Crear un buscador léxico polígrafo para un corpus multilingüe en lenguas amerindias: el caso la base de datos LANGAS

Élodie Blestel, Stéphane Fouelefak

CREAR UN BUSCADOR LÉXICO POLÍGRAFO PARA UN CORPUS MULTILINGÜE EN LENGUAS AMERINDIAS: EL CASO DE LA BASE DE DATOS LANGAS*

Élodie Blestel
EA7345 CLESTHIA & UMR 7227 CREDA/Université Sorbonne Nouvelle

Stéphane Fouelefak
Télécom SudParis

Creating a Polygraph Lexical Search Engine for a Multilingual Corpus in Native American Languages: The Case of the LANGAS Database

Abstract: Some Amerindian languages widely spread before colonial times (Quechua, Aymara, Guaraní, Tupí) turned into vehicular languages under Spanish and Portuguese colonization since the sixteenth century. They have been written and called "general languages". They became the main means of communication between the indigenous population and Europeans; therefore, they were used to foster new economic and administrative spaces. Thus, these "general languages" served as interfaces between the colonial administration and the indigenous community, becoming written languages and giving birth to a rich and varied textual production. LANGAS database, created in 2011 (http://www.langas.cnrs.fr/), arose from the need to study these little-known documents, particularly the political vocabulary of these languages, as well as the natives' modes of expression of institutions and new concepts during the late colonial and the early republican eras.

Our purpose is to expose the problems of the large variation of the graphs in these linguistic varieties to implement a search engine on the LANGAS database. This aim involves translating each grapheme to computer language and establishing correspondences, which requires a few operations and carries out specific problems at the same time. After presenting the contents of the database and the way we process those manuscripts, we will identify the technical challenges we have to face, and we will also present a possible

* El proyecto colectivo e interdisciplinar LANGAS "Lenguas Generales de América del Sur" (Centro de investigación CREDA, UMR 7227) es financiado por la Agencia Nacional de Investigación francesa (2011-2016) y dirigido por la Dra. Capucine Boidin (IHEAL/Université Sorbonne Nouvelle) y el Dr. César Itier (INALCO): les agradecemos por sus lecturas críticas y sus consejos durante la redacción de este trabajo.

attempt of rationalization which is still under development: the “canonical” and “topological” searches.;

Keywords: language corpora; digital archive; lingua franca; colonial America.

La colonización española y portuguesa de América del Sur tuvo como efecto ampliar la implantación geográfica de algunos idiomas de vasta difusión prehispánica (principalmente el quechua, el aimara, el tupí y el guaraní) que se convirtieron en los principales medios de comunicación entre indígenas y europeos y permitieron vertebrar nuevos espacios económicos y administrativos. Fue así como estas “lenguas generales” sirvieron de interfaces entre la administración colonial y los indígenas, convirtiéndose en lenguas escritas y dando luz a una producción textual rica y variada. La base de datos LANGAS, creada en 2011 (<http://www.langas.cnrs.fr/>), surgió de la necesidad de estudiar estos documentos poco conocidos, en particular el vocabulario político de estos idiomas y los modos indígenas de expresión de instituciones y conceptos nuevos en las épocas colonial tardía y republicana temprana.

CUADRO 1
Página principal de la base de datos LANGAS (<http://www.langas.cnrs.fr/>)

Nuestro propósito es exponer los problemas que plantea la gran variación de las grafías en estas variantes lingüísticas antiguas para implementar un buscador en la base de datos. Esto supone traducir cada uno de los grafemas al lenguaje informático y establecer correspondencias, lo cual requiere unas cuantas operaciones y conlleva problemas específicos.

Tras presentar el contenido de la base y la forma con la cual procesamos los manuscritos, identificaremos los desafíos técnicos que se nos plantean y presentaremos un posible intento de racionalización que aún está en elaboración: la búsqueda “canónica” y la búsqueda “topológica”.

1. Presentación de la base de datos LANGAS

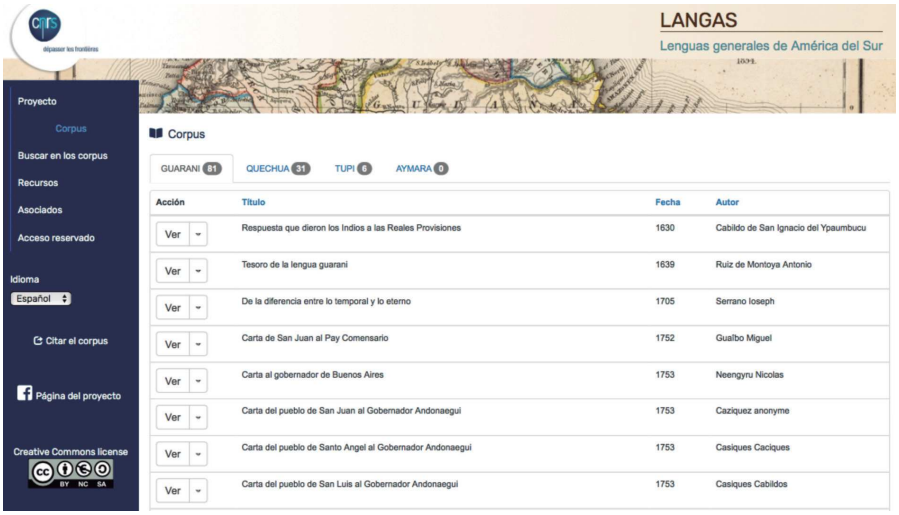
1.1. *El proyecto LANGAS*

El proyecto LANGAS ("LANGues générales d'Amérique du Sud" / Lenguas generales de América del Sur) reúne a un grupo de investigadores –historiadores, antropólogos, filólogos y lingüistas[1]– que se dedican al estudio de los documentos escritos que aparecieron en las épocas colonial y republicana temprana en lo que denominamos las "lenguas generales de Sudamérica", es decir, las principales lenguas vehiculares indígenas de uso extendido en esa zona (tupí, guaraní, quechua, aimara) que sirvieron de vehículo de comunicación entre hablantes de diversos idiomas amerindios y los europeos, lo que desembocó en la creación de nuevos espacios económicos y administrativos, así como en la evangelización de los indígenas.[2] Este proyecto, localizado en el Centro de Investigación del CREDA (IHEAL-Universidad Sorbonne Nouvelle)[3] y financiado por la Agencia Nacional de Investigación francesa, consiste en estudiar y comparar documentos escritos en esas lenguas –documentos que hasta ahora se hallaban dispersos en varios archivos– con el fin de contribuir a la historia social, semántica y cultural de las mismas. Para lograrlo, hemos implementado una base de datos multilingüe libremente consultable en línea (<http//:www.langas.cnrs.fr>) en la que vamos archivando los textos que venimos reuniendo desde el inicio del proyecto en 2011. En su fase actual, la base cuenta con unos cien documentos paleografiados y traducidos, los cuales van siempre acompañados de una ficha técnica con unas cien entradas en las cuales se especifican, en la medida de lo posible, todos los datos de los que disponemos para cada documento (fuente, lugar, fecha, número de folletos o páginas, editor si lo hay, etc.) y cada autor (trato, nombre, lugar de nacimiento/fallecimiento, cargo, etc.). Así, se puede consultar un abanico de textos de toda índole (metalingüística, administrativa, religiosa) provenientes de los siglos XVI, XVII y XVIII: actualmente están en consulta 31 documentos en lengua quechua fechados entre 1560 y 1823, 81 en lengua guaraní fechados entre 1630 y 1813 y seis en lengua tupí fechados entre 1575 y 1686. La base ha sido pensada para permitir búsquedas léxicas con el fin de poder comparar el uso de conceptos en diferentes textos de una misma lengua o en textos de distintas lenguas generales para contribuir al entendimiento de los procesos que luego darían lugar a los procesos independistas, por ejemplo. No obstante, el buscador no se

1 La lista de los miembros del equipo se puede consultar en <http://www.langas.cnrs.fr/#/description>.
2 Ver al respecto Estenssoro/Itier (2015).
3 El proyecto cuenta también con el apoyo del INALCO (Institut des Langues et Civilisations Orientales) de París.

limita a hacer posibles investigaciones semánticas, ya que ofrece una muestra hasta ahora inédita de los cambios fonológicos y morfosintácticos que se han dado en cada una de las lenguas en diacronía, por lo cual interesa tanto a antropólogos e historiadores como a (socio)lingüistas que tengan interés en estudiar la evolución de dichas lenguas.

