Actas del XXXV Simposio Internacional de la Sociedad Española de Lingüística, editadas por Milka Villayandre Llamazares, León, Universidad de León, Dpto. de Filología Hispánica y Clásica, 2006. ISBN: 84-690-3383-2. Publicación electrónica en: http://www3.unileon.es/dp/dfh/SEL/actas.htm

EDICIÓN DIGITAL PARA EL ANÁLISIS LINGÜÍSTICO AUTOMÁTICO DEL CORPUS BONAPARTE[*]

ROSA MIREN PAGOLA
CARMEN ISASI
JULEN ERRASTI
PATRICIA FERNÁNDEZ
Universidad de Deusto

1. DESCRIPCIÓN DEL PROYECTO

No es necesario justificar, a día de hoy, el interés de las aplicaciones, de lo que, bajo una denominación muy común en el ámbito italiano, se puede citar como "Informática Humanística"[1]. Es bien sabido que se trata de un marco en el que, entre otras cuestiones, los problemas filológicos de la edición de textos se abordan cada vez con más frecuencia desde nuevas posibilidades como la marcación. Del mismo modo, es también evidente el engrosamiento del capítulo de trabajos basados en corpus en el ámbito mismo del español[2] y el desarrollo de nuevas implicaciones metodológicas, como puede ser la formalización gramatical.

De ambas vertientes, es decir, la nueva Filología y la Lingüística de Corpus, participa lo que hemos dado en llamar el proyecto

[*] Esta investigación se realiza con la ayuda de la Dirección de Euskara del Departamento de Cultura de la Diputación Foral de Bizkaia.

[1] Con este mismo título, por ejemplo, la miscelánea de Fiormonte (2003), y en el ámbito de la lengua española, muy recientemente la de Lucía Megías (2005).

[2] *Vid*. el capítulo sobre Lingüística de corpus de Abaitua (2002).

Bonaparte. Por un lado, en efecto, se trata de preparar una edición digitalizada de un corpus diacrónico y dialectal en lengua vasca, el corpus *Bonaparte*, por medio del lenguaje de marcación XML (*Extensible Markup Language*). Este metalenguaje estándar es el más apto para el intercambio y publicación de datos, pues permite "organizar jerárquicamente todas las unidades informativas de un documento mediante estructuras lógicas" (Abaitua *et al*. 2003:16).

Para esta primera tarea, es decir, la consecución de una edición enriquecida y navegable, partimos del supuesto de que "una edición digital con fines de consulta o estudio y aun de simple lectura debería atender, como es obvio, a los mismos requisitos filológicos de cualquier edición 'escolar', y participar, en este sentido, de muchas tomas de decisión comunes a las ediciones convencionales" (Isasi *et al*. en prensa). Además, al partir de una edición en papel, se ha contado con la ventaja de disponer ya de unos criterios y una reflexión filológica bien elaborada.

Por otra parte, esta edición constituye el soporte para el segundo componente del trabajo: el desarrollo de un sistema de etiquetado en el nivel morfológico.

Conviene subrayar que el proyecto aporta diversos aspectos innovadores: se trata de la aplicación sistemática de un proceso de edición –desde un equipo interdisciplinar– a un corpus de importancia indiscutible en la tradición documental del euskara; pero también, por añadidura, de una propuesta de análisis que implica lingüísticamente las vertientes dialectal y diacrónica de esta lengua.

2. DESCRIPCIÓN DEL CORPUS

El corpus del proyecto o *Corpus Bonaparte* recibe su denominación del fondo de manuscritos pertenecientes a la colección personal del Príncipe Luis Luciano Bonaparte. Este fondo documental vasco, el más importante desde el punto de vista diacrónico y dialectológico, que se encuentra repartido en los archivos de las Diputaciones de Bizkaia, Gipuzkoa y Navarra, fue recopilado primero por el propio L. L. Bonaparte en los viajes que realizó al País Vasco, y de los numerosos obsequios que recibió para poder llevar a cabo sus estudios lingüísticos; el corpus se incrementó

también notablemente con los encargos de traducción que él mismo solicitó a diversos colaboradores.

Uno de los principales objetivos de esta recopilación fue el trazado de un mapa lingüístico que recogiera las delimitaciones de los dialectos, subdialectos y variedades del euskara. A este propósito, y siguiendo la metodología dialectológica habitual a mediados del siglo XIX, Bonaparte se rodeó de estrechos colaboradores que le facilitaron traducciones en distintas variedades lingüísticas por él seleccionadas[3]. Los textos objeto de estas traducciones fueron fundamentalmente la *Doctrina Cristiana*, y libros y partes concretas del Antiguo y Nuevo Testamento.

Si bien es cierto que el método de la traducción de textos religiosos para el análisis dialectológico ha tenido numerosos detractores, y que tampoco Bonaparte se ha librado de importantes críticas de lingüistas vascos por haberlo seguido, no puede negarse que en realidad los resultados de su compilación fueron excelentes. En efecto, la minuciosa clasificación en la que llegó a distinguir ocho dialectos, veinticinco subdialectos y casi cincuenta variedades, mantenida sin apenas modificación hasta nuestros días, fue el resultado de la indudable uniformidad de información que le ofrecieron los mencionados materiales.

En cuanto a los manuscritos de obsequio, conviene decir que proceden de diferentes orígenes, autores, dialectos y variedades tipológicas, y que se encuentran entre ellos autores conocidos, incluso relevantes en la literatura vasca, como Eusebio M.ª de Azkue, Juan Antonio Moguel, Jean Martín Iribarren, Jean Baptiste Dasconaguerre o Joaquin Lizarraga. En otros casos, como en los de los sermones de Oñate o de Baztán, aun tratándose de autores desconocidos, el uso de la lengua es de indudable calidad.

Entre los manuscritos del Fondo Bonaparte, existen textos en verso, en prosa y de diferentes tipologías, incluida la epistolar. La mayor parte de los documentos es, como se ha dicho, de tipo

[3] Los principales colaboradores y los dialectos a los que tradujeron respectivamente son: Jose Antonio Uriarte (bizkaíno); Jose Antonio Uriarte (gipuzkoano literario) y Claudio Otaegui (gipuzkoano-Zegama); Bruno Echenique (alto-navarro-Baztan); Pedro Jose Samper (salacenco); Pedro Prudencio Hualde Mayo (roncalés); Jean Pierre Duvoisin (labortano); Iribarnegaray (bajo-navarro meridional-Baigorri); Salaberri D'Ibarrole (bajo-navarro oriental), Abbe Cazenave (bajo-navarro oriental-Garazi); Jean Baptiste Archu y Emmanuel Inchauspe (suletino).

religioso, y, entre ellos, los más numerosos corresponden a las traducciones del catecismo y del Evangelio de San Mateo. Respecto al uso lingüístico, también hay diferencias. En algunos casos es más formal y cuidado, mientras que en otros se recoge el habla espontánea y popular, pero en todos ellos se reflejan los caracteres propios del habla local. Cronológicamente, un importante número de manuscritos pertenece a la segunda mitad del siglo XVIII, si bien la parte más voluminosa data del XIX.

En todo caso, el fondo tiene un innegable interés lingüístico y dialectológico, porque permite contar con una gran riqueza de testimonios diacrónicos de todas las variedades del euskara –algunas de ellas hoy desaparecidas en la lengua hablada– y sin los cuales el conocimiento de la dialectología vasca hubiera resultado mucho más limitado. Por todo ello, hemos considerado que se trataba de un corpus idóneo para realizar una descripción diacrónica y dialectal del euskara mediante una herramienta de análisis automático.

3. OBJETIVOS DEL PROYECTO

Los objetivos del proyecto, se cifran, por una parte, en la confección de una edición digital que incluya completa descripción bibliográfica de los textos, reseña de las fuentes y ediciones anteriores, información documental de cada texto (origen, tipología, cronología, tema, dialecto, etc.), distinción de las diversas estructuras textuales (capítulos, secciones, párrafos, versículos) y datos contenidos en las notas a pie de página de la edición en papel (cambios, correcciones, adiciones…).

Por otra parte, y desde una perspectiva puramente lingüística, se pretende conseguir una herramienta de análisis automático del euskara diacrónico y dialectal que permita señalar las diferencias entre dialectos, cotejar sus principales rasgos diferenciadores y destacar las características/estructuras morfológicas de la lengua de las distintas épocas.

En función de estos objetivos, pues, se viene desarrollando una serie de tareas tanto en el plano editorial como en el de la introducción de metadatos lingüísticos. Estas tareas son las que a continuación presentamos.

4. UNA MARCACIÓN ORIENTADA AL ANÁLISIS LINGÜÍSTICO

Como ya se ha dicho, para la anotación textual se ha empleado el lenguaje de marcación XML a través del estándar de etiquetado TEI (*Text Encoding Initiative*)[4]. Esta convención internacional e interdisciplinar de anotación humanística pretende facilitar la representación de textos literarios y lingüísticos con vistas a la preservación, enseñanza e investigación en estas áreas.

En concreto, se han aplicado los capítulos dedicados a: la estructura de los documentos TEI (Capítulo 3: *Structure of the TEI Document Type Definition*), para la anotación de las distintas divisiones estructurales; la cabecera TEI (Capítulo 5: *The TEI Header*), para la inserción de los datos contenidos en la introducción de la edición en papel; y la transcripción de fuentes primarias (Capítulo 18: *Transcription of Primary Sources*), para marcas como *sic*.

Para la inclusión de los datos bibliográficos y las descripciones de los manuscritos, fuente de los textos digitalizados, se siguieron, en un primer momento, las directrices del proyecto *TEI Master* (*Manuscript Access through Standards for Electronic Records*)[5]. Los contenidos de esta propuesta han sido modificados y añadidos durante el año 2005 a la guía general de TEI en su versión P5 (Capítulo 13: *Manuscript Description*). Por ello, una de las tareas actualmente en desarrollo es la adaptación del etiquetado *Master* a la revisión de TEI P5.

Cabe destacar que se ha mantenido la lengua estándar del etiquetado, es decir, el inglés, excepto en los valores de los atributos no definidos en las directrices, que han sido traducidos al euskara como primer paso hacia la creación de una propuesta de etiquetado en lengua vasca.

A su vez, se está elaborando una taxonomía propia del corpus digital. Esta clasificación, contenida en la cabecera general del proyecto, permitirá establecer jerarquías y relaciones entre los distintos textos digitalizados atendiendo a los criterios de: dialecto, subdialecto, variedad y tipología textual. No presentamos aquí dicha taxonomía ya que se encuentra aún en fase de desarrollo.

---

[4] URL: http://www.tei-c.org

[5] URL: http://www.tei-c.org/Master/

4.1. *Las etiquetas*

El etiquetado aplicado a cada texto se divide en dos grandes bloques: la cabecera y el etiquetado estructural.

La cabecera, colocada al comienzo de cada texto, incluye toda la información bibliográfica y documental contenida en las introducciones de la edición en papel. Además, se ha confeccionado una cabecera general del proyecto, que contiene completa información bibliográfica de ediciones precedentes, una breve descripción del proyecto y la taxonomía digital ya mencionada.

Un aspecto importante que nos hemos visto obligados a resolver ha sido el de la inclusión de los distintos niveles de edición en la descripción bibliográfica de la fuente, ya que nuestra digitalización toma los textos de una edición previa en formato PDF (Pagola 2004), que procede, a su vez, de la edición en papel realizada por Rosa Miren Pagola (Pagola *et al*. 2004) de los manuscritos del Fondo Bonaparte.

El etiquetado estructural, que sigue las directrices de la TEI para la anotación de textos en prosa, varía en función de la tipología textual. Así, dentro del cuerpo (< *body* >) de cada texto (< *text* >) podemos encontrar un título del documento (< *docTitle* >); una división por capítulos (< *div type* = "*kapitulua*" *n* = "" >) acompañada del número que le corresponda y de un título (< *titlePart type* = "*kapitulua*" >); y en cada capítulo, un listado de versículos (< *list type* = "*bertsikulua*" >), cada uno de ellos numerado (< *item n* = "" >) y precedido, si fuese el caso, por un pequeño encabezado (< *head* > / < *headItem* >) que da información o recoge un resumen de su contenido. Estas etiquetas no han de aparecer obligatoriamente, y en casos como el padre nuestro pueden ser sustituidas en su conjunto por un único párrafo (< p >), como en el caso del *Aita Gurea*.

En esquema, el etiquetado estructural es el siguiente:

```
<text>
        <body>
                <div type="kapitulua" n="" > // <p></p>
                        <titlePart type="kapitulua" ></titlePart>
                        <head> / <headItem>
                        <list type="bertsikulua" >
                            <item n="" >
                            …
                        </list>
                </div>
                …
        </body>
</text>
```

Otras etiquetas utilizadas en la anotación de los textos son:

- *<note id="" ></note>*: para la inserción de las notas a pie de página de la edición en papel. Esta etiqueta aparece adosada a la palabra de la cual se realiza la anotación, en el mismo lugar que en la edición de la que procede. El identificador indica el número de nota del documento;
- *<sic></sic>*: en sustitución de la marca (*sic*);
- *<hi rend="lodi" ></hi>*: para indicar que el pasaje etiquetado aparece resaltado en el manuscrito original.

4.2. *Textos etiquetados*

Han sido treinta los textos anotados –siguiendo el etiquetado mencionado– durante la primera fase del proyecto. Recogemos aquí una relación de los mismos y los dialectos, variedades y subvariedades en los que se encuentran:

- *Salomonen Kanten Kanta* (Cantar de los Cantares de Salomón), en *Bizkaiera ekialdekoa*, *Bizkaiera orokorra*, *Ekialdeko behe-nafarrera*, *Gipuzkera hegoaldekoa*, *Gipuzkera orokorra*, *Iparraldeko goi-nafarrera*, *Mendebaldeko behe-nafarrera*, *Zuberera*.
- *Aita Gurea* (padre nuestro), en *Aezkera*, *Ekialdeko behe-nafarrera (Garazi Amikutze)*, *Gipuzkera hegoaldekoa (Zegama)*, *Gipuzkera hegoaldekoa*, *Gipuzkera iparraldekoa*, *Gipuzkera nafarrokoa*, *Iparraldeko goi-nafarrera (Ultzamakoa)*, *Lapurtera orokorra*,

Mendebaldeko behe-nafarrera (Baigorrikoa), *Mendebaldeko behe-nafarrera (Lapurdikoa)*, *Zaraitzera (Eiaurreta)*, *Zuberera*.

- *San Juan Apostoluaren Apokalipsia* (Apocalipsis de San Juan), en *Bizkaiera ekialdekoa*, *Bizkaiera orokorra*, *Ekialdeko behe-nafarrera*, *Gipuzkera orokorra* (2 textos), *Iparraldeko goi-nafarrera*, *Lapurtera orokorra* (2 textos), *Mendebaldeko behe-nafarrera*, *Zuberera*.

4.3. *Del etiquetado estructural al lingüístico*

No podemos olvidar que el objetivo principal de esta edición es la preparación textual del corpus para su ulterior procesamiento lingüístico automático. Por tanto, con la anotación descrita se ha perseguido, por una parte, poder contextualizar las búsquedas definidas en función del etiquetado lingüístico; y por otra, establecer relaciones entre segmentos de los distintos textos que permitan cotejar y facilitar el estudio de la lengua, tanto en su vertiente diacrónica como en la dialectal.

5. La anotación lingüística

El proceso mediante el cual se han ido estableciendo las características lingüísticas más revelantes en el nivel morfológico para la descripción de los diversos dialectos del euskara, ha sido laborioso y complejo, e indudablemente, hay que decir que el proyecto todavía se encuentra en una fase intermedia. Este proceso ha sido el fundamento para las especificaciones computacionales definidas para la automatización del análisis lingüístico y ha requerido la resolución de problemas y toma de decisiones en diversos niveles.

5.1. *Tratamiento del nivel gráfico fonético*

En primer lugar, fue necesario realizar un exhaustivo análisis de las grafías, especialmente complejo ya que los materiales corresponden al habla de las diversas variedades del euskara, e indirectamente, por tanto, a una gran gama polimórfica en la

articulación de los sonidos. Se suma a esta dificultad la falta de normalización ortográfica[6] que dificulta la interpretación fonológica a través de ese polimorfismo. En esta intersección de lo gráfico y lo fonético, constituyen también aspecto de dificultad específica en este corpus los casos de reducciones de los segmentos vocálicos: síncopas, aféresis, apócopes; y de reducciones o cambios de las secuencias: monoptongaciones, armonías, etc., que llegan a alcanzar un volumen inconmensurable.

Por todo ello, en una primera etapa se recogieron todas las variantes gráficas de cada término, aunque sin proceder al análisis de las evoluciones fonéticas[7], con la intención de fijar una base sólida para la anotación de cuestiones morfológicas, nivel en el que se asienta la parte central del proyecto.

5.2. *Análisis morfológico*

5.2.1. *Cuestiones previas*

El análisis morfológico pretende alcanzar la automatización en el proceso de etiquetado de los fenómenos pertinentes atendiendo a su valor y función concreta.

Se ha tomado en este estadio del proyecto como documento base para el análisis un texto que tiene la máxima representación en todos los dialectos: el *Cantar de los Cantares* del Antiguo Testamento, con 7 versiones en otras tantas variedades del euskara.

5.2.2. *El procedimiento*

La metodología seguida en cuanto al tratamiento y extracción de datos morfológicos ha sido:

---

[6] Esta heterogeneidad deriva, desde luego, de la dispersión cronológica del corpus, pero también de la presencia de distintas tradiciones ortográficas. En efecto, tendríamos los sistemas del castellano, fundamentalmente, en la parte correspondiente a Hegoalde, y del francés y gascón para los textos de Iparralde.

[7] Es decir, el etiquetado reconocerá que "naiz" es una determinada forma verbal que se realiza en unos determinados dialectos y que, además, tiene otras variantes con su correspondiente distribución dialectal "naz", "niz" e, incluso, "naz" o "nax".

1) Definición de una serie de *modelos de conjugación* para las distintas categorías gramaticales, discriminando diferentes modelos para cada subdialecto en casos excepcionales.
2) Lematización manual de los términos que aparecen en cada uno de los textos y posterior asociación a cada uno de los modelos. De este modo, se obtienen todas las formas conjugadas de cada uno de los lexemas, clasificadas y con el resto de información añadida: número, persona, caso. Se ha seguido un proceso paralelo en las formas verbales, atendiendo en este caso a cuestiones de modo, tiempo, número, persona, registro y género (únicamente dentro del registro coloquial).
3) Elaboración de esquemas morfológicos de los lemas-modelo a partir de la recogida de testimonios textuales. Se ha empleado este método y no el de la reconstrucción, ya que si bien este último completaría todos los paradigmas, obligaría a trabajar sobre la analogía y, por tanto, restaría fiabilidad a las formas.
4) El etiquetado lingüístico se ha ido configurando en dos fases. En la primera, se ha confeccionado un listado básico de atributos; mientras que la segunda ha supuesto la distribución de estos atributos por categorías.
5) Para la anotación hemos desarrollado nuestro propio repertorio de etiquetas manteniendo la sintaxis de XML.

5.2.3. *Restricciones en este nivel*

En esta primera fase, se ha pospuesto todo aquello que representa análisis sintáctico, incluidas las conjunciones (*juntagailuak*) y los conectores (*lokailuak*) en su función. Igualmente, se ha dejado para un momento posterior, la interpretación semántica que requiere el análisis de algunas formas morfológicas; y también aspectos como las posposiciones, interjecciones, etc., que serán tenidas en cuenta más adelante. Actualmente, se está procediendo a la comprobación y validación de todos los textos y al estudio de los problemas de desambiguación.

5.2.4. *Elementos analizados*

Por todo lo expuesto, los atributos y categorías analizadas hasta el momento son las que se señalan a continuación.

El *sustantivo* con la distinción de número singular, plural, indefinido y plural cercano. Respecto a la declinación, se recogen todos los casos posibles de aparición: absolutivo, ergativo, dativo, genitivo de posesión, partitivo; y todos los de lugar y modo: ablativo, inesivo, alativo, alativo de dirección, alativo de conclusión, destinativo, causativo, sociativo, instrumental, y prolativo. Igualmente se recoge la distinción entre sustantivos animados e inanimados.

El *adjetivo* también con la distinción de número singular y plural, así como el indefinido y el plural cercano y todos los casos que ya se han señalado para el *sustantivo*.

Respecto al *adverbio*, se han atendido las dos posibilidades que ofrecen, recogiendo tanto los que aceptan morfemas de la declinación como los no declinables. Para los declinables, en lo que se refiere al número y la declinación, se ha seguido idéntica clasificación que para *sustantivos* y *adjetivos*. En los adverbios que no se declinan no se ha hecho ninguna subclasificación.

Se han distinguido los diferentes tipos de *pronombres*: personales, indefinidos, reflexivos y recíprocos. Atendiendo, también, a aquellos que pueden funcionar como determinantes[8].

En cuanto a los *determinantes*, la categorización comprende el artículo y los demostrativos en sus tres grados: primero (*hau*), segundo (*hori*) y tercer grado (*hura*).

Las *conjunciones* se han recogido y clasificado dentro de una única categoría, sin hacer distinciones entre ellas, ya que el hacerlas conllevaría entrar en el nivel de la sintaxis. También los *conectores* y las *posposiciones* se han recogido y clasificado en sendas categorías, sin proceder, por el momento, a otras subdivisiones. Finalmente, se han recogido bajo una única categoría las formas que no han tenido ubicación apropiada entre las citadas.

En lo que concierne al *verbo*, se han distinguido las formas *no-personales* y las *personales*. En las primeras, la clasificación y análisis comprende el *participio*, la *forma nominal* y la *raíz verbal*. En el *participio* la subdivisión categorial se ha hecho discriminando el *primer participio* (*etorri*), el *segundo* (*etortzen*) y el *tercero* (*etorriko*). En las *formas personales* se ha analizado el *modo* en sus

[8] Los *pronombres relativos* no son objeto de análisis en esta primera fase porque comprenden el nivel sintáctico.

diferentes clases: indicativo, condicional, condicional consecuente, potencial, subjuntivo e imperativo; el *tiempo*: presente, pasado, hipotético y futuro; la *persona*: marcas de concordancia personal con el sintagma nominal en caso absolutivo (*Nor*), ergativo (*Nork*) y dativo (*Nori*); el *tratamiento*: el neutro, por un lado, y el alocutivo (*hika*) con la distinción de género, masculino y femenino, y las formas alocutivas respetuosas (*zuka* y *xuka*).

6. FUTUROS TRABAJOS

Las tareas fijadas por ahora para la continuación del proyecto se centran en completar el análisis morfológico[9], iniciar el estudio del nivel sintáctico, dar solución a los problemas que han ido surgiendo y tener en cuenta los subdialectos y variedades que no han sido atendidos por el momento. Son estas tareas las que actualmente ocupan la labor investigadora de los integrantes del grupo TesiTek de la Universidad de Deusto.

REFERENCIAS BIBLIOGRÁFICAS

ABAITUA, J. (2002): "Tratamiento de corpora bilingües", en M. A. Martí y J. Llisterri (eds.), *Tratamiento del lenguaje natural*, Barcelona: Edicions Universitat de Barcelona, 61-90.

ABAITUA, J. *et al*. (Julio-Septiembre 2003): "Contenidos y metacontenidos en la edición digital", *Letras de Deusto*, 100, Bilbao: Universidad de Deusto, 16, 11-52.

FIORMONTE, D. (2003): *Informatica umanistica*. *Dalla recerca all insegnamiento*, Roma: Bulzoni Editore.

ISASI, C.; FERNÁNDEZ, P. y PÉREZ ISASI, S. (en prensa): "Philological Issues Regarding a Multiple Plurilingual Digital Edition", *Variants*.

LUCÍA MEGÍAS, J. M. (2005): *Manual de informática humanística*, Madrid: Castalia.

[9] En este nivel aún quedan aspectos por tratar como la sobredeclinación, la onomástica, las lexías complejas o la subclasificación de conjunciones, conectores y posposiciones.

PAGOLA, R. M. (ed.) (2004): *Bonaparte Ondareko Eskuizkribuak. Bilduma osoaren Edizio Digitala*, Bilbo: Deustuko Unibertsitatea.

PAGOLA, R. M. *et al*. (eds.) (1992-1999): *Bonaparte Ondareko eskuizkribuak*, Bilbo: Deustuko Unibertsitatea.