UN CORPUS DE LENGUA HABLADA: EL CASO DE FONATARI

*Alexander Iribar Ibabe*Laboratorio de Fonética
Universidad de Deusto

1. INTRODUCCIÓN: LOS CORPUS ORALES

El concepto de corpus lingüístico resulta ciertamente heterogéneo. Un corpus puede adoptar muchas formas, y muchas colecciones de datos lingüísticos pueden ser consideradas como un corpus. Sin pretender en modo alguno examinar exhaustivamente esta cuestión, conviene, para centrar debidamente el objeto de estas líneas, mostrar someramente la mencionada heterogeneidad del concepto de corpus. Para ello, seguiremos una de sus muchas clasificaciones posibles: la propuesta por Torruella & Llisterri (1999). Según estos autores, un corpus, ya sea textual u oral, puede clasificarse según distintos criterios:

- Según el porcentaje y la distribución de los distintos tipos de texto¹:
 1. Corpus grande: no pone límite al número de sus registros, que en cualquier caso ha de ser muy elevado, para evitar problemas de representatividad, equilibrio, etc.
 2. Corpus equilibrado: contiene distintas variedades de textos, en parecidas proporciones.
 3. Corpus piramidal: organizado en estratos, cada vez con más variedades pero con menos textos de cada una de ellas.
 4. Corpus monitor: contiene un volumen textual menor, pero renovado periódicamente.
 5. Corpus paralelo: colección de textos traducidos a una o varias lenguas.
 6. Corpus comparables: selección de textos parecidos en varias lenguas (no son traducción directa).

1. Debe entenderse que los textos pueden ser, en principio, escritos u orales.

7. Corpus multilingües: textos de diferentes lenguas, sin ningún criterio específico de selección.
 8. Corpus oportunista: recoge los textos disponibles.
- Según la especificidad de los textos:
 1. Corpus general: pretende reflejar la lengua común en su sentido más amplio.
 2. Corpus especializado: recoge textos de una determinada variedad o tipo particular de lengua.
 3. Corpus genérico: recoge un único tipo o género de textos.
 4. Corpus canónico: recoge todos los textos que componen la obra de un autor.
 5. Corpus periódico o cronológico: recoge textos de años o épocas concretas.
 6. Corpus diacrónico: recoge textos de diferentes etapas o épocas, para analizar posibles evoluciones en la lengua.
 - Según la cantidad de texto que se recoge en cada documento:
 1. Corpus textual (*whole texts corpus*): los textos son íntegros.
 2. Corpus de referencia (*reference corpus*): formado por fragmentos de los textos.
 3. Corpus léxico (*samples corpus*): fragmentos muy pequeños y de longitud constante, de interés léxico.
 - Según la codificación y la anotación:
 1. Corpus simple: no codificado ni anotado.
 2. Corpus codificado o anotado: los textos van acompañados de etiquetas de distinto tipo.
 - Según la información que acompaña a los textos:
 1. Corpus documentado: cada documento posee su propia cabecera de descripción (*header*) o un archivo DTD (*document type definition*) asociado.
 2. Corpus no documentado.

Más centrados en la tipología de los corpus orales, Torruella & Llisterri (1999) distinguen tres grandes tipos, según su finalidad:

1. Los orientados a la descripción fonética de la lengua.
2. Los que se utilizan para el desarrollo de sistemas en el ámbito de las tecnologías del habla.

3. Los propiamente denominados corpus orales, consistentes en transcripciones de lengua hablada.

Los corpus orales desarrollados para el campo de las tecnologías del habla son, tal vez, los más controlados desde el punto de vista técnico, y pueden estar diseñados para muy diversas aplicaciones. Por ejemplo, Listerri et al. (2005) distinguen hasta ocho posibles finalidades de este tipo de corpus oral:

1. Corpus generales para el desarrollo de aplicaciones en tecnologías del habla.
2. Corpus para el desarrollo de sistemas de conversión texto-voz.
3. Corpus para el desarrollo y evaluación de sistemas de reconocimiento de habla.
4. Corpus para la identificación automática del locutor.
5. Corpus para la identificación automática de la lengua.
6. Corpus para el desarrollo de sistemas de diálogo.
7. Corpus para la traducción automática del habla.
8. Corpus para la recuperación de información.

Frente a este tipo de corpus del ámbito de la tecnología del habla, cuya concepción teórico-metodológica les otorga unas características algo más definidas, los corpus orales de habla son un amplio terreno de aplicación que entronca en realidad con los trabajos de campo de la dialectología clásica e incorpora reflexiones de otras muchas disciplinas, como la fonética experimental, la propia tecnología del habla, la sociolingüística, la etnolingüística y la lingüística de corpus.

Un ejemplo de lo anterior está en la evolución experimentada por los tradicionales atlas lingüísticos, hoy convertidos en atlas multimedia.² Detrás de los primeros estaba una profusa colección de transcripciones de habla (conocidas como corpus de lengua oral o *spoken language corpora*); detrás de los segundos están ahora los corpus orales (*speech corpora*, aunque este término se emplee a veces únicamente para referirse a los corpus de aplicación tecnológica), con muy diversas concreciones metodológicas.³

2. J.A. González Salgado proporciona una lista de los atlas lingüísticos accesibles en internet en la dirección <http://www.geolectos.com/atlas.htm>.

3. La distinción entre los corpus de lengua oral (*spoken language corpora*), que coleccionan transcripciones ortográficas de habla real, y los corpus orales (*speech corpora* o *speech databases*), que consisten en la propia señal sonora, se ha realizado desde el punto de vista de los primeros especialistas en tecnología del habla que crearon recursos orales específicos para ese campo, con el afán de distinguirlos de aquellos otros creados desde una perspectiva netamen-

En definitiva, el concepto de corpus oral es heterogéneo, puesto que confluyen en él múltiples disciplinas, con sus concepciones teórico-metodológicas y sus intereses particulares. Por tanto, los corpus orales existentes presentan una gran variedad. En este trabajo nos centraremos en los corpus orales existentes en euskara, con especial atención al ofrecido en el portal de fonética vasca FonAtari.

2. LOS CORPUS ORALES DEL EUSKARA

Existe un buen número de colecciones de habla en euskara. En este apartado efectuaremos un repaso somero, comenzando por una serie de trabajos que pueden considerarse como precedentes de los corpus más modernos.