CUADRO 2
Muestra del corpus en lengua guaraní
(<http://www.langas.cnrs.fr/#/consulter_corpus/liste/1>)



1.2. *Procesamiento de los manuscritos*

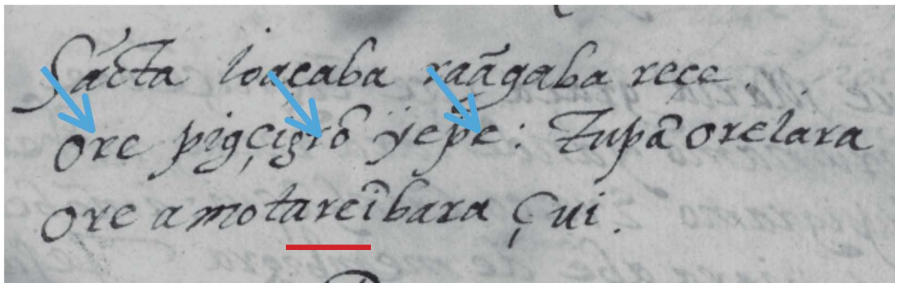
Cada uno de los documentos presentes en la base pasa por un proceso de localización en distintos archivos (museos, archivos y bibliotecas nacionales de Europa y Sudamérica), digitalización, reclasificación (según los criterios de las fichas técnicas de la base), transcripción diplomática, transcripción paleográfica, transliteración y traducción. Describimos brevemente cada etapa a continuación.

1.2.1. Localización, reclasificación y transcripción diplomática

A partir de los conocimientos historiográficos actuales, se localizan fondos de archivos susceptibles de conservar documentos en lenguas amerindias generales. Misiones de archivo *in situ* permiten entonces una investigación sistemática

y el hallazgo de manuscritos e impresos hasta hoy no identificados.[4] Luego, procedemos a la paleografía diplomática de los documentos respetando la compaginación (disposición, párrafos, saltos de línea, saltos de página, etc.) y la grafía. La reproducción de las grafías implica a menudo una interpretación para decidir cuáles son pertinentes y cuáles no, como en el ejemplo siguiente:[5]

CUADRO 3
Muestra de grafía en un manuscrito original en guaraní



Consideramos que los dos primeros signos diacríticos presentan diferencias mínimas. Los representamos en la paleografía diplomática con el mismo diacrítico llamado "macron acute", que corresponde a una sola entidad html (᷄). En cambio, el tercero es diferente y necesita otra entidad html.[6] El resultado es:

Sa̅cta ioaçaba raa̅gaba reçe
Ore pigçigrô yepe: Tupâ oreiara
Ore amotareîbara çui.

Dependiendo del estado en el cual se encuentran los documentos, a veces hace falta añadir secuencias (letra, sílaba, palabra) que no están en el documento original, lo cual se señala mediante el uso de corchetes [ ]. Cuando encontramos una secuencia cuya lectura o interpretación es dudosa, la señalamos mediante el

4 A modo de ejemplo, un manuscrito monolingüe en guaraní de 280 páginas, ver al respecto Adoue/Orantin/Boidin (2015).

5 El ejemplo elegido es un manuscrito identificado casualmente hace pocos años en la biblioteca de la Universidad de Oxford por Vivian Kogut. Es un manuscrito monolingüe de 222 páginas, en "tupi", con los títulos en portugués (MS. Bodl. 617). Titulado *Doutrina christaã na língua brasilica* es un manuscrito anónimo, sin fecha, que fue robado por Thomas Lodge en la biblioteca del colegio jesuítico de Santos en 1591. Cândida Barros y Ruth Monserrat, junto con Jessica Moreira están analizando el manuscrito. Ruth Monserrat y Capucine Boidin tomaron las decisiones relativas a la paleografía a insertar en la base Langas.

6 â = ȃ ê = ȇ î = ȋ ô = ȏ û = ȗ

uso de paréntesis ( ). Si, en cambio, no logramos leer la secuencia, usamos puntos suspensivos entre paréntesis (…). Por último, si a pesar de que una laguna sea evidente dado el contexto, no tenemos ninguna hipótesis acerca de ello, lo señalamos mediante puntos suspensivos entre corchetes […].

