16. Corpus informatizado de la lengua ibérica y herramientas de trabajo

In memoriam, Frederic Santaeulària Roig

David Folch Florez

16.1. Currículum

a) Estudios y actividad en este tema

- David es licenciado en Farmacia y en Física por la UB. Colaborador CTTC-UPC (ingeniería).
- Es director del Grup de Recerca de l'Institut d'Estudis Ibers".
- Autor junto a Carme Jiménez Huertas del corpus informatizado en lengua ibérica.

b) Publicaciones

- Es escritor y colaborador en periódicos digitales

c) Contacto

- folch.florez[arroba]cofb.net

16.2. Resumen

El estudio de la lengua ibérica dispone de nuevas herramientas informáticas que pueden ayudar a profundizar en su comprensión.

En esta conferencia presentaremos el corpus informatizado, realizado por David Folch y Carme Jiménez Huertas. Este corpus consiste en un fichero en formato hoja de cálculo (excel) que se puede descargar gratuita y libremente de la página web www.ibers.cat, así como la fuente "iberian" necesaria para su lectura.

En este fichero están recogidas más de 2800 epigrafías (en más de 4000 líneas), todas en signario ibérico nor-oriental. Esto supone como mínimo el 90% de todas las inscripciones encontradas hasta el 2015. Los datos están ordenados partiendo de la clasificación de

Untermann, más todos los registros posteriores, con una identificación unívoca y fácil de seguir online.

De cada pieza se hace una mínima descripción física y se han anotado todas las dificultades de lectura o interpretación. El fichero consta de otro apartado, más reciente, con las cadenas de caracteres repetidas en los textos, marcadas en un código de colores y que se explicará su utilización en la charla.

Por último, se mostrará el magnífico buscador on-line desarrollado por Joan Vilaseca y sus aplicaciones prácticas, esperamos que con grandes resultados.

16.3. Ponencia

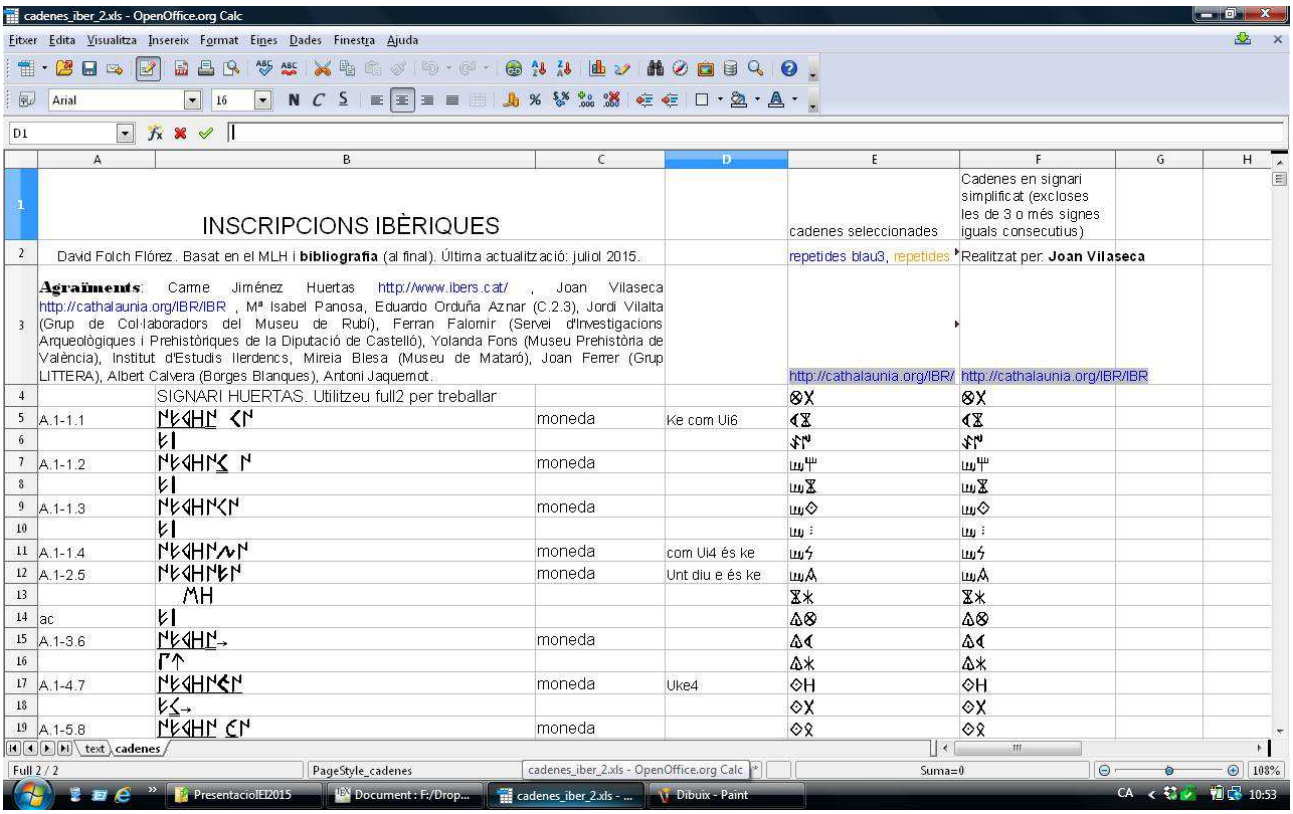
En el arduo camino hacia el conocimiento completo de la cultura ibérica es fundamental contar con herramientas potentes y de largo alcance para descifrar los textos ibéricos, que todavía ocultan su significado. En este artículo presentaremos el corpus informatizado en signario nororiental y sus posibles formas de uso, y esbozaremos unas etapas que nos permitan avanzar en la comprensión de los textos, sin dejar de lado el objetivo final, esto es, entenderlos.

a) Herramientas básicas de trabajo

Desde el estudioso alemán J. Untermann0, se han ido recopilando todos los textos ibéricos (epigrafías) encontrados y en mayor o menor medida se han ido publicando para su análisis posterior. Todas estas referencias acaban estructurando un corpus bastante considerable, pero que todavía no ha sido descifrado.

Para este objeto, y conociendo la enorme potencialidad del tratamiento informático, decidimos trasladar todas (o la mayoría) de las epigrafías al formato informático, tecleando y repasando uno por uno todos los símbolos (letras) ibéricos. Para mantener un orden y facilitar la identificación y uso, cada entrada dispone de un código lógico unívoco, una descripción somera del soporte material donde está el texto y unos comentarios con las diversas (y no pocas) dificultades del traslado del texto real a su imagen informática.

Todo esto con una característica que consideramos esencial: los textos no están transliterados, es decir, no se usa el alfabeto latino para representar (aproximadamente) el sonido de los símbolos ibéricos (más o menos conocido) sino que se usa el signario ibérico nororiental cuya fuente informática ha sido creada por Carme Jiménez Huertas1. Más adelante explicaremos porqué creemos que es esencial.



Otra característica a destacar de este corpus es su acceso gratuito y universal. Tanto la fuente Iberian de Carme J. Huertas como el archivo en hoja de cálculo de todas las inscripciones se pueden descargar del la página web www.ibers.cat tantas veces como se desee. Todo el que quiera puede desde ahora trabajar con la mayoría de textos en su signario original y a tan sólo un par de clics.

El corpus informatizado, actualizado en julio de 2016, contiene más de 2800 epigrafías, aproximadamente 4200 líneas de texto, que representan como mínimo el 90% de todas las epigrafías encontradas hasta el 2015. Es propósito de los autores actualizarlo cada año, a través de las publicaciones de referencia que se explicitan en el archivo.

Por si sólo, por su completitud, su facilidad de uso y su fidelidad (salvo error o interpretación) a los textos originales, este corpus es una buena herramienta de consulta. Pero también abre la puerta a la aplicación de metodologías de trabajo imposibles de realizar sobre papel, entre las que proponemos una en particular:

b) Metodología de las cadenas de caracteres repetidas

Antes de entrar en detalle en nuestra propuesta2, analizaremos algunos aspectos generales sobre el descifrado de la lengua ibérica. Una de las formas tradicionales de estudio de los textos es afrontar una de las inscripciones y analizarla en base a conocimientos de otras lenguas y de trabajos previos de otras inscripciones. Es decir, podríamos decir que es un método secuencial. Al disponer de un corpus informatizado, esto es, un archivo con todas las inscripciones, se abre la puerta a un análisis de forma global (en paralelo) de todos los textos a la vez, y que puede ser auto-contenido, sin referencia previa a otras lenguas ¿Será ésta la clave del descifrado? Lo que es seguro es que es una forma alternativa y potente de trabajar.

En el “ataque” a un determinado texto ibérico hay que tener siempre en cuenta la pérdida o error en la información que hay desde la pieza original hasta el texto en papel o en soporte informático. Analizamos tres transformaciones que se llevan a cabo, algunas siempre y otras muy frecuentemente: la primera, la interpretación de los signos. Es decir, qué símbolo ibérico concreto es el que está escrito en la pieza original. En este punto, debemos otorgar toda la confianza a los que estudian la pieza (mediante lo que se llama autopsia) y que proceden a publicar el resultado. En esta transformación, poco podemos hacer.

La segunda, la transliteración. Tal cómo ya habíamos comentado es el cambio del símbolo ibérico a un equivalente (aproximado) en alfabeto latino. Esto produce una pérdida de información que quizá nos pueda deparar sorpresas en el futuro. Cuando no es posible usar el signario ibérico, porque no hay (o había) la fuente informática, es lógico o incluso preciso usar un sistema alternativo para publicar los textos, pero puede dificultar la comprensión final del texto. Cada vez más, los autores, ante la falta de una fuente aceptada por todos, adoptan el método alternativo de asignar un código a cada símbolo (por ejemplo a1, para un tipo determinado de “a” ibérica) que evita la pérdida de información a costa de una lectura farragosa, entre otros problemas menores*. En nuestro corpus, con signario original, evitamos el problema.

Finalmente, la segmentación, esto es, la separación de los textos en “palabras”. Recordamos aquí que es norma habitual de los íberos escribir “todo seguido”, es decir, sin separación entre las unidades morfo-sintácticas UMS (fragmentos con sentido o “palabras”), con la consecuente complicación en la interpretación del texto. Huelga decir cuántas interminables discusiones se pueden realizar y se realizan ante la segmentación de un texto determinado y cuán diferente resultado se deriva, con siempre elaborados argumentos de cada uno de los autores. Éste es también un punto clave y es aquí donde aplicamos lo que ya hemos comentado del trabajo global o en paralelo sobre todos los textos.

*Por ejemplo, si disponemos del signario original, es más fácil visualmente hacer suposiciones sobre errores en la transcripción de signos.

La idea básica y simple es trabajar con las cadenas de caracteres repetidas que se encuentren en los textos, con la esperanza, que luego habrá que confirmar o desmentir, que sean UMS o “palabras” con sentido: el trabajo lo realiza inicialmente el ordenador, de forma eficiente pero “ciega”, con lo que se necesita este pos-procesado humano, que inevitablemente puede ser origen de errores. Buscamos obtener una imagen como ésta:

Presentamos aquí los pasos de la metodología propuesta y los resultados obtenidos hasta el estudio de cadenas de 4 o más caracteres, datos que podéis analizar en el mismo archivo del corpus:

1- Recopilación de las cadenas de caracteres repetidas. Realizado ya por el informático Joan Vilaseca.
2- Estudio de estas cadenas: determinación de cuáles son realmente unidades morfo-sintácticas UMS
3- Estudio de las características y formas de uso de estas UMS dentro de los textos: prefijos, sufijos, variaciones, combinaciones, etc.
4- Estudio de las cadenas cortas que queden delimitadas por las UMS
5- Agrupación según las características de las UMS
6- Elaboración del léxico: asignación de categorías tipo nombre, verbo, etc. a cada grupo, si procede. Y, si es posible, traducción.

La mejor manera de explicar el proceso es con un ejemplo en castellano. Imaginemos tres epigrafías diferentes, escritas a la manera ibérica, sin separaciones, sin acentos y con textos un tanto forzados para simplificar el estudio:

a- LACASA**MAS**ALTA**DEL**CAMINO
b- UNATORREALTACAE
c- PUERTADELACASA**AL**CAMINO
...

LACASAMASAL~~TADEL~~CAMINO

En su análisis ciego, el ordenador encontraría las siguientes candidatas a “palabra”: CASA, ALTA, CAMINO. Como sabemos castellano, estamos contentos ya que fácilmente hemos determinado las palabras repetidas. Pero hay trampa: el ordenador también habrá dado la opción TADEL que aparece en a) y c) y que sabemos perfectamente que no tiene sentido, con lo que deberíamos rechazarla. Este es el eslabón más débil del método, el que requiere de la pericia del estudioso y el que otra vez puede dar lugar a diferentes interpretaciones. Pero cabe decir que en nuestro trabajo, y a medida que se buscan una por una las “palabras” encontradas en cada texto, y se analizan, es “fácil” proceder a la eliminación de muchas candidatas, por comparación de longitud de cadena, de frecuencia de uso, de aparición unívoca de otra opción, etc. Es decir, TADEL es solo posible si ALTA y PUERTA no son “palabras”. Cualquiera que sea la decisión tomada es reversible y está consignada en nuestro trabajo.
Como beneficio extra, y marcado en negrita en nuestro ejemplo, aparecen cadenas lo suficientemente cortas para ser candidatas a partículas o modificadores con sentido (ejemplo: MAS, DEL, AL y hasta LA).

En las siguientes imágenes os mostramos cómo quedan algunos textos una vez identificadas las UMS (recordad que sólo hemos trabajado hasta cadenas de 4 o más caracteres, quedando pendiente realizar el estudio para cadenas más cortas). En algunos casos, los resultados son muy alentadores.

Dediquemos unas explicaciones al código de colores usado: los tonos azulados, lilas y derivados dan cuenta de las UMS ya determinadas como tal y marcadas en los textos. Los tonos naranja sirven para resaltar fragmentos compartidos con otras “palabras”, con la posibilidad que por sí solos ya signifiquen algo. Los tonos en verde son UMS que continúan en la línea de abajo. Otro ejemplo:

En el archivo encontraréis todos los textos, más las cadenas de caracteres encontradas a ciegas informáticamente, más las cadenas seleccionadas como UMS. A partir de aquí, abrimos el escenario para completar el trabajo y seguir los siguientes pasos de la metodología a todos aquellos que quieran participar.

c) El buscador de Joan Vilaseca

Debemos mencionar una herramienta fundamental desarrollada por Joan Vilaseca en su web www.cathalaunia.org y que amplifica el uso del corpus. Su autor presenta y ya da cumplidas instrucciones sobre su uso en su página web. Aquí comentaremos que el corpus es totalmente online; que tiene un buscador en el que se pueden teclear los símbolos y como resultado presenta las inscripciones; que marca las cadenas de caracteres repetidas (sin filtrar); que abre abanicos con diferentes textos semejantes, y muchas otras funcionalidades sorprendentes. En definitiva, una herramienta imprescindible para realizar las búsquedas en nuestro estudio, que por sí sola llenaría un artículo y presentación completa por parte de su autor, y que os animamos a utilizar por su potencia, simplicidad y profundidad de resultados.