2.1. *Precedentes: las primeras colecciones de euskara*

Por una serie de razones que no es el momento de tratar (la reforma ilustrada, el apologismo vasco, el gusto por lo rústico del rococó, el romanticismo, el nacionalismo), los vascos han valorado desde antiguo (dos siglos largos, al menos) su patrimonio etnográfico; así, las danzas, los cantos, las tradiciones, han sido valoradas —de hecho, sobrevaloradas no pocas veces⁴—, recogidas y estudiadas.⁵ De todos los elementos patrimoniales, el más valorado ha sido, desde siempre, la propia lengua.⁶

te lingüística. Sin embargo, una división tajante entre ambos enfoques nunca ha sido fácil de trazar, puesto que, por una parte, resultó habitual prácticamente desde el primer momento que los corpus con una finalidad tecnológica incorporaran también algún tipo de transcripción del material sonoro, y, por otra, los corpus de lengua oral han pretendido desde hace décadas facilitar el acceso al material grabado original. En definitiva, la diferencia entre los dos productos nunca ha sido fácil de definir, y el desarrollo tecnológico de los últimos años vuelve cada vez más imprecisa esa línea divisoria.

4. En palabras de Ansorena (1993:11): “Gure herri izaera, guru kultura kinka larrian bizi zenez, askoz indartsuago zen batek inguratzen zuelako, defendatzeko modurik onena apologiarena okurritu zitzaigun aspaldi euskaldunoi. Eta benetan nolakoak garen aztertu beharrean, zeinen onak garen frogatzera zuzendu genituen geure ahaleginak” (como nuestra cultura vivía un momento crítico, rodeada por otra mucho más fuerte, el mejor modo de defenderla que se nos ocurrió a los euskaldunes fue el apologismo. Y en lugar de estudiar cómo somos realmente, dedicamos nuestros esfuerzos a probar qué excelentes somos).
5. En ocasiones, esos materiales etnográficos, que podían además haber sido objeto de manipulaciones, reconstrucciones y *purificaciones*, eran —y siguen siendo— devueltos a la población, en un constante intento de *revasquizarla*.
6. No en vano el vasco ha denominado a su propia tierra como *Euskal Herria*, esto es: ‘el pueblo del euskara’.

Este interés ofrece unos primeros resultados especialmente valiosos en el campo de la etnomusicología. Se trata de una serie de Cancioneros o Colecciones Musicales, cuyo interés primordial es, obviamente, el musical, pero que también presentan los textos de las canciones que se recogen. Y esas canciones han sido siempre recogidas por los distintos autores mediante encuestas a informantes, es decir, se trata de expresiones de habla real (aunque sometidas en ocasiones a correcciones y estilizaciones). Por tanto, pueden considerarse como las primeras recogidas ordenadas de materiales orales en euskara.⁷

El primero de esta lista de estudiosos que recogieron muestras etnográficas en distintas zonas del País Vasco es el alemán Wilhelm von Humboldt (1767-1835), quien lo visitó en 1799 y 1801. No interesa tratar aquí la teoría humboldtiana del vascoiberismo, sino recordar los papeles musicales que reunió, de gran interés musicológico, y que contienen varias canciones cuyos textos no han sido tratados por el momento desde un punto de vista lingüístico.⁸

Aproximadamente un cuarto de siglo después, en 1826, se publica la *Colección Musical* de Iztueta-Albéniz. Juan Ignacio Iztueta, pionero de los estudios folklóricos en Europa, había publicado dos años antes su famoso *Gipuzkoako Dantza Gogoangarriak* (Danzas memorables de Gipuzkoa), donde describía minuciosamente las danzas tradicionales que se bailaban en su provincia. Dos años después, con la ayuda del músico Pedro Albéniz, publica en esta Colección las músicas de dichas danzas, con sus letras correspondientes.

En 1870, el zuberotarra J.D.J. Sallaberry publica en Bayona sus *Chants populaires du Pays Basque*, una colección de 50 romances y canciones en dialectos suletino, bajo-navarro y labortano, a veces con variantes en más de un dialecto, y siempre traducidos al francés. En la introducción (*Quelques observations indispensables au sujet de l'orthographe employée dans ce recueil*), Sallaberry ofrece algunas de las primeras informaciones sobre la fonética vasca.

Ya en el siglo XX, la Diputación de Bizkaia impulsó mediante un concurso la confección de los dos Cancioneros más importantes del País Vasco: el

7. En realidad, siguiendo una estricta línea cronológica, no todos los Cancioneros son las primeras muestras de material oral. Los trabajos de Louis Lucien Bonaparte, por ejemplo, son anteriores en varias décadas a los Cancioneros más importantes. De todo ello se tratará más adelante.

8. Entre la *Collectanea Linguistica* de Humboldt se encuentra el trabajo "La música en el País Vasco", localizado a principios del siglo XX en la *Preussische Staatsbibliothek* (biblioteca del Estado de Prusia), que presenta varias canciones recogidas supuestamente por el autor. Los datos pueden encontrarse en RIEZU (1971).

del músico y lingüista Resurrección María de Azkue, con 1.989 canciones, y el más cuidadoso del músico franciscano José Gonzalo Zulaika, Aita Donostia, que reunía 499 canciones.

En realidad, ambos musicólogos dedicaron décadas a la recopilación y al análisis de las canciones populares vascas (continuaron de hecho trabajando hasta el fin de sus días), y publicaron un buen número de colecciones diversas, con miles de piezas recogidas. La obra de Azkue es más monumental, mientras que la de Aita Donostia (continuada tras su muerte por su colaborador, el P. Riezu) es más rigurosa desde el punto de vista musicológico.

En resumen, la labor de recogida etnográfica plasmada en los distintos Cancioneros,⁹ además de su evidente interés musicológico, nos ha legado una gran cantidad de material lingüístico perteneciente a todos los dialectos de la lengua, que puede en cierto modo considerarse como la primera gran colección o corpus textual del euskara.

Dejando ya de lado los trabajos musicológicos, el gran nombre para el euskara en el siglo XIX es el del francés Louis Lucien Bonaparte (1813-1891), el primer gran estudioso de la dialectología vasca. Desde 1856 hasta 1869, Bonaparte realizó cinco viajes al País Vasco para recoger material lingüístico de primera mano. Pero, sobre todo, organizó una tupida red de colaboradores en todo el País, a quienes encargó traducciones del Catecismo y versiones de los Evangelios.¹⁰ Fruto de este trabajo, publicó numerosos estudios sobre el euskara, entre los que destacan las sucesivas propuestas de clasificación dialectal, la cuarta de las cuales, de 1869 (*Cartes des Sept Provinces Basques montrant la délimitation actuelle de l'Euskara et sa division en dialectes, sous-dialectes et variétés*), ha sido comúnmente aceptada hasta hoy en día.

El ingente material lingüístico recogido por L.L. Bonaparte ha permanecido inédito hasta los años finales del siglo pasado, cuando Rosa Miren Pagola afrontó en la Universidad de Deusto la tarea de completar su edición crítica. Después de años de trabajo en equipo, la colección de *Bonaparte ondareko eskuizkribuak* ('manuscritos del fondo Bonaparte') culminó en 26 volúmenes, publicados desde 1992 hasta 1999, con más de 12.000 páginas.

En 2004, los 26 volúmenes se publicaron en formato digital (como colección de PDF). Actualmente, Rosa Miren Pagola desarrolla, también en la Universidad de Deusto, el proyecto *Bonaparte sarean*, que pretende publicar en internet una versión completa del fondo Bonaparte, en formato XML y si-

9. La lista de los Cancioneros Vascos es mucho más amplia. Se han reseñado únicamente los primeros y los más importantes.

10. Como curiosidad, cabe recordar que Bonaparte encargó también una versión al gallego del Evangelio de San Mateo, que publicó en 1871 en Londres bajo el título de "Observaciones comparativas sobre la pronunciación gallega, asturiana, castellana y portuguesa".

guiendo el estándar de etiquetado TEI P5. De esta manera, un corpus lingüístico recogido hace 150 años se convertirá en un modelo de aplicación tecnológica.

2.2. Los primeros registros orales

Los primeros registros orales de euskara fueron realizados en París durante la Exposición Universal de 1900. Las grabaciones, correspondientes a los dialectos suletino y guipuzcoano, consistieron en cinco canciones y tres lecturas de ciertos textos de los Evangelios de San Juan y San Lucas. El material ha sido remasterizado y editado por Eusko Ikaskuntza en 2000.¹¹ Su calidad, de todas maneras, no es demasiado alta.

En 1913, el etnógrafo y lingüista alemán Rudolf Trebitsch (1876-1918) realizó grabaciones gramofónicas en varias localidades vascas para la colección de la Fonoteca de Viena. En realidad, Trebitsch tenía dos intereses, el lingüístico y el musical, de modo que obtuvo dos tipos de material: grabaciones lingüísticas y musicales.

Las grabaciones lingüísticas¹² de Trebitsch ocupan 64 discos o cilindros, duran unos 130 minutos y contienen ejemplos de todos los dialectos, incluido el roncalés. Aparece una gran variedad de textos orales: cuentos, historias, anécdotas, cantos.¹³

El gran musicólogo americano Alan Lomax (1915-2002) realizó, durante los años 1952 y 1953, numerosas grabaciones por la geografía española, que reunió bajo el título de *Spanish Recordings* (1953). Buena parte de este material ha sido reeditado en formato CD en los últimos años, y el resto espera su pronta publicación.¹⁴ Los materiales correspondientes al País Vasco ocupan dos discos compactos y contienen 71 grabaciones, instrumentales y cantadas, recogidas en Bizkaia, Gipuzkoa y Nafarroa.

11. Puede leerse una reseña sobre este material en Mazusta (2001).

12. Las grabaciones musicales ocupan 14 discos o cilindros, duran un total de 32 minutos y contienen grabaciones vocales e instrumentales. Su contenido se presenta minuciosamente en López de Arana (2004).

13. Todo este material ha sido cuidadosamente examinado en una serie de once artículos por José María Etxebarria, quien, en 1984 y junto a Pierre Xarriton, copió el material en la Fonoteca de Viena y lo depositó en la Biblioteca de Euskaltzaindia, la Academia de la Lengua Vasca. La serie completa de artículos se reseña en la bibliografía final.

14. Los discos editados son: Aragón and València; The Basque Country: Bizcay and Guipuzcoa; The Basque Country: Navarre; Extremadura; Galicia; Ibiza and Formentera; The Pityusic Islands; Mallorca: The Balearic Island. Próximamente serán editados los volúmenes siguientes: La Mancha; Mallorca; Castilla-León: The Maragatos; Andalucía: Granada & Sevilla; Andalucía: Jerez & Tarifa; Murcia; Cantabria; Asturias.

Además de estas tres colecciones reseñadas, existe un material extraordinariamente heterogéneo y notablemente disperso de grabaciones radiofónicas y gramofónicas, que esperan una labor de índole archivística, en primer lugar. Algunos materiales de estas características pueden encontrarse, por ejemplo, en Eresbil (Archivo Vasco de la Música).

A partir de los años sesenta y setenta, las grabaciones aumentan vertiginosamente. Un gran número de estudiosos de la cultura vasca, euskalzales (amantes del euskara) y estudiosos de la lengua se lanzan a grabar cintas de *cassette* con distintos materiales, según los intereses de cada caso particular. El resultado de esta actividad de varias décadas es una ingente cantidad de material recogido, cuya explotación —al menos lingüística— es escasísima, y cuya accesibilidad es muy limitada.

2.3. Colecciones modernas

De las diversas colecciones modernas o corpus orales de euskara, destacaremos los más interesantes:

EIBARTARREN AHOTAN

El archivo oral *Eibartarren Ahotan* ('en boca de los eibarreses') es un proyecto surgido en 1999 en el ayuntamiento de Eibar (Gipuzkoa). Consiste en entrevistar a personas mayores del municipio, recoger en vídeo historias, anécdotas y sucesos de la primera mitad del siglo XX, y divulgar en internet toda esa información recogida oralmente en el habla eibarresa.

Actualmente, la web de *Eibartarren Ahotan*¹⁵ contiene 21 grabaciones de otros tantos informantes, ordenadas por temas. Cada grabación viene acompañada de su transcripción ortográfica.

HABE-IKASBIL

HABE es un Centro oficial del Gobierno Vasco dedicado básicamente a la euskaldunización de adultos. En su completo portal en internet, posee una sección denominada Ikasbil,¹⁶ en la que ofrece una gran cantidad de recursos educativos, muchos de ellos con el correspondiente material de audio o de vídeo.

Los componentes principales son los siguientes:

- *Dokuteka* ('centro de documentación') – *Entzungaiak* ('audios'). Más de 2.500 textos orales, de entre dos y seis minutos de duración, toma-

15. URL: http://www.egoibarra.com/Egoibarra/Ikerketak/Eibartarren_Ahotan.

16. URL: <http://www.ikasbil.net/jetspeed/>.

dos de los medios de comunicación (Radio Euskadi, preferentemente), y ordenados por niveles de dificultad. No están transliterados.

- *Dokuteka – Bideoak*. Más de 2.000 vídeos, tomados también de la televisión vasca, con las mismas características que los audios anteriores.
- *Aisialdirako* ('para el tiempo libre') – *Irrati nobelak*. Cinco telenovelas, ordenadas por capítulos o escenas. Cada uno de ellos contiene el texto completo y la versión de audio correspondiente.
- *Aisialdirako – Kantu-kantari*. Más de 90 canciones tradicionales, con letra, música (formatos wav y midi) y partitura (formato pdf).
- *Aisialdirako – Txistekak*. Más de 200 chistes, con texto y audio. El audio corresponde a grabaciones enviadas por los propios usuarios del portal.
- *Aisialdirako – Ipuinak*. Dieciséis cuentos, con su texto y su audio correspondiente, leídos por locutores profesionales.

OHIKO EUSKAL MINTZAMOLDEEN ANTOLOGIA

La *Antología del Habla Usual Vasca* ha sido realizada por Euskaltzaindia, la Academia de la Lengua Vasca, con parte de los materiales grabados para la elaboración del Atlas Lingüístico del País Vasco. Consiste en una colección de 145 grabaciones realizadas en otros tantos pueblos de la geografía del euskara, entre los años 1986 y 1992. La Antología presenta un extracto de cada grabación¹⁷ de aproximadamente minuto y medio, acompañado de su transliteración.

La Antología se presenta en formato CD, pero también está accesible en internet.¹⁸

BIZKAIERAREN FONOTEKEA

Bizkaieraren Fonotekea es un sólido y completo corpus oral sobre el dialecto vizcaino del euskara, creado por el grupo de investigación Aholab.¹⁹ Nació primeramente como corpus oral, aunque desde hace algún tiempo incorpora también registros de vídeo.²⁰

17 Las grabaciones completas, que pueden consultarse en la Bilioteca de Euskaltzaindia, suman más de 4.000 horas.

18 URL: <ftp://www2.euskaltzaindia.net/Antologia/Antologia.pdf>.

19 URL: <http://aholab.ehu.es/aholab/>. Aholab es un grupo de investigación ubicado en Laboratorio de Procesado de Señal del Departamento de Electrónica y Telecomunicaciones de la Facultad de Ingeniería de Bilbao, perteneciente a la Universidad del País Vasco. También pertenece al grupo el lingüista Iñaki Gaminde, autor de la inmensa mayoría de las grabaciones.

20. Puede encontrarse más información sobre este corpus en Hernáez (2002) y Sánchez et al. (2001, 2002).

El corpus está accesible en internet,²¹ pero también puede adquirirse en ELRA²² recogido en cinco CD's. Contiene más de 11.000 palabras, casi 2.000 frases y 1.000 textos. La base de datos contiene los archivos de audio o vídeo, más una colección de archivos SGML con la descripción técnica, las marcas de tiempo para cada nivel de marcado y la transcripción fonética, de acuerdo con el estándar TEI P4.

3. LOS CORPUS ORALES DEL PORTAL FONATARI

El portal de fonética vasca *FonAtari: Euskal Fonetikaren Ataria*²³ pone a disposición del usuario una colección de corpus orales del euskara, dentro de una oferta más amplia de información y servicios acerca de la lengua vasca.

3.1. Descripción general del portal de FonAtari

FonAtari nació con la pretensión de constituirse en el punto obligado de referencia para todo el que desee alguna información, del tipo que sea, sobre el nivel fónico del euskara. Para ello, trata de reunir un buen número de componentes, materiales y servicios, en un entorno multimedia diseñado para facilitar el crecimiento futuro del *sitio* con nuevos materiales y servicios.

Por tanto, ni el diseño del Portal ni la lista de los materiales que contiene pueden considerarse como definitivos, puesto que deben estar sujetos a cambios y actualizaciones periódicas y constantes.

Toda la información contenida en FonAtari está estructurada en varias áreas generales. En cada una de ellas coexisten tres tipos de materiales:

- Materiales de creación propia del Laboratorio de Fonética de la Universidad de Deusto, alojados en su propio servidor.
- Materiales creados por otras personas, pero alojados en el servidor del Laboratorio de Fonética.

21. URL: <http://bizkaifon.ehu.es/>.

22. European Language Resources Association. URL: http://catalog.elra.info/product_info.php?products_id=6&osCsid=68879de413546e0054a1bada09d46bcb.

23. El portal *FonAtari: euskal fonetikaren ataria* es un proyecto llevado a cabo en el Laboratorio de Fonética de la Universidad de Deusto por el equipo de investigación *Fonetiker*, compuesto por Ana Elejabaitia, Rosa Miren Pagola (la directora del Proyecto *FonAtari*) y Alexander Iribar. Para su desarrollo operativo, ha contado con la financiación del Departamento de Cultura de la Excma. Diputación Provincial de Bizkaia.

- Enlaces a otras páginas pertinentes, creadas por otras personas y residentes en otros servidores.

En cuanto a la lengua de comunicación, FonAtari es un portal tetralingüe. El idioma principal es el euskara, pero, por su carácter divulgativo, está también disponible en inglés y en las dos lenguas en contacto con el euskara, esto es, español y francés. En cualquier caso, la lengua de cada material concreto suministrado por FonAtari será la original de su correspondiente autor.

Actualmente, FonAtari consta de seis secciones principales:

1. Euskara
2. Euskararen Hotsak (sonidos del euskara)
3. Dokumentu basea (base documental)
4. Ahoskerak (corpus orales)
5. Kideak (socios)
6. Berriak (noticias)

Vamos a examinar someramente cada una de estas secciones:

SECCIÓN 1: INFORMACIÓN GENERAL DEL EUSKARA

En esta sección se ofrece información general acerca del euskara, cuya función primordial es proporcionar un marco adecuado para el resto de los materiales del Portal, más centrados en el nivel fónico de la lengua. Así, se ofrece información sobre al menos las siguientes cuestiones:

- La situación jurídica y social del euskara
- Breve historia externa del idioma
- Los distintos intentos de clasificación del euskara
- La fragmentación dialectal del euskara
- Principales características lingüísticas (morfológicas, sintácticas, etc.)
- Principales características fonéticas y fonológicas

Estas cuestiones ya se tratan parcialmente, con más o menos acierto, en algunas páginas existentes en la red. FonAtari ofrece un guión principal de elaboración propia, más una breve bibliografía, en el que se incardinan los enlaces y comentarios oportunos.

Por otra parte, se ha encargado la elaboración de algunos materiales específicos sobre algunos de estos subapartados a reconocidos expertos en las materias correspondientes, como un modo de aumentar tanto el nivel científico como el prestigio del Portal.

SECCIÓN 2: ANÁLISIS FONÉTICO-FONOLÓGICO DE LOS SONIDOS DEL EUSKARA

Esta sección, desarrollada específicamente en el Laboratorio de Fonética, presenta una descripción minuciosa de los sonidos de todos los dialectos vascos, tanto desde el punto de vista acústico como articulatorio.

Los análisis se basan en una colección de muestras de habla de todos los dialectos vascos, que permite la aparición y el consiguiente análisis de todos los sonidos propios de la lengua.²⁴

Mediante un mapa dialectal y distintas tablas fonéticas, el usuario selecciona un elemento fónico de un determinado dialecto. El Portal presenta entonces una grabación —de audio o de vídeo, según el tipo de descripción que se haya previamente elegido, acústica o articulatoria— de una palabra que contiene el elemento seleccionado, acompañada de una completa ficha explicativa y de una gráfica de análisis acústico.

La ficha explicativa contiene los siguientes datos:

- Informante: nombre, edad, localidad.
- El elemento seleccionado, la palabra que lo contiene y su transcripción fonética.
- Los datos numéricos básicos para su descripción acústica.²⁵
- Comentarios

La gráfica del análisis acústico incluye oscilograma, curva de *pitch*, espectrograma de banda ancha y transcripción fonética.

SECCIÓN 3: BASE DOCUMENTAL

Por medio de un buscador bibliográfico, se accede a una base documental que pretende reunir todas las referencias bibliográficas existentes sobre el nivel fónico de la lengua. Para manejar adecuadamente el volumen de información reunido, el buscador cuenta con una completa colección de criterios de búsqueda:

- Tipo de asiento bibliográfico: libro, capítulo de libro, artículo de revista, tesis doctoral, etc.
- Soporte: papel, CD, internet, etc.
- Dialecto.
- Tema: general, vocalismo, consonantismo, prosodia.
- Sincronía / Diacronía.
- Otros: título, autor, año.

24. El material de audio pertenece al corpus *Euskararen Hotsak*, que se presentará más adelante.

25. La naturaleza de estos datos depende, obviamente, del elemento en cuestión

Cada búsqueda genera una página de resultados donde, de forma resumida, aparece el asiento o asientos localizados.²⁶ El usuario puede pinchar sobre el título para acceder a una ficha completa sobre el asiento seleccionado.

SECCIÓN 4: CORPUS ORALES DEL EUSKARA

Éste es uno de los apartados centrales del Portal. En rigor, no se trata de un corpus oral único, sino de una colección de corpus de distintos autores y características. La unidad lingüística de los corpus puede ser de tres tipos: palabras, frases o habla continua.

La sección dispone de un único buscador exhaustivo para todos los materiales orales, con siete criterios diferentes. Cada ejemplo seleccionado suministra, junto con el audio, una completa ficha con los datos correspondientes, entre ellos una cuidada transcripción fonética.

Esta sección de Corpus orales del euskara se expondrá con más detalle en los apartados siguientes.

SECCIÓN 5: SOCIOS

La sección de socios consiste actualmente en un foro de debate sobre cuestiones de fonética vasca. En un futuro, se pretende reformar esta sección y dotarla de mayor cercanía para con los usuarios, con más herramientas y posibilidades.

SECCIÓN 6: NOTICIAS

Finalmente, FonAtari cuenta con dos secciones más o menos habituales en este tipo de Portales:

La sección de Noticias (novedades, libros, cursos, becas, etc.). Desde hace algunos años, las noticias de FonAtari están sindicadas con otros *sitios* web, y pueden a su vez sindicarse.

La sección de enlaces (*links*). En especial, este segundo apartado cobra una importancia cada vez mayor, debido al aumento imparable de los recursos *on line*, de modo que su relación con la sección bibliográfica es cada vez más estrecha.

3.2. Los corpus orales de FonAtari: filosofía general

Como se ha dicho, el portal FonAtari no contiene, en sentido estricto, un solo corpus oral, sino una colección de corpus orales, accesibles todos ellos mediante una herramienta unitaria.

26. La información suministrada en esta página de resultados es: título; autor; fecha; revista (para los artículos).

La filosofía que sustenta esta colección es simple. Como se ha dicho en un apartado anterior, la recogida de material lingüístico ha sido, en los últimos años, una constante preocupación de casi todos los estudiosos del euskara, en general, y de su nivel fónico, en particular. Así, a la hora de plantearse un trabajo serio de descripción lingüística, la primera tarea consistía casi siempre en la adquisición de un material lingüístico *ad hoc*. De hecho, en numerosas ocasiones dicha adquisición suponía un porcentaje muy elevado del trabajo que debía realizarse (y del presupuesto disponible); incluso, a veces, el trabajo no pasaba más allá de esta fase inicial de *recogida de material*.

De esta manera, en las últimas décadas, se han recogido cientos de horas de grabaciones lingüísticas, varias de ellas al menos de enorme interés. Ahora bien, ¿dónde están esas grabaciones? Nadie lo sabe: la mayoría, en manos de sus dueños, lo que vale tanto como decir que están perdidas, puesto que resultan completamente inaccesibles.

De las grabaciones más *afortunadas* (aquéllas que han dado frutos en forma de publicaciones, científicas o divulgativas) apenas si están accesibles algunas de las más recientes.²⁷

En resumidas cuentas la situación, al menos antes de FonAtari, era la siguiente:

- La recogida de materiales realizada hasta el momento había sido, en cierta medida, estéril, puesto que no había generado transferencia de conocimiento, o en un grado muy pequeño.²⁸
- La mayor parte del material lingüístico, independiente del rendimiento que hubiese producido, se mantenía inaccesible para la comunidad científica.
- De esta manera, la tarea de recogida del material suponía muchas veces el primer gran escollo —tantas veces insalvable— a la hora de iniciar un proyecto de investigación.²⁹

FonAtari pretende contribuir en la superación de este problema, haciendo que los esfuerzos de la comunidad investigadora puedan optimizarse,

27. Las más características son las descripciones de *El habla de...* realizadas por Iñaki Gaminde, que se reseñan en el apartado bibliográfico, ya que van casi siempre acompañadas por un CD con material lingüístico. Gran parte de ese material, por otra parte, ha sido recogido en la web *Bizkairaren Fonotekea*.

28. Por supuesto, huelga decir que pueden señalarse numerosas excepciones, puesto que pueden citarse varias descripciones valiosas basadas en material lingüístico recogido por el o los autores.

29. Nuevamente, es obvio que ciertos enfoques y ciertos proyectos precisarán siempre de un material lingüístico recogido *ad hoc*.

puesto que las grabaciones realizadas por un investigador pueden ser reutilizadas por otro.³⁰

Para eso, las colecciones tienen que estar disponibles en un único lugar, y accesibles mediante una herramienta de búsqueda única (de acuerdo a unos criterios unitarios). Por su parte, este *recipiente* tiene que ofrecer una garantía de seriedad científica, de modo que para los investigadores pueda resultar rentable depositar sus fondos en él.

3.3. Los corpus de FonAtari: diseño técnico

La base de datos oral de la colección de corpus de FonAtari consta de cerca de 10.000 archivos de sonido, en un doble formato: WAV para las palabras y las frases y MP3 para los textos. Las especificaciones técnicas de los archivos dependen de la calidad original de la grabación,³¹ pero la mayoría de ellos se ofrecen a 44 ó 22 KHz., 16 bits, mono. Los archivos MP3 se han convertido a partir de archivos WAV a 128 Kbps. Los archivos se ordenan en el servidor de acuerdo al conocido procedimiento del *árbol de directorios*.

Una base de datos en MySQL contiene todos los datos relacionados con cada uno de los registros. La gestión de todo el sistema se realiza por medio de un *software* específico de administración accesible mediante una intranet privada.³²

En cada registro oral, la información que se suministra al sistema por medio de la intranet es la siguiente:

1. Nombre del registro: la palabra o la frase para estos dos tipos de material, y un título descriptivo para los textos.
2. Transliteración: se trata de una transcripción ortográfica, en la versión dialectal normalizada correspondiente, que incorpora además la mayor cantidad posible de información fonética añadida.
3. Transcripción fonética: se trata de una transcripción fonética semies-trecha, que no incorpora ninguna información prosódica. Sigue los criterios de la IPA (93) y utiliza las fuentes *SIL Manuscript IPA 93*.
4. Autor de la transliteración.

30. Además, la multiplicidad de los intereses de los distintos investigadores puede generar corpus con características también distintas. Poder disponer de varios corpus, con características diferentes, puede suponer un enriquecimiento añadido para una nueva investigación.

31. Las grabaciones realizadas por el equipo Fonetiker son siempre con las especificaciones conocidas habitualmente como *calidad CD*.

32. La intranet controla todas las secciones del portal. Aquí se expondrá únicamente lo relacionado con la colección de corpus orales.

5. Autor de la transcripción.
6. Corpus: conduce a otra base de datos con los datos pertinentes (nombre, autores, depositario y fecha de recogida).
7. Unidad (palabra – frase – texto).
8. Tipo de grabación (lingüística leída – lingüística guiada – lingüística semidirigida– lingüística libre – etnográfica –literatura oral – otros).
9. Informante: conduce a otra base de datos con los siguientes datos:
 - a. Identificador: nombre.³³
 - b. Año de nacimiento.
 - c. Edad en el momento de realizar la encuesta.
 - d. Sexo.
 - e. Lengua materna.
 - f. Alfabetización en euskara.
 - g. Nivel de estudios (primario – medios – superiores).
 - h. Dialecto. FonAtari maneja la clasificación dialectal clásica de Louis Lucien Bonaparte, a la que añade la variedad estándar (conocida como *euskara batua*) y una variedad indeterminada para los casos de difícil adscripción dialectal.
 - i. Localidad de origen.³⁴
10. Fichero de grabación: suministra al programa el *path* del fichero de audio en cuestión.

Por último, el portal de FonAtari dispone de una herramienta unitaria para la consulta web de todo el material. Por medio de una serie de desplegables, el usuario solicita al sistema un listado del material al que se desea acceder. Antes de nada, puede escogerse entre buscar por registros o buscar por corpus.

Los desplegables de la búsqueda de registros son los siguientes:

1. Dialecto/Localidad. El usuario, en primer lugar, escoge entre las dos posibilidades. Si se selecciona la primera, aparecen los dialectos señalados anteriormente. Si se selecciona la segunda, se escoge primeramente la provincia, y después una localidad de la lista de localidades cubiertas por los informantes disponibles.
2. Tipo de grabación.

33. Los informantes se identifican habitualmente con iniciales, salvo en el caso de personas distinguidas por alguna razón (escritores, intelectuales, etc.), que aparecen con el nombre completo.

34. Conviene distinguir separadamente dialecto y localidad, porque la equivalencia entre ambos no siempre es exacta, puesto que hay un buen número de localidades fronterizas donde conviven hablantes de diferentes dialectos.

3. Unidad lingüística.
4. Informante: se puede seleccionar el sexo, el grupo de edad (hasta 30, de 30 a 60, mayor de 60), la lengua materna y el grado de instrucción de los informantes. También se puede localizar un nombre (o conjunto de iniciales) concreto.
5. Contexto fonético. Esta herramienta es de un gran interés, puesto que permite localizar los registros que contengan un determinado contexto fonético, que puede además expresarse en caracteres ortográficos o en caracteres fonéticos (por medio de la fuente señalada). De esta manera, el corpus de FonAtari permite con gran facilidad su utilización para estudios fonéticos.

Entendido como corpus unitario, FonAtari posee, pues, una sólida estructura que permite ordenar los materiales que alberga de un modo adecuado para los fines que pretende.

3.4. Los corpus de FonAtari: fondos actuales

FonAtari contiene, a día de hoy, nueve corpus orales. Sus características principales se especifican a continuación:

1. CORPUS ARNEFE
 - Autores: equipo Fonetiker (1998-2001)
 - Material: palabras
 - Informantes: 13
 - Dialectos: 3
 - Registros: 7.097
 - Grabación: digital – Laboratorio
 - Descripción: inventario exhaustivo de los elementos y posiciones del euskara
 - Observaciones: parte del material del proyecto ARNEFE³⁵
2. CORPUS EUSKARAREN HOTSAK
 - Autores: equipo Fonetiker (1995-1996)
 - Material: palabras – texto (leído y libre)
 - Informantes: 8

35. El proyecto ARNEFE (“Aplicación de redes neuronales para la evaluación automática del nivel fónico del euskara”) fue desarrollado por el equipo Fonetiker en el Laboratorio de Fonética durante los años 2001-2003.

- Dialectos: 8
- Registros: 321
- Grabación: digital – de campo
- Descripción: inventario fonético (palabras), más un texto leído equi-parable (parábola), más una pequeña conversación libre, en todos los dialectos
- Observaciones: versión web del CD-Rom *Euskararen Hotsak*³⁶

3. CORPUS EUSKALKIZ EUSKALKI

- Autora: Rosa Miren Pagola
- Material: texto
- Informantes: 13
- Dialectos: 7
- Registros: 40
- Grabación: analógica – de campo
- Descripción: audio de un trabajo publicado previamente en libro y *cassette*, nuevamente transcrito
- Observaciones: implementación en desarrollo³⁷

4. CORPUS RMP

- Autora: Rosa Miren Pagola
- Material: palabras – frases – texto
- Informantes: 22
- Dialectos: 4
- Registros: 543
- Grabación: analógica – de campo
- Descripción: selección del material de audio para la tesis *Euskal fonetika Nafarroan*,³⁸ nuevamente transcrito
- Observaciones: implementación en desarrollo

5. CORPUS DU-KO DIALEKTOLOGIAKO ESKOLAK

- Autora: Rosa Miren Pagola – alumnos (1981-1991)
- Material: palabras – frases – texto
- Informantes: 42
- Dialectos: 3
- Registros: 1.663

36. Elejabeitia; Iribar & Pagola (1998).

37. Los corpus en los que se señala la “implementación en desarrollo” constan de un número variable de registros en diferentes fases de tratamiento, que esperan su incorporación a la base de datos.

38. Pagola (1992).

- Grabación: analógica – de campo
- Descripción: selección de trabajos tutelados de Dialectología Vasca, nuevamente transcritos
- Observaciones: implementación en desarrollo

6. CORPUS ESEUSCA

- Autora: Carmen Isasi – Alexander Iribar (2006-)
- Material: texto
- Informantes: 3
- Dialectos: 1
- Registros: 23
- Grabación: digital – de campo
- Descripción: parte de euskara del corpus bilingüe ESCA, dedicado al estudio del seseo castellano del País Vasco³⁹
- Observaciones: en desarrollo

7. CORPUS EITB

- Autora: EITB – Euskal Irrati Telebista (2000-)
- Material: texto
- Informantes: 2
- Dialectos: 2
- Registros: 12
- Grabación: digital – Radio Euskadi
- Descripción: archivo sonoro de EITB. Entrevistas a personajes ilustres y famosos de la cultura vasca
- Observaciones: implementación en desarrollo

8. CORPUS EIBARKO EUSKARA

- Autora: Nerea Areta (1996-1999)
- Material: palabras – texto
- Informantes: 4
- Dialectos: 1
- Registros: 111
- Grabación: analógica – de campo
- Descripción: selección del material de audio para la tesis *Eibarko Euskara*,⁴⁰ nuevamente transcrito
- Observaciones: implementación en desarrollo

39. La parte equivalente en castellano del corpus ESCA se denomina ESESCA. Todo el corpus ESCA está actualmente en fase de desarrollo.

40. Areta (2003).

9. ERRONKARIKO DOROTEO DE MIGUEL

- Autores: Luis Ángel Isasi – Javier Lauzirika (1977)
- Material: texto
- Informantes: 1
- Dialectos: 1
- Registros: 2
- Grabación: analógica – de campo
- Descripción: antigua grabación de uno de los últimos hablantes del dialecto roncalés, nuevamente transcrita
- Observaciones: calidad muy deficiente

Como puede observarse del listado precedente, FonAtari está en fase de implementación de una gran cantidad de material, correspondiente a diversos corpus. No obstante, cuenta además con nuevos corpus que incorporar a sus fondos, correspondientes a estudios realizados en diferentes localidades del País Vasco.

La siguiente tabla resume los actuales fondos orales de FonAtari. Conviene destacar que estos fondos se incrementan semanalmente.

	Palabras	Frases	Textos	Registros
ARNEFE	•			7.097
Euskararen Hotsak	•		•	321
Euskalkiz Euskalki			•	40
RMP	•	•	•	654
DU-ko Dialekologiako Eskolak	•	•	•	1.663
ESEUCA			•	23
EITB			•	12
Eibako euskara	•	•		111
Erronkariko Doroteo de Miguel			•	2
				9.923

BIBLIOGRAFÍA

- Ansorena, J. Ignacio (1993): *Euskal Kantak*. Donostia: Erein.
- Ansorena, J. Ignacio (1995): “La colección musical de Iztueta y Albéniz”, *Txistulari* 163, 29-46.
- Arete, Nerea (2003): *Eibarko euskara*. Bilbao: Universidad de Deusto.

- Azkue, R. M (1968): *Cancionero popular vasco*. Bilbao: La Gran Enciclopedia Vasca.
- Elejabeitia, A.; A. Iribar & R. M. Pagola (1998): *Euskararen Hotsak*, CD-ROM. Bilbao: Universidad de Deusto.
- Etxebarria, J. M.^a (1984): "Vienako euskal grabaketak", *Euskera* 29, 395-397.
- Etxebarria, J. M. (1990a): "Las grabaciones en euskera de la fonoteca de Viena. Cuentos y anécdotas en dialecto vizcaíno", *Etniker – Bizkaia aldizkaria* 9, 187-204.
- Etxebarria, J. M. (1990b): "Las grabaciones en euskera de la fonoteca de Viena. II. Grabaciones en dialecto labortano", in *Anuario de Euskal Folklore* 36, 59-64.
- Etxebarria, J. M. (1990c): "1913ko Vienako euskal grabaketak. III Erronkarierra", *IKER* 6, 161-176.
- Etxebarria, J. M. (1990d): "Euskalkietan egindako lehenengo grabaketak", *IKER* 7, 607-618.
- Etxebarria, J. M. (1991): "Euskalki guztien etno-testuak", in *Eskola Apunteak* 17, Bilbao: Labayru Ikastegia.
- Etxebarria, J. M. (1993a): "Las grabaciones de Viena en euskera. Dialectos Alto navarro septentrional y meridional", *Mundaiz* 45, 161-176.
- Etxebarria, J. M. (1993b): "Las grabaciones de Viena en euskera. Dialecto guipuzcoano", *Mundaiz* 47, 25-50.
- Etxebarria, J. M. (1994): "Las grabaciones de Viena en euskera. Bajo navarro occidental y Bajo navarro oriental", *Mundaiz* 47, 45-76.
- Etxebarria, J. M. (1995): "Vienako euskal grabaketak", *Enseiucarrean* 11, 93-114.
- Etxebarria, J. M. (1996): "Vienako euskal grabaketak. Zuberera", *Enseiucarrean* 12, 177-199.
- Gaminde, Iñaki (1999): *Bakio berbarik berba*. Bilbao: BBK.
- Gaminde, Iñaki (2000): *Zamudio berbarik berba*. Bilbao: BBK.
- Gaminde, Iñaki (2001a): *Meñaka berbarik berba*. Meñaka: Ayuntamiento de Meñaka.
- Gaminde, Iñaki (2001b): *Leioa berbarik berba*. Leioa: Ayuntamiento de Leioa.
- Gaminde, Iñaki (2001c): *Mungia berbarik berba*. Mungia: Ayuntamiento de Mungia.
- Gaminde, Iñaki (2002a): *Arrieta berbarik berba*. Arrieta: Ayuntamiento de Arrieta.
- Gaminde, Iñaki (2002b): *Jatabe berbarik berba*. Jarabe: Ayuntamiento de Jatabe.
- Gaminde, Iñaki (2003): *Zaldibar berbarik berba*. Zaldibar: Ayuntamiento de Zaldibar.

- Gaminde, Iñaki (2004): *Berango berbarik berba*. Berango: Ayuntamiento de Berango.
- Hernández, I. et al. (2002): "BIZKAIFON: A sound archive of dialectal varieties of spoken Basque", in *Proc. of the 3th Int. Conf. on Languages Resources and Evaluation*, Las Palmas de Gran Canaria. Accesible también en <http://bips.bi.ehu.es/~inma/publicaciones/LREC2002.pdf>.
- Iztueta, J. I. & P. Albéniz (1826): *Euscaldun Anciña anciñaco ta are lenda-bico etorquien Dantza on iritci pozcarri gaitzic gabecoen Soñu gogo-angarrac beren itz neurtu edo Versoaquin*. Ed. crítica de J. Ignacio Ansoarena, *Txistulari* 163, 1995.
- Mazusta, Aintzina (2001): "Euskarazko Lehen Soinuak", *Revista Internacional de Estudios Vascos* 46/1, 364-365. Accesible también en <http://www.euskomedia.org/PDFAnlt/riev/4603640365.pdf>.
- López de Arana, Inaxio (2004): "Rudolf Trebitsch-en grabazio musikalak", *Sancho el Sabio* 21, 49-89.
- López de Arana, M. & I. López de Arana (1998): "Rudolf Trebitsch: Euskal hizkuntzaren eta musikaren grabazioak (1913)", *Fontes linguae vasconum* 30/77, 97-118. Accesible también en http://dialnet.unirioja.es/servlet/fichero_articulo?codigo=26260&orden=57319.
- Llisterri, J.; M. J. Machuca; C. de la Mota; M. Riera & A. Ríos (2005) : "Corpus orales para el desarrollo de las tecnologías de habla en español", *Oralia. Análisis del discurso oral* 8, 289-325.
- Pagola, R. M. (1992): *Euskal fonetika Nafarroan*, 2 t. Pamplona: Gobierno de Navarra, Dpto. de Educación y Cultura.
- Pagola, R. M. (2004): *Bonaparte ondareko eskuizkribuak*. Bilbao: Universidad de Deusto [CD-ROM].
- Pagola, R. M.; S. González; I. Iribar & J. J. Iribar (1992): *Bonaparte ondareko eskuizkribuak. Bizkaiera*. 3 vols. Bilbao: Universidad de Deusto.
- Pagola, R. M.; A. Beola; I. Iribar & J. J. Iribar (1993): *Bonaparte ondareko eskuizkribuak. Gipuzkera*. 6 vols. Bilbao: Universidad de Deusto.
- Pagola, R. M.; E. Alkiza; A. Beola; I. Iribar & J. J. Iribar (1994): *Bonaparte ondareko eskuizkribuak. Lapurtera*. 2 vols. Bilbao: Universidad de Deusto.
- Pagola, R. M.; E. Alkiza; A. Beola; I. Iribar & J. J. Iribar (1995): *Bonaparte ondareko eskuizkribuak. Iparraldeko Goi-Nafarrera*. 3 vols. Bilbao: Universidad de Deusto.
- Pagola, R. M.; E. Alkiza; A. Beola; I. Iribar & J. J. Iribar (1996): *Bonaparte ondareko eskuizkribuak. Hegoaldeko Goi-Nafarrera*. 4 vols. Bilbao: Universidad de Deusto.
- Pagola, R. M.; I. Iribar & J. J. Iribar (1997a): *Bonaparte ondareko eskuizkribuak. Aezkera*. Bilbao: Universidad de Deusto.

- Pagola, R. M.; I. Iribar & J. J. Iribar (1997b): *Bonaparte ondareko eskuizkribuak. Erronkariera*. Bilbao: Universidad de Deusto.
- Pagola, R. M.; I. Iribar & J. J. Iribar (1997c): *Bonaparte ondareko eskuizkribuak. Zaraitzera*. Bilbao: Universidad de Deusto.
- Pagola, R. M.; I. Iribar & J. J. Iribar (1999a): *Bonaparte ondareko eskuizkribuak. Mendebaldeko Behe-Nafarrera*. Bilbao: Universidad de Deusto.
- Pagola, R. M.; I. Iribar & J. J. Iribar (1999b): *Bonaparte ondareko eskuizkribuak. Ekialdeko Behe-Nafarrera*. 2 vols. Bilbao: Universidad de Deusto.
- Pagola, R. M.; I. Iribar & J. J. Iribar (1999c): *Bonaparte ondareko eskuizkribuak. Zuberera*, 2 vols. Bilbao: Universidad de Deusto.
- Riezu, P. Jorge (1971): "Material folklórico de la Collectanea Linguistica de Humboldt. Canción del vino", *Boletín de la Institución Sancho el Sabio* XXV, 91-115.
- Sánchez, J. M. et al. (2001): "Base de datos oral y textual para el euskara" in *XVI Simposio URSI*, Madrid, 551-552 [CD-Rom]. Accesible también en http://bips.bi.ehu.es/~inma/publicaciones/URSI2001_J_AURKEZPENA.pdf.
- Sánchez, J. M. et al. (2002): "Implementación de base de datos oral y textual para el euskara", in *XVII Simposio URSI* [CD-Rom]. Accesible también en http://bips.bi.ehu.es/~inma/publicaciones/URSI2002_S.pdf.
- Sallabery, J. D. J. (1870 [1992²]) : *Chants populaires du Pays Basque*. Nîmes: Lacour.
- Trebitsch, Rudolf (1913 [2003]): *The collections of Rudolf Trebitsch. Basque Recordings 1913* [2 CD]. Wien: Österreichischen Akademie der Wissenschaften.
- Torruella, J. & J. Listerri (1999): "Diseño de corpus textuales y orales", in G. Clavería; C. Sánchez & J. Torruella (eds.): *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Barcelona: Universidad de Barcelona / Ed. Milenio, 45-77.
- Zulaika, J. G. (Aita Donostia) (1994): *Cancionero vasco*. Donostia: Eusko Ikaskuntza.