1.2.2. Transcripción paleográfica utilizada en la base

La segunda etapa consiste en adaptar la paleografía diplomática a las necesidades de la base. Eliminamos la composición original del documento y solo conservamos las particularidades gráficas, la puntuación, la segmentación y las abreviaciones presentes en los originales. La diferencia con la paleografía diplomática estriba, pues, en el hecho de que en esta segunda paleografía, la disposición del texto se ve modificada, ya que suprimimos los saltos de página, los saltos de línea, las marcas de párrafos y la segmentamos en pequeños párrafos que ponemos en correspondencia, ya sea con la traducción en español de la época,[7] ya sea con la traducción que hacemos nosotros en español actual. El resultado en la base es el siguiente:

CUADRO 4
Ejemplo de transcripción paleográfica en la base

2 Sãcta ioçaba raãgaba reçe Ore pigçigrô yepe: Tupâ oreiara Ore amotareĩbara çui. Persinar. Tuba, taigra, Tupâ spû Santo rerapupe. Amen Jesu. (f.1r)

1.2.3. Transliteración

Sigue la etapa de la transliteración que consiste en la substitución del sistema gráfico de los textos paleografiados por otro más sistemático y riguroso. El quechua, por ejemplo, presenta un sistema consonántico muy diferente al del español. Esto explica que encontremos sistemas gráficos muy poco precisos en los textos antiguos: no traducen ninguna oposición fonológica y a veces presentan una proliferación de dígrafos que no corresponden a nada sistemático. De ahí la necesidad de esta etapa de transliteración: no solamente para que estos textos

7 En este caso, las traducciones en español de la época se tratan de la misma forma, esto es, suprimiendo los elementos relativos a la materialidad del texto.

sean leíbles, sino para formular una hipótesis sobre lo que puede haber sido el estado fonológico de la lengua según la época, el lugar e incluso la procedencia social del autor, permitiendo así que los textos, inicialmente redactados en sistemas gráficos muy dispares, puedan dar lugar a un trabajo de comparación.

El sistema de reglas de transliteración es distinto para cada una de las lenguas del corpus y ha ido evolucionando con el tiempo y con los paleógrafos. Para el tupí, por ejemplo, se considera por el momento que no es necesario agregar una transliteración. Como hemos dicho, las transliteraciones son solamente hipótesis sobre los distintos estados de lengua, tomando en cuenta las variaciones dialectales, sociolectales e incluso diacrónicas. Estas hipótesis se fundan en un análisis filológico de cada texto, con lo cual, aunque facilita el acceso a los textos, no se puede considerar como substituto a las paleografías de los textos originales.

1.2.4. Traducción

Finalmente, todos los textos aparecen junto con al menos una traducción al español. Cuando disponemos de la traducción en el español de la época, la agregamos. Cuando no, proponemos nuestra traducción. En el caso de algunos documentos, ofrecemos tanto la traducción de la época como una propuesta de retro-traducción,[8] de modo que los manuscritos se recopilan en tablas de hasta cuatro columnas (transcripción paleográfica / transliteración moderna / traducción en español moderno / transcripción en español original) (cuadro 5).

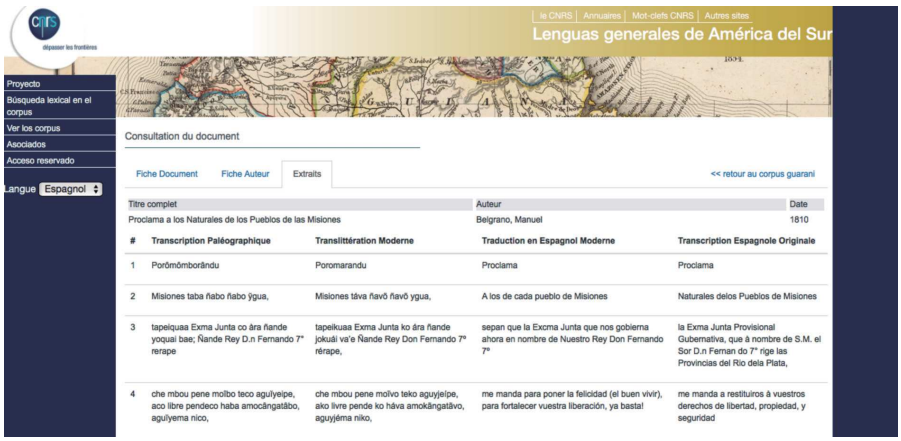
2. Implementar un buscador polígrafo para un corpus multilingüe en lenguas amerindias: ejemplo del corpus guaraní

2.1. *De la necesidad de un buscador polígrafo*

Para que la base de datos sea de uso más fácil, se ha implementado un buscador que permite hacer búsquedas léxicas en las paleografías. Como hemos expuesto anteriormente, las grafías de los textos paleografiados son muy heterogéneas, lo que obviamente se explica por la diversidad de las grafías en los documentos originales, pero también por la inserción de nuevos elementos durante el proceso de transcripción paleográfica.

8 En particular cuando se trata de textos originalmente en castellano y vertidos a lenguas amerindias. Ver los documentos en guaraní de la época independentista.

CUADRO 5
Ejemplo de manuscrito procesado en lengua guaraní:
"Proclama a los Naturales de los Pueblos de las Misiones"
(<http://www.langas.cnrs.fr/#/consulter_document/extraits/3>)



Varios motivos explican la gran variación patente en los documentos originales. Uno de ellos es el hecho de que todas las lenguas del corpus fueron transcritas en alfabeto latino, el cual ya no correspondía a ningún sistema fonológico de ninguna lengua de la época colonial: en efecto, para las lenguas europeas del siglo XVII, el alfabeto latino ya no correspondía sino a una solución gráfica solo parcialmente fonológica para la transcripción de las lenguas (español, portugués, italiano, etc.). Si añadimos a ello el hecho de que muchos sonidos amerindios no existían en las lenguas del viejo continente, entendemos cuán dificultosa podía resultar su transcripción gráfica. A eso hay que sumar dos factores más: como lo menciona Marc Thouvenot (1992: 45), en esa época se admitía mucho más fácilmente que una misma palabra tuviera varias grafías y, por otra parte, esa gran variabilidad también dependía del grado de educación y de la sensibilidad del oído del transcriptor.

Otra dificultad para implementar un buscador es la variación introducida en las transcripciones paleográficas: como hemos expuesto arriba, utilizamos corchetes, paréntesis y puntos suspensivos cuando tenemos dudas o lagunas en el texto original. Esto constituye un obstáculo más para la implementación de un buscador eficiente en la base.

Como consecuencia de esto, la gran variabilidad de los elementos presentes en el corpus se hace patente en distintos aspectos: los grafemas utilizados, la segmentación de las palabras, la puntuación y la presencia de algunas variantes morfofonéticas debidas a variaciones idiolectales, sociolectales, dialectales y

diacrónicas. No podemos considerar que se neutralizan estas variaciones con el proceso de transliteración, ya que este solamente constituye una hipótesis y, además, las reglas de transliteración son distintas según las lenguas del corpus. Es el motivo por el cual nos parece imprescindible que el buscador rastree las transcripciones paleográficas. Ahora, cuando un usuario formula una búsqueda en estas, tiene que poder elegir entre formular una búsqueda "exacta" (es decir, buscar solamente los elementos tal y como los entra en el buscador) o bien buscar –¡y encontrar!– todas las variantes posibles de una secuencia dada, sin que las variaciones mencionadas arriba constituyan un obstáculo. Para lograr este objetivo, la creación de un buscador polígrafo consiste en asociar un criterio de entrada –es decir, una secuencia en una sola grafía, con una segmentación única– con fragmentos que contienen el conjunto de sus variantes gráficas, sin importar la segmentación y la puntuación adoptadas. A esto tenemos que añadir otra dificultad: puede darse el caso de que el usuario no especialista produzca variaciones que no existen en el corpus. Esta posibilidad también tiene que ser considerada como otra variante posible.

2.2. *Primeros intentos de racionalización*

Tenemos pues dos tipos de variaciones: las variaciones internas al corpus paleográfico y las variaciones externas a este, ya que estas últimas son introducidas por el usuario de la base. En las versiones del buscador que se están implementando, dejamos de lado el problema de la segmentación temporalmente. Dos tipos de búsquedas polígrafas están en elaboración: la búsqueda canónica y la búsqueda topológica. El primer tipo de búsqueda, "canónica", considera el texto como evento de lenguaje: se basa en una caracterización de la noción de variante gráfica como resultado de un razonamiento en parte externo al texto. El segundo tipo de búsqueda, "topológica", considera el texto como un flujo de caracteres y define la noción de variante gráfica a partir de un solo flujo de caracteres. Tras presentar el sistema actual, presentaremos en lo que sigue la estrategia que estamos adoptando y cada una de estas soluciones técnicas.

2.2.1. El sistema actual

Desde un punto de vista técnico, el sistema actual está constituido de una aplicación web clásica escrita en PHP y alojada en un servidor Apache. Los fragmentos de texto se salvaguardan en una base de datos MYSQL. Es pues la base de datos la que ofrece los servicios de búsqueda de bajo nivel. Los desarrollos informáticos que estamos implementando ahora consisten entonces en una utilización parti-

cular de estos servicios de bajo nivel con la ayuda de programas escritos en PHP. Habitualmente, el sistema funciona de la manera siguiente:

– El usuario introduce un criterio de búsqueda, por ejemplo, “porokuaita” (‘mandamiento’ en guaraní);
– Se genera una consulta mediante los programas PHP que se somete a la base de datos: select * from extracts where contenu rlike ‘porokuaita’;
– El resultado generado son los fragmentos que contienen el flujo de caracteres ‘porokuaita’ (34 fragmentos);
– Los fragmentos que contienen las variantes “porocuaita” y “poroquaita” no son encontrados.

Sin embargo, otra posibilidad existe: consiste en la utilización del punto como expresión de una alternativa gráfica o segmentacional. Veamos un ejemplo:

– Un criterio de entrada puede ser, por ejemplo, “poro.uaita”;
– Los programas PHP generan una consulta: select * from extracts where contenu rlike ‘Poro.uaita’;
– Se encuentran los resultados que contienen “Porouaita”, “Porokuaita”, “Porocuaita” e incluso “Poro uaita” (lo que constituiría una segmentación alternativa) y “Porozuaita”, si existiesen (64 fragmentos).

Ahora bien, esta utilización del punto solo hace variar un solo carácter: cuando la secuencia es de tamaño *N*, esta expresión habría que introducirla una vez. Pero podríamos necesitar hacer variar dos caracteres en una misma secuencia. En este caso, si la secuencia es de tamaño *N,* el número de posibilidades sería entonces de N^2, etc. Entendemos que esta solución no es satisfactoria, incluso si no es técnicamente imposible. Además, con este sistema, la base de datos no es asequible a todo público, ya que el usuario tiene que anticipar cuáles pueden ser las variaciones gráficas utilizando el punto donde le parece que los textos del corpus presentarán grafías heterogéneas. Nosotros, en cambio, contemplamos la posibilidad de que incluso un usuario no especialista encuentre todas las variantes gráficas, sin importar el criterio de entrada, es decir, sin importar la grafía y la segmentación que utiliza para formular su consulta en el buscador.

2.2.2. La búsqueda canónica

Para lograr nuestro objetivo, hace falta un enfoque más global. Partimos de una constatación muy simple: los manuscritos fueron escritos según lo que los autores escuchaban, o creían escuchar. Formulamos entonces la hipótesis de que

todas y cada una de las variaciones pueden relacionarse con su correlato fonológico. Nuestra estrategia consiste entonces ya no en trabajar a partir de los grafemas, sino a partir de su proyección fonológica. Dicho de otra manera, todas las entradas formuladas en el buscador mediante cierta transcripción gráfica tienen una sola traducción fonológica. Como consecuencia de ello, decidimos transponer la búsqueda del espacio de los grafemas al de los fonemas –lo que viene a constituir el espacio "canónico"–, para deshacernos de una parte importante de las variaciones gráficas. Para conseguirlo, elaboramos tablas de correspondencias fonema < > grafemas según el modelo que sigue:

CUADRO 6
Ejemplo de tabla de correspondencias fonema < > grafemas en guaraní

Fonema	Realizaciones fonéticas	Reglas para el generador		
		Grafía actual	Grafías antiguas prototípicas	Ocurrencias paleográficas menos frecuentes
Vocales orales				
/i/	[i]	-i- -í-	-i- (*rupi*) -í- (*cunumí*) -y- (*ychupe*) -î- (*îru*, *mî*)	-ì- (*oroì*) -ý- (*mondǐý*) -ỳ-
	[j] (semi-cons.)	-j-	-y- (*yaiquaa*)	
	[i̯] (semi-voc.)	-i-	-y- (*aypo*) -i- (*acoi*) -ĩ- -ỹ- (*ruguâỹ*)	-ĭ- (*rũĭ*) -î- (*oñoîrârô*) -î- (*acoî*)
	[ĩ] si contexto nasal		-î- (*oquirîrî*)	
	[j̰] entre dos vocales o en ataque silábico	-j-	-y- *(aguĭye)*	-i- (*aguĭiebe*)
/ʔi/	[ʔi] o [ʔĩ]	-'i- -í-	-ý- (*Paý*) -y- (*hey*) -í- (*heí*) -ì- (*heì*)	-'i- -ỳ-

A partir de estas tablas, se elaboran dos programas:

– El primer programa relaciona las grafías con sus correlatos fonológicos;

– El segundo programa hace lo mismo para los textos de la base. Pero como el texto contiene varias secuencias (o palabras), elegimos una sola alternativa para cada secuencia. Esto significa que cuando los miembros del equipo suben la versión paleográfica de un texto a la base, salvaguardamos su proyección fonológica. Luego, cuando el usuario entra un criterio escrito, este es transformado en un criterio fonológico.

Podemos resumir el proceso de la manera siguiente:

– Inserción o modificación de un texto: se salvaguarda el texto original y una de sus variantes canónicas (transcrita en fonemas);

– Tratamiento del criterio: se transforma el criterio del usuario (una variante gráfica) en una lista de criterios (el conjunto de contrapartidas canónicas);

– Consulta: se rastrea el conjunto de criterios canónicos en las versiones canónicas de los textos;

– Resolución de correspondencia: si un texto canónico coincide con la búsqueda canónica por un término en posición “p”, entonces el fragmento gráfico correspondiente coincide con la búsqueda en posición “p”.

Ahora, esta estrategia que consiste en proyectar el espacio escrito en el espacio fonológico presenta dos limitaciones. La primera tiene que ver con el hecho de que algunas contrapartidas fonológicas no tienen sentido: algunas de ellas no podrían ser pronunciadas o no podrían existir en las lenguas del corpus. Tenemos que mejorar este método añadiendo restricciones para lograr que sea más eficiente.

La segunda limitación es que esta solución tiene como fundamento el que todo texto es un evento de lenguaje: consideramos que todas y cada una de las variaciones gráficas caben en una sola variante fonológica. Sin embargo, esto no es totalmente cierto: según la variante gráfica con la que entramos a la base, nos damos cuenta de que hay diferencias en su proyección fonológica. Pues en teoría, a partir de la primera variante gráfica, tenemos la certeza de que vamos a encontrar todas las demás variantes. Pero a partir de la segunda variante, si hacemos un mero cálculo de recuento, solo tendríamos un 50 % de probabilidad de encontrar la primera variante, y a partir de la tercera, solo un 25 % de probabilidad de encontrar la primera y la segunda otra vez. En resumidas cuentas, con el sistema canónico, es posible que el buscador no rastree algunas secuencias (o palabras). Para remediarlo, hemos decidido completar este sistema de búsqueda con otro que enfoca el texto de manera intrínseca, es decir, como una simple secuencia de caracteres. Esta segunda búsqueda –la llamamos “topológica”– nos

permite verificar en cada momento que la búsqueda canónica no omite ninguna secuencia de caracteres.

2.2.3. La búsqueda topológica

La búsqueda topológica se basa en el concepto de distancia de Levenshtein. La distancia de Levenshtein entre dos palabras corresponde al número mínimo de operaciones (eliminación, inserción, substitución) que permiten transformar una cadena de caracteres en otra. Por ejemplo, en guaraní, las variantes gráficas <rembiaïhu>, <rembiahu>, <rembiaihu> están a una distancia de Levenshtein de 1 unas de otras.

La búsqueda topológica toma en cuenta dos criterios de entrada: una secuencia (la variante que se busca) y un entero de búsqueda (que corresponde a la distancia de Levenshtein). El entero es *a priori* 0, 1, 2 o 3, lo que significa que se rastrean las variantes gráficas que se encuentran entre las variantes de Levenshtein de orden máximo 3. Por ejemplo:

– El usuario introduce un criterio: (“Porokuaita”, 1). Esto se puede interpretar literalmente como la búsqueda de los fragmentos que contienen la cadena de caracteres “Porokuaita” o toda otra cadena de caracteres distante de 1 según la distancia de Levenshtein;

– La consulta la generan los programas PHP y se somete a la base de datos formulada de la siguiente manera: select * from extraits where (contenu rlike ‘.orokuaita’ or contenu rlike ‘P.rokuaita’ or contenu rlike ‘Po.okuaita’ or contenu rlike ‘Por.kuaita’ or contenu rlike ‘Poro.uaita’ or contenu rlike ‘Porok.aita’ or contenu rlike ‘Poroku.ita’ or contenu rlike ‘Porokua.ta’ or contenu rlike ‘Porokuai.a’ or contenu rlike ‘Porokuait.’ or contenu rlike ‘.Porokuaita’ or contenu rlike ‘Porokuaita.’).

Esta búsqueda nos permite verificar en cada momento que la búsqueda canónica no deja nada de lado, ya que las variantes fonológicas caben supuestamente en las variantes del orden 1, 2 y tal vez 3.

Conclusión

Para implementar el buscador de la base de datos LANGAS y solucionar el problema de la gran variación de las grafías, partimos primero de un sistema de búsqueda “canónica”, que consiste en proyectar las entradas gráficas y los textos en un espacio fonológico. Este sistema se funda en el postulado de que cada

manuscrito constituye un evento de lenguaje y permite reducir una gran cantidad de variantes a una sola unidad fonológica. Este sistema necesita mejoras, en particular en cuanto a restricciones sobre las secuencias fonológicas que no tienen sentido. Como complemento a este primer sistema de búsqueda, elaboramos otro, llamado "topológico", que considera el texto ya no como evento de lenguaje, sino simplemente como secuencia de caracteres. Basado en la distancia de Levenshtein, permite verificar que la búsqueda canónica no deja ninguna secuencia de lado. Estos dos sistemas son los dos que se están elaborando en la fase actual del proyecto LANGAS.

Otro desafío que queda por resolver es el problema de la variación de segmentación. Efectivamente, razonamos hasta ahora en términos de secuencias o palabras cuando en realidad la segmentación de estas también experimenta variaciones. La solución por la cual estamos optando ahora es considerar la cuestión de la variación de segmentación como un problema aparte y tratar de resolverla antes del de la variación gráfica. En efecto, como los espacios constituyen estorbos para la implementación del buscador, lo que pensamos hacer es suprimirlos y proceder a una segmentación automática. Tendremos así tres versiones de los textos: una versión paleográfica, la versión segmentada automáticamente (y de manera constante en todo el corpus) de esta y su proyección fonológica. Nos parece que resolveremos así el problema de la variación gráfica y el de la variación de segmentación a la vez. Ahora, nos queda todavía mucho por hacer: crear los algoritmos que permiten una segmentación automática de los textos requiere un análisis pormenorizado de la estructura de cada una de las lenguas del corpus. Es un desafío más para los miembros del equipo LANGAS y a eso nos dedicaremos en los próximos meses.

Referencias

ADOUE, Cecilia/ORANTIN, Mickaël/BOIDIN, Capucine (2015): "Diálogos en guaraní, un manuscrito inédito de las reducciones jesuitas de Paraguay (siglo XVIII)" [en línea], en: *Nuevo Mundo Mundos Nuevos* (Debates) 15. <https://nuevomundo.revues.org/68665> (15 enero 2016).

BOIDIN, Capucine/CHAMORRO, Graciela/MÉRET, Géraldine (2014): "Introducción al dossier 'Fuentes en lenguas amerindias de América del Sur'" [en línea], en: *Corpus* 4, 2. <http://corpusarchivos.revues.org/1335> (15 enero 2016).

ESTENSSORO, Juan Carlos/ITIER, César (dirs.) (2015): "Introducción al dossier 'Langues indiennes et empires dans l'Amérique du Sud coloniale'", en: *Mélanges de la Casa de Velázquez* 45, 1, pp. 9-14.

THOUVENOT, Marc (1992): "Temoa", en: *Amerindia* 17, pp. 45-68.