guaraní — parece ser que hay una zona limítrofe bastante más definida y, pese al trasvase lingüístico, se puede identificar con claridad hasta que punto la cultura nativa sobrevive en la lengua de los colonizadores y, más aún, de que forma la misteriosa supervivencia del Guaraní denota la existencia de un valioso legado cultural arraigado al código lingüístico, y que se remonta a tiempos inmemoriales. Por ello, es imprescindible llevar a cabo estudios que rescaten y preserven dicho legado, para que se resguarden los valores espirituales y culturales que subyacen bajo el guaraní paraguayo.

Por otro lado, concomitantemente a este proceso de rescate histórico-cultural, una de las tareas más urgentes es la sistematización y normalización del código lingüístico. Es decir, para que la lengua guaraní deje de ser motivo de apasionadas — y a veces irracionales — discusiones respecto de los aspectos normativos que otorgan a una determinada variedad hablada el estatus de estándar — clasificada por muchos como purista y artificial —, en contraposición a "sus desviaciones", que deslindan de la norma académica y, pese a que son el fiel retrato de una lengua viva, son tratadas como aberraciones prosaicas — tildada como lengua inculta del pueblo llano —, es imprescindible, además de la aprobación de la sonada "Ley de Lenguas", que propone la creación de una "Academia de la lengua guaraní", aunar esfuerzos para trabajar sobre una normalización consensuada de la lengua Guaraní hablada en Paraguay. En resumen, además de exigir la utilización del guaraní en todos los ámbitos, invirtiendo en el proceso de funcionalización de la lengua autóctona y llevándola a todos los estamentos de la sociedad — con o sin neologismos, préstamos lingüísticos etc. —, para que, por ejemplo, los medios masivos de comunicación, el mundo empresarial y el Gobierno "de facto" la utilice como otra lengua del Estado, es preciso ponernos de acuerdo y analizar en profundidad el amplísimo repertorio léxico — activo y pasivo — del guaraní contemporáneo.

4. CORACON: *corpus* de referencia del guaraní contemporáneo

En consonancia con lo expuesto anteriormente, y convencidos de que el guaraní todavía goza de buena salud en lo que concierne a sus tradiciones y costumbres, estamos llevando a cabo un trabajo de investigación que va dirigido a rescatar los valores culturales de la lengua guaraní, que son seña de identidad de una extensa comunidad idiomática, la de los paraguayos. Los objetivos del proyecto *AVAKOTEPA* es recopilar un *corpus* oral y textual de referencia del guaraní hablado en el Paraguay que sirva de base empírica para investigaciones de índole lingüística, y que den a conocer, con mayor amplitud, algunos rasgos y aspectos representativos de la cultura guaranítica; asimismo, dicho proyecto conformará un marco lingüístico, a través de su amplia y representativa base de datos, imprescindible en el proceso de normalización de la lengua Guaraní. Dicho *corpus* comenzó a elaborarse en abril de 2008 y se enmarca en el proyecto *AVAKOTEPA*, que nace con la idea de poner a disposición de la comunidad científica un nuevo recurso que pueda ser accesible a través de *internet*. La finalidad del CORACON es, por tanto, facilitar la obtención de datos para el estudio de aspectos morfológicos, sintácticos y léxicos de la lengua guaraní. El "*Corpus* de Referencia del Guaraní Contemporáneo" se compone de dos partes: el CORACON, *corpus* oral que recoge muestras reales de habla que ocurren en interacciones y conversaciones auténticas y son almacenadas en una base de datos textual acompañada de su respectivo archivo de audio, y el COTRACON, *corpus* textual que reúne un conjunto de textos lingüísticos y reales almacenados en formato electrónico. En ambos *corpora* se siguen las recomendaciones hechas por Eagles y, *a posteriori*, se pretende adoptar las normas de codificación definidas en el TEI P4. El CORACON abarcará un espacio cronológico que va desde 1950 hasta la actualidad. Las bases de datos que componen dicho *corpus* son abiertas, es decir, están

diseñadas para albergar los últimos 58 años del guaraní, de modo que vayan actualizando sus materiales con el paso del tiempo.

4.1. Representatividad y tipología del *CORPUS* DE REFERENCIA GUARANÍ

Uno de los caballos de batalla a la hora de recopilar un *corpus* son los criterios que deben guiar su diseño para que éste sea realmente representativo. Por ello, en la recopilación del CORACON nos hemos hecho tres preguntas básicas: ¿Qué variedades de uso de la lengua guaraní debemos incluir? ¿En qué proporción? ¿Cuál debe ser el tamaño del *corpus* para que, realmente, represente el patrimonio de la cultura guaranítica y recoja los distintos usos que los paraguayos hacen de su lengua? Por ello, para que este *corpus* pueda convertirse en el "*corpus* de referencia del guaraní contemporáneo", estamos siguiendo las máximas establecidas para la creación de *corpus* lingüísticos, es a saber, cantidad, calidad, simplicidad y documentación. Los textos lingüísticos que conforman el CORACON están siendo seleccionados según los criterios establecidos por las organizaciones y los estándares internacionales de referencia y con la suficiente amplitud para que puedan considerarse como representativos del uso lingüístico del guaraní actual. Respecto del tamaño, en un primer momento, contará con una amplitud mínima de 5 millones de palabras y recogerá desde conversaciones informales hasta presentaciones de índole más académica — no leídas. Principalmente a partir de grabaciones espontáneas recogidas en la calle, en el autobús, en los comercios etc., diálogos familiares espontáneos y/o dirigidos, grabaciones en el ámbito educativo (Charlas en Guaraní, clases y diálogos entre alumno/ alumno y profesor/ alumno). Ya en la segunda fase del proyecto, tanto las grabaciones para conformación del *corpus* oral como los documentos que compondrán la base de datos textual, abarcarán todos los ámbitos descritos a continuación.

5. COTRACON: tipología adaptada del CREA

El COTRACON se compone de cuatro grandes bloques de materiales: LIBROS, PRENSA, MISCELÁNEA y LENGUA, que conforman la parte escrita del *corpus*. El apartado LENGUA se compone de las transcripciones de la lengua hablada, y está directamente vinculado con el *corpus* ORAL y sus áreas temáticas. Los "Libros y Prensa" se dividirán en dos grandes bloques: ficción y no ficción. Cada uno a su vez se subdividirá en siete grandes hipercampos formados por las siguientes áreas temáticas: 1) ciencias y tecnología; 2) ciencias sociales, creencias y pensamiento; 3) política, economía, comercio y finanzas; 4) artes; 5) ocio y vida cotidiana; 6) salud; 7) ficción.

El apartado MISCELÁNEA se agrupa en dos bloques: impresa y no impresa. Cada bloque lo forman diferentes tipos textuales divididos por áreas temáticas, como vemos a continuación:

8.1 Impresa
8.1.1 Resúmenes, actas, congresos, cursos
8.1.2 Programas, espectáculos, universidades, cursos, afiches cine, teatro
8.1.3 Boletines
8.1.4 Propaganda, información turística, guías consumo, informaciones varias
8.1.5 Instrucciones, juegos, electrodomésticos
8.1.6 Prospectos, medicinas
8.1.7 Invitaciones, convocatorias
8.1.8 Impresos oficiales y no oficiales
8.1.9 Folletos varios, religión, política, filosofía, ética

8.2 No Impresa
8.2.1 Notificaciones oficiales: multas, correos, hacienda
8.2.2 Cartas: personales, empresas, propaganda
8.2.3 Circulares internas
8.2.4 Mensajes de correo electrónico
8.2.5 Exámenes
8.2.6 Páginas *web*
8.2.7 Propaganda
8.2.8 Otros

Cabe señalar que el apartado LENGUA constituye uno de los bloques más importantes del citado "*Corpus* de referencia". En él se recogen las transcripciones de las muestras de lengua hablada del *corpus* ORAL, y cubre las siguientes áreas temáticas: 1) radiofónico o televisivo; 2) noticias; 3) reportajes; 4) entrevistas; 5) debates; 6) tertulias; 7) documentales; 8) retransmisiones deportivas; 9) magacines; 10) revistas deportivas; 11) variedades; 12) sorteos, concursos.

6. CORACON: criterios para la recogida de muestras orales

El en proceso de recopilación de los registros orales se están siguiendo los siguientes criterios a la hora de seleccionar las muestras de habla que integrarán la base de datos de audio: 1) oralidad; 2) espontaneidad; 3) adecuación; 4) representatividad; 5) autenticidad; 6) niveles diastrático, diafásico y diatópico; 7) variedades: coloquial y/o académica. El grupo de informantes está formado por hombres y mujeres de todas las zonas geográficas de Paraguay, en una distribución y participación equitativa por género — 50% cada —, y divididos en cinco grupos de edad (5 a 12 años; 13 a 24; 25 a 35; 36 a 65; y más de 65 años). Respecto de la metodología que se está empleando en la grabaciones, básicamente son: diálogos dirigidos entre uno o dos informantes y el investigador; entrevistas libres con intervención del encuestador; elocuciones formales; declaraciones semi-dirigidas; diálogo libre entre dos informantes; grabación secreta de un diálogo espontáneo; y elocuciones en actitudes formales (clases, conferencias, discursos etc.).

6.1. Descripción del *Corpus* Oral

Dicho *Corpus* se compondrá de la transcripción ortográfica de grabaciones de lengua hablada en Guaraní recogidas en diversas situaciones de habla (diálogo

dirigido, entrevistas, elocuciones formales, grabaciones secretas de diálogos espontáneos, grabaciones de llamadas telefónicas, programas de radio etc.).

Es importante mencionar que las muestras de género radiofónico y/o televisivo que formarán parte del *corpus* oral, se clasificarán siguiendo los criterios generales definidos en los apartados anteriores mencionados, y se agruparán de acuerdo a los siguientes géneros: 1) noticias; 2) reportajes; 3) entrevistas; 4) debates; 5) tertulias; 6) documentales; 7) retransmisiones deportivas; 8) magacines; 9) revistas deportivas; 10) variedades; 11) sorteos y concursos.

7. Conclusiones

En este trabajo hemos presentados los avances del proyecto de investigación *AVAKOTEPA*[2] *AVAÑE'Ẽ KO'ÂGAGUA OJEPURUHAÍCHA TETÃ PARAGUÁIPE*, cuyo objetivo principal es recopilar el primer *corpus* de referencia del guaraní contemporáneo (CORACON) y ponerlo a disposición de la comunidad científica en una plataforma flexible que pueda ser accesible a través de *internet*. Dicha base de datos pretende convertirse en una fuente empírica fiable para el estudio de cualquier aspecto relacionado con la lengua guaraní, ya sea con fines lingüísticos, lexicológicos o terminológicos. En este sentido, nuestro proyecto no sólo revalorizará el legado de cultura guaranítica mediante la elaboración de un inventario de su patrimonio, sino que también contribuirá a la mejor comprensión de los estudios existentes a través de la promoción de la investigación académica y la enseñanza de la lengua Guaraní con rigor científico. Además, aportará los elementos necesarios e imprescindibles en el proceso de normalización de lengua guaraní, permitiendo, inclusive, la recopilación del DRAC (Diccionario de Referencia de Guaraní Contemporáneo). Aunque este proyecto todavía está en ciernes y en la actualidad no se pueda acceder a la base de almacenamiento — interfaz de consulta — del *corpus*

de referencia, ya contamos con una base de datos de aproximadamente 500.000 formas gráficas — de la muestra textual — y más de ciento cincuenta horas de grabaciones de entrevistas, ambas en lengua Guaraní.

Referencias

BIBER, D.; CONRAD, S.; REPPEN, R. *Corpus linguistics. Investigating language structure and use.* Cambridge: Cambridge University Press (Cambridge Approaches to Linguistics), 1998.

CHURCH, K.; MERCER, R. Introduction to the special issue on computational linguistics using large corpus*. Computational Linguistics*, v. 19, n. 1, p. 1-24, 1993.

EAGLES. *Corpus typology*: a framework for classification. Informe interno n. 2.1 preparado por John M. Sinclair. Birmingham: Universidad de Birmingham, Corpus Linguistics Group, 1994.

_____. *Text corpus working group reading guide*. Documento Eagles (Expert Advisory Group on Language Engineering). EAG-TCWG-FR-2, 1996a.

_____. *Preliminary recommendations on corpus typology*. Documento Eagles (Expert Advisory Group on Language Engineering). EAG-TCWG-CTYP/ P, 1996b.

KENNEDY, G. *An introduction to corpus linguistics*. London: Longman (Studies in Language and Linguistics), 1998.

LAGER, T. *A logical approach to computational corpus linguistics.* (Tesis doctoral) — Gothenburg Monographs in Linguistics 14, Department of Linguistics, Gothenburg University, Sweden, 1995.

LEECH, G. The state of the art in corpus linguistics. In: AIJMER, K.; ALTENBERG, B. (Eds.). *English corpus linguistics. Studies in honour of Jan Svartvik.* London: Longman, 1991. p. 8-29.

_____. Corpus and theories of linguistic performance. In: SVARTVIK, J. (Ed.). *Directions in corpus linguistics*: proceedings of Nobel symposium 82. Berlin and New York: Mouton de Gruyter, 1992. p. 125-148.

LEECH, G.; FLIGELSTONE, S. Computers and corpus analysis. In: BUTLER, C. S. (Ed.). *Computers and written texts.* Oxford: Basil Blackwell, 1992. p. 115-140.

McENERY, T.; WILSON, A. *Corpus linguistics.* Edinburgh: Edinburgh University Press (Edinburgh Textbooks in Empirical Linguistics), 1996.

_____. *Corpus linguistics.* 2nd ed. Edinburgh: Edinburgh University Press, 2001.

MERCOSUR. *Tratado de Asunción.* En Acta n. 8/ 95. Montevideo, 30 nov. 1995.

SINCLAIR, J. *Corpus, concordance, collocation.* Oxford: Oxford University Press (Describing English Language), 1991.

SPERBERG-McQUEEN, C. M.; BURNARD, L. (Eds.). *TEI P4: guidelines for electronic text encoding and interchange*. Text encoding initiative consortium. XML Version.

Oxford, Providence, Charlottesville, Bergen, 2002. Disponible en: <http://www.tei-c.org/P4X/>.

TOGNINI-BONELLI, E. Towards translation equivalence from a corpus lingüistics perspective. *International Journal of Lexicography*, v. 9, n. 3, p. 197-217, 1996.

Notas

[1] Véase Tratado de Asunción: <http://www.parlamento.gub.uy/htmlstat/pl/tratados/trat16196.htm>.

[2] El director del proyecto es el Doctorando José Maria Rodrigues, Director del Área de Investigación del Centro de Postgrado e Investigación de la Facultad de Filosofía y Ciencias Humanas da la Universidad Católica "Nuestra Señora de la Asunción", y experto en lingüística de *corpus* y lingüística aplicada.