Análisis automático del diccionario Hauta-Lanerako Euskal Hiztegia

Arriola, J.M., Artola, X., Soroa, A.
Departamento de Lenguajes y Sistemas Informáticos (Universidad del País Vasco)
Facultad de Informática, Apdo. 649 20080 Donostia
E-mail: jibaregj@si.ehu.es

Area temática: Lexicografía computacional.

Resumen

El propósito de la siguiente comunicación es el de dar a conocer la labor desarrollada en el proceso de análisis del Hauta-Lanerako Euskal Hiztegia (HLEH) [Sarasola, 84/95]. Para ello se describirán las características más importantes del mismo, para pasar a continuación a detallar las fases concernientes a la preparación del diccionario fuente en MRD y la confección de la gramática que refleje la estructura del mismo.

Introducción

El objetivo fundamental del presente trabajo es el de describir los pasos dados de cara a la construcción de una Base de Datos Léxica (BDL) a partir de un Diccionario de uso humano en soporte magnético (Machine Readable Dictionaries, MRD): el Hauta-Lanerako Euskal Hiztegia (HLEH), con la finalidad de facilitar el acceso así como la explotación de la información léxica, para su posterior integración y aprovechamiento en el enriquecimiento de la Base de Datos Léxica para el Euskara (EDBL) diseñado como soporte para diversas aplicaciones, entre las cuales podemos señalar el corrector ortográfico para el euskara *Xuxen* [Agirre et al., 92]. Se enmarca por tanto en el ámbito de la reutilización de los recursos léxicos existentes como una de las posibles respuestas a las necesidades que genera el desarrollo de una aplicación de dimensión real. La utilización de diccionarios de uso humano en soporte informático para la construcción de lexicones ha constituido la aportación más significativa en esta línea [Amsler 81], [Calzolari, Picchi 86], [Boguraev et al. 91], [Castellón 93].

1. Estudio de las características del diccionario fuente

Este apartado constituye un paso previo básico para acometer el proyecto de construcción de la BDL como sostén para el proceso de enriquecimiento de la EDBL; así se estudiaron cuidadosamente tres aspectos fundamentales:

A. La aplicabilidad de la información contenida en el HLEH para el enriquecimiento de la EDBL.

La información que incluye en estos momentos la EDBL para las más de 60.000 entradas es de carácter léxico-morfológico fundamentalmente (se está introduciendo cierta información sintáctica). Se consideró como fuente idónea el HLEH al ser un diccionario normativo y repositorio del euskara estándar[1], para la obtención de nuevas entradas así como la información básica correspondiente a las mismas: categoría y subcategoría básica en el caso de los verbos. Evidentemente ésta es la información menos costosa de obtener. En un futuro próximo, se tratará de explotar la información no explícita del campo de las definiciones (para ir dotando a la EDBL de información semántica) y los ejemplos (para el estudio de los patrones básicos de subcategorización de los verbos, aportando a la EDBL una información sintáctica más completa) .

1Hay que señalar que no toda las entradas tienen el mismo peso, de ahí el calificativo de *Hauta-Lanerakoa.*

B. Cómo se encuentra estructurada dicha información y los criterios de los que se han servido los lexicógrafos para ello: ordenación de homógrafos, acepciones, etc.

Si bien la estructuración de la información y los criterios seguidos para ello no siempre son coincidentes con los de la EDBL, se ha dado una interrelación entre ambos. Así por ejemplo, las entradas verbales se han desarrollado en la EDBL siguiendo el criterio del HLEH. En otros casos, como la ordenación de las acepciones, en el HLEH se han seguido criterios cronológicos que no son muy adecuados para la EDBL. De cara a una futura integración plena de la información extraída a partir de la BDL, todas estas características se tuvieron en cuenta en el diseño de la EDBL [Agirre et al., 95].

C. Herramientas disponibles para adquirir dicha información y su posterior representación.

No nos planteamos la construcción de herramientas de cobertura general; la filosofía básica tanto para el análisis y acceso al diccionario como para la representación de la información obtenida ha sido la de aprovechar las experiencias existentes en este campo, principalmente la del grupo ACQUILEX, el cual se plantea la construcción de herramientas potentes y aplicables a cualquier diccionario.

1.1. Características del diccionario fuente

Se trata de un diccionario monolingüe en euskara que es presentado por su autor como un diccionario general que sirve de apoyo al afianzamiento del euskara estándar y para cuya confección se ha sustentado principalmente en la tradición literaria. Siguiendo los criterios señalados en el prefacio, pasaremos a continuación a mostrar las partes de las que se componen los artículos y los rasgos más significativos de las mismas. He aquí un artículo, tal como aparece en la versión del MRD:

mehartu, mehar, mehartzen. da-du ad. (*XVII ea., 1759). Mehar edo meharrago bihurtu. Ik. **estutu; mehetu.** *Bidea mehartzen den tokian. Sutegian goritzen eta biguintzen bada burdina, errazki mehartzen da ingudean. Ez dio maitasun horrek gogoa mehartu batere. Hizkuntzaren lurrak eta mugak mehartu eta murriztu ahala, hedatzen eta zabaltzen ari dirá literaturarenak.*

1.1.1. Entrada

Las entradas aparecen en negrita. Siguiendo la tradición lexicográfica desarrollada por los lexicógrafos vascos, normalmente los derivados tienen su propia entrada, si bien a veces aparece la raíz junto a su derivado. En el caso de los verbos se ordenan en función del participio que irá seguido de las formas no aspectuales. Otro punto a señalar es el de que junto a la forma estándar aparecen sus variantes y en otros casos es la propia variante la constituyente de una entrada precedida por un símbolo (* o cruz) que indica que no pertenece al euskara estándar e irá seguido de un símbolo de relación que remite a la forma correcta correspondiente.

Cuando existen diferentes homógrafos se identifican por medio de un número en negrita.

1.1.2. Categoría

En este campo podrán aparecer tanto categorías gramaticales como rasgos morfológicos (Sing., Pl., ize., adj., ad., etc.). Dicha información aparece normalmente después de la entrada. En las correspondientes a verbos y adverbios, antes de la categoría aparece una información que enriquece la primera. Así para los verbos se señala una subcategorización básica que indica por medio de una etiqueta (explicada en el prefacio del diccionario) los complementos que exige dicho verbo para la concordancia. Y en cuanto a los adverbios se señala el tipo de adverbio (si es de modo, etc.).

1.1.3. Fecha

Se precisa la fecha de aparición de la entrada en algún diccionario o texto literario, y en ocasiones se recogen las variantes precedentes y de cuando datan éstas.

1.1.4. Definición

En las entradas con una sola acepción este campo se compondrá del texto de la definición. Cuando hay más de una acepción los lexicógrafos se sirven de diversos símbolos para diferenciar entre las distintas acepciones: mediante números, mediante letras mayúsculas y en negrita para agrupar éstas bajo sentidos generales, etc. También se marcan a través de

símbolos (#1,#2, etc.) las matizaciones de algunas acepciones.

El criterio de ordenación de las acepciones es cronológico. Ciertas palabras componentes del texto de la definición van acompañadas de un número que indica el sentido en el que están siendo empleadas si es que éste se aleja del sentido más usual.

1.1.5. Relaciones

El campo de las relaciones se configura con los códigos lexicográficos (Ik., Ant.) que sirven para expresar relaciones de sentido, ya sean de sinonimia o antonimia.

1.1.6. Ejemplo

Este campo está constituido por texto en cursiva y aparece normalmente tras el de la definición. A través del mismo se tratará de fijar el uso más frecuente, fijándose aquellos contextos en los que puede aparecer dicha palabra y remarcándose las palabras con las que tiene una relación más estrecha.

1.1.7. Abreviaturas

Constan de dos tipos básicos, las de uso y las temáticas (aparece una relación de las mismas en el prefacio). Dentro de las de uso se podrían incluir: las que señalan el dialecto al que pertenecen, el tipo de registro (familiar, elevado, etc.) y el nivel de utilización de las mismas (poco utilizadas a lo largo de la tradición literaria o en la actualidad, etc.). También podemos encontrarnos con ciertas abreviaturas que se escapan de los dos grupos principales, como son las que nos indican si una entrada o acepción son de creación reciente.

1.1.8. Subentradas

Aparecen en letra mayúscula y pueden considerarse como auténticas entradas pero dentro de otra entrada. Son formas declinadas o construcciones (sustantivo + verbo, sustantivo + adjetivo, etc.) que por su significado o uso especial son remarcadas en mayúscula.

1.1.9. Explicaciones gramaticales

A lo largo de los diversos tipos de entrada, nos podemos encontrar con explicaciones tales como en el caso del verbo *adiskidetu* (hacer las paces), para el cual se indica entre paréntesis que el objeto de dicho verbo ha de ser de persona e ir expresado en el caso asociativo (-kin).

2. Tratamiento del MRD

El diccionario fuente, al estar dirigido al uso humano, presenta la problemática de las versiones impresas, la carencia de una debida estructuración de la información. Se trata por tanto del fruto de una labor lexicográfica en la que la informática ha jugado un papel auxiliar como ayuda para el lexicógrafo a la hora de configurar la estructura de los artículos, sirviéndose para ello meramente de diversos procesadores de texto. Todo ello conlleva que nos hayamos tenido que enfrentar a un continuum textual sobre el que se habrán de realizar las siguientes labores:

2.1. Etiquetado

El etiquetado tiene como objetivo la segmentación del texto fuente categorizando aquéllos elementos que nos serán útiles para identificar los artículos y los campos de los que se componen los mismos. Dichos elementos son de dos tipos:

A. Códigos correspondientes a la información lexicográfica.

Son los identificadores de los que se sirve el lexicógrafo para especificar las diversas partes de un artículo: número de acepción, grupo de acepción, etc. A continuación describiremos las etiquetas que hemos empleado:

- AZH / AZB: inicio y final del campo delimitado por el número de acepción.

- AXH / AXB: inicio y final del campo delimitado por los símbolos (#1, #2, etc.) que expresan matices de la acepción.

- AMH / AMB: inicio y final del campo delimitado por letras mayúsculas en negrita, que indican una serie de acepciones agrupadas bajo un mismo sentido más general.

- IKH / IKB: inicio y final del campo de relación que se expresa por medio del símbolo Ik. en negrita y que remite a entradas sinónimas.

- ANH / ANB: inicio y final del campo de la relación de antonimia que se expresa por medio del símbolo Ant. en negrita.

- HZH / HZB: inicio y final del identificador de homógrafo.

- DH / DB: inicio y final del identificador correspondiente a la fecha.

- GUR : indica que estamos ante una forma no estándar.

El etiquetado del campo de relación resultó problemático debido a que es utilizado con dos finalidades bien distintas: para relacionar sentidos y para remitir a otras entradas no ya en función del sentido sino de la forma. Este último uso se detecta en aquellas entradas que no son consideradas como euskara estándar y que irán precedidas en el campo correspondiente a la entrada por un asterisco o una cruz, por lo que en este contexto la misma etiqueta adquirirá una semántica diferente en la gramática. Si bien la dificultad anterior pudo ser solventada en la fase de etiquetado se prefirió dejarlo en manos de la fase de construcción de la gramática para no aumentar el número de etiquetas. El problema se agrava en el caso de los símbolos empleados como indicadores de matices en una acepción, puesto que además pueden aparecer como indicadores del cambio de categoría, si ésta se ha producido como consecuencia del paso de una acepción a otra. También pueden aparecer delante del número de acepción cuando la nueva acepción es debida a un cambio de campo (temático, etc.). Si bien a nivel de etiquetado se fija una sola etiqueta, todas estas circunstancias habrán de tenerse en cuenta a la hora de confeccionar la gramática.

B. Códigos tipográficos correspondientes al estilo (o tipo de letra).

Los códigos tipográficos que hemos etiquetado son los correspondientes a los tipos de letra negrita y cursiva, que se indican por medio de las siguientes etiquetas:

- LH / LB: inicio y final de negrita.

- EH / EB: inicio y final de cursiva.

La negrita se usa en la entrada, en el número identificador de homógrafo y en los códigos lingüísticos empleados en el campo de la acepción (número de acepción, subacepción, etc.) y los de relación. La distinción entre los diversos elementos en negrita no entraña dificultad, los identificadores de homógrafo y número de acepción han de ser números, y en el caso de las subacepciones son letras en minúscula seguidas de un paréntesis o bien, si se trata de grupos de acepciones, además de ir en negrita han de ir en mayúscula y seguidos de un punto.

La cursiva a su vez se emplea para indicar diferentes elementos: categoría, abreviaturas de uso, abreviaturas temáticas, ejemplos y el participio de los verbos nominales. La distinción entre categorías y abreviaturas, y el resto resulta sencilla al pertenecer éstas primeras a un conjunto cerrado. Ahora bien, la distinción entre ejemplos y el participio de las entradas de verbos nominales resulta un tanto más compleja y hay que tener en cuenta la posición en la que aparece el participio (el cual irá siempre seguido de la categoría de sustantivo verbal).

También hemos creado etiquetas para diferenciar los elementos que van en mayúscula:

- MH / MB: inicio de mayúscula y final de mayúscula.

El resto se etiquetará por defecto como texto con la etiqueta TX. Todas estas etiquetas permiten ir distinguiendo las diversas partes de las que se compone un artículo. Evidentemente también hemos de fijar las fronteras entre los distintos artículos, apoyándonos para ello en los delimitadores de párrafo de los que consta cada uno de éllos. Así pues les asignaremos las etiquetas:

- SH / SB: inicio y final de artículo.

Tanto para los códigos reseñados en A como en B, se observa una problemática común: la aparición de códigos no previstos o con una finalidad distinta a la que les fue asignada inicialmente.

Por medio del programa de etiquetado se marca el inicio y el final de cada campo asignándole una etiqueta; posteriormente se pasa a una estructura lispificada para que pueda servir como entrada del analizador de entradas escrito en Prolog. Veamos un ejemplo del resultado de la fase de etiquetado y su posterior transformación en estructura lispificada:

• Entrada etiquetada:

[SH][LH]mehartu,mehar,mehartzen.[LB]da-du[EH]ad.[EB][DH](*XVIIea.,1759)[DB].Mehar edo meharrago bihurtu. [IKH]Ik.[LH]estutu; mehetu.[LB][IKB][EH]Bidea mehartzen den tokian. Sutegian goritzen eta biguintzen bada burdina, errazki mehartzen da ingudean. Ez dio maitasun horrek gogoa mehartu batere. Hizkuntzaren lurrak eta mugak mehartu eta murriztu ahala, hedatzen eta zabaltzen ari dira literaturarenak. [EB][SB].

• Entrada etiquetada y dispuesta en forma de estructura lispificada:

e([[SH,''],[LH,''],[tx,'mehartu,mehar,mehartzen.'],[LB,''],[tx,'da-du'], [EH,''],[tx,'ad.'],[EB,''], [DH,''], [tx,*XVIIea.,[tx,'*XVIIea.,1759'],[DB,''],[tx,'.Mehar edo meharrago bihurtu.'], [IKH,''],[tx,'Ik.estutu; mehetu.'],[IKB,''],[EH,''], [tx,'Bidea mehartzen den tokian. Sutegian goritzen eta biguintzen bada burdina, errazki mehartzen da ingudean. Ez dio maitasun horrek gogoa mehartu batere. Hizkuntzaren lurrak eta mugak mehartu eta murriztu ahala, hedatzen eta zabaltzen ari dira literaturarenak. '] ,[EB,''],[SB,'']]).

2.2. Construcción de la gramática para el análisis

La gramática ha sido expresada en forma de una gramática de claúsulas definidas (DCG) que recoge la estructura de las entradas y para cuyo desarrollo ha sido necesario un proceso previo de etiquetado tal como se ha descrito en el punto precedente.

2.2.1. Estructura general de los artículos

En este punto presentaremos la estructura global de cada artículo apoyándonos para ello en el siguiente metalenguaje:

• <A>/<B> A o B , nodos alternativos.

• <A> <B> A y B, A seguido de B.

• [<A>] nodo opcional.

• <*NULL> nodo vacío.

• <*A> nodo terminal.

La forma de las reglas es la siguiente:

<ELEMENTO>=<EL1> <EL2>...<ELN>.

He aquí una pequeña gramática a modo de ilustración de la estructura general de los artículos:

<ARTICULO> =<ENTRADA_DIC> [<RELACIONES>] <CATEGORIA>
[<*FECHA>] [<DEFINICION_EJEMPLOS>] .

<ENTRADA_DIC> = [<NUMERO_DE_HOMOGRAFO>]
[<ENTRADA_CADUCA> / <ENTRADA_ESTÁNDAR>].

<NUMERO_DE_HOMOGRAFO> = <*HZH> <*NUMERO> <*HZB>.

<ENTRADA_CADUCA> = <*GUR> <*LH> <*ENTRADA> <*LB>.

<ENTRADA_ESTÁNDAR>= <*LH> <*ENTRADA> <*LB>.

<CATEGORIA> = [<*SUBCATEGORIA>] <CAT>.

<CAT> = <*EH> <*CAT> <*EB>.

<DEFINICION_EJEMPLOS> = <ACEPCION> [<EJEMPLOS>]
<DEFINICION_EJEMPLOS> / <*NULL>.

<ACEPCION>=[<NUMERO_ACEPCION>][<GRUPO_ACEPCION>]
<*TEXTO_DE_DEFINICION> [RELACIONES] .

<NUMERO_ACEPCION> = <*AZH> <*NUMERO_ACEPCION> <*AZB>.

<GRUPO_ACEPCION> = <*AMH> <*GRUPO_ACEPCION> <*AMB>.

<RELACIONES> = [<RELACION_SIN> / <RELACION_ANT>]
<RELACIONES> [EJEMPLOS]./<*NULL>.

<RELACION_SIN> = <*IKH> <*ENTRADA_REFERENCIADA> <*IKB>.

<RELACION_ANT> = <*ANH> <*ENTRADA_REFERENCIADA> <*ANB>.

<EJEMPLOS> = <*EH> <*TEXTO_EJEMPLO> <*EB>.

2.2.2. Gramática

A continuación iremos explicando brevemente las reglas de la gramática simplificada escrita a modo de ilustración:

• La primera regla de la gramática establece la estructura general de todo artículo, que se compondrá de la forma correspondiente a una entrada de diccionario seguida del campo opcional de relación, la categoría, la fecha (que será también opcional), la definición y los ejemplos, así como de los distintos tipos de

abreviaturas que pueden aparecer o no precediendo a éstos.

<ARTICULO> =<ENTRADA_DIC> [<RELACIONES>] <CATEGORIA>

[<*FECHA>] <DEFINICION_EJEMPLOS>.

• Entendemos <ENTRADA_DIC> como componente central del artículo, recogiéndose en la misma el número identificador de homógrafo (cuando ello sea pertinente) y la consiguiente forma bien sea ésta estándar o caduca:

<ENTRADA_DIC>=[<NUMERO_DE_HOMOGRAFO>]

[<ENTRADA_CADUCA> / <ENTRADA_ESTÁNDAR>].

• A través de la siguiente regla se puede ver el campo <RELACIONES> que tiene la finalidad de poner en relación una entrada de diccionario con una entrada o más de otros artículos. Dicha referencia puede deberse a razones de carácter semántico, estableciéndose relaciones sinonímicas o antonímicas, o a razones referentes a la vigencia de la forma, relacionándose la forma caduca en cuestión con la correspondiente forma estándar.

<RELACIONES>=[<RELACION_SIN>/<RELACION_ANT>]
<RELACIONES> /<*NULL>.

<RELACION_SIN> = <*IKH> <*ENTRADA_REFERENCIADA> <*IKB>

<RELACION_ANT> = <*ANH> <*ENTRADA_REFERENCIADA> <*ANB>

• La regla para el campo categoría ofrece el nodo opcional de subcategoría que aparecerá en el caso de los verbos indicándose los complementos exigidos para la concordancia. También se han considerando como elementos integrantes de <SUBCATEGORIA> aquellos que nos informan del tipo de adverbio (de modo, lugar, etc.). La categoría, <*CAT>, nos proporciona las categorías gramaticales clásicas (verbo, sustantivo, adjetivo, etc.), junto con información morfológica (singular, plural).

<CATEGORIA> = [<*SUBCATEGORIA>] <CAT>.

<CAT> = <*EH> <*CAT> <*EB>

• El campo denominado <*FECHA> aporta información referente a la fecha en la que la entrada de diccionario apareció en la literatura tradicional o en el documento que se trate, así como las posibles variantes de la misma y sus fechas de aparición. Dada la complejidad de dicha información, se consideró como solución más razonable no analizarla y agruparla en un campo común.

• El cuerpo de la definición y los ejemplos viene reflejado por la siguientes reglas:

<DEFINICION_EJEMPLOS>=<ACEPCION>[<MATICES>] [<EJEMPLOS>].

<DEFINICION_EJEMPLOS> / <*NULL>

<ACEPCION>=[<NUMERO_ACEPCION>] [<GRUPO_ACEPCION>]

<*TEXTO_DE_DEFINICION> [RELACIONES] [EJEMPLOS].

<MATICES>= <*AXH> <*TEXTO_MATIZACION> <*AXB>

<NUMERO_ACEPCION> = <*AZH> <*NUMERO_ACEPCION> <*AZB>

<GRUPO_ACEPCIONES>=[<GRUPO_ACEPCION>]
<*TEXTO_DE_DEFINICION> .

<GRUPO_ACEPCION> = <*AMH> <*GRUPO_ACEPCION> <*AMB>

Como puede observarse en la regla inicial de la gramática no todos los artículos se componen de acepción y ejemplos, se trata por tanto de un nodo opcional. Ahora bien cuando aparece ofrece las siguientes peculiaridades: si hay más de una acepción ésta irá precedida por un número, y en aquéllos artículos que contienen varias acepciones que se pueden agrupar bajo un mismo sentido general aparecerá un indicador del grupo de acepción precediendo a las mismas. Dichos nodos son a su vez opcionales puesto que hay artículos que tienen una sola acepción, que se compone del texto de definición y opcionalmente de ejemplos. Cada acepción puede estar seguida del campo recursivo de las relaciones pero no necesariamente. Otra información enriquecedora la constituye el nodo opcional de los matices. A través del mismo se matiza la acepción por medio de ejemplos o frases explicativas. Los ejemplos se componen de texto comprendido entre las etiquetas delimitadoras del mismo; este nodo es opcional.

2.3. Análisis

El resultado del proceso será el mismo texto del diccionario, organizado según una estructura predefinida y etiquetada por campos de la que se han eliminado dichas etiquetas porque ya no resultan de interés.

Por ejemplo:

(artikulua([artikulua([lema([],[],mehartu,mehar,mehartzen.)],[artikulu_biz([kategoriak([azpikategoria(da-du)],[],[kategoria(ad.)],[])],[data_zer([data (*XVIIea.,1759)],[])],[],[],[],[],[definizio_adib_ani([],[], [definizio_adib([],[definizioa(.Mehar edo meharrago bihurtu.)],[],[],[],[])],[])])],[erlazioak ([erlazioa(Ik.estutu; mehetu.,[adibidea(Bidea mehartzen den tokian. Sutegian goritzen eta biguintzen bada burdina, errazki mehartzen da ingudean. Ez dio maitasun horrek gogoa mehartu batere. Hizkuntzaren lurrak eta mugak mehartu eta murriztu ahala, hedatzen eta zabaltzen ari dira literaturarenak.)])],[])])],[]))

El resultado del análisis en bruto resulta prácticamente ilegible, por lo que lo expresaremos del siguiente modo:

•entrada: [2] mehartu, mehar, mehartzen

•subcategoría: da-du

•fecha: *XVIIea, 1759

•definición: Mehar edo meharrago bihurtu.

•relación: IK. estutu, mehetu.

•ejemplo: Bidea mehartzen den tokian. Sutegia goritzen eta biguintzen bada burdina, errazki mehartzen da ingudean. Ez dio maitasun horrek gogoa mehartu batere. Hizkuntzaren lurrak eta mugak mehartu eta murriztu ahala, hedatzen eta zabaltzen ari dira literaturarenak.

Veamos otro ejemplo:

(artikulua([artikulua([lema([homografo_zenb(1)],[],sortu,sor,sortzen.,[])],[artikulu_biz([kategoriak([azpikategoria(da-du)],[], [kategoria (ad.)],[])], [data_zer([data((1545))] ,[])] ,[kategoriak ([],[],[])],[],[],[],[],[],[],[definizio_adib([adiera_ikur(1.)],[],[],,[kategoriak([],[],[])],[],[definizioa(Izatea, bizia hartu edo eman; (artean ez zen zerbait) egin.)],[adibidea(Ezerezetik sortu. Lurreko animaliak....,[])],[],[],[], [adiera_xehetasunak([adixe_ikurra(#2)],[kategoriak([],[],[])],[],[definizioa(Amaren sabalean gorpuztu.)],[], [adibidea(Salomonek sortu zuen Booz; Boozek sortu zuen Obed, Rutgandik. Bekatu gabe sortua. Kristo gure jauna ez zelako sortu, ezta ere jaio, beste gizonak bezala. Sabelean sortu eta izango duzu seme bat.,[])],[kategoriak ([],[],[])],[],[],,[definizio_adib ([adixe_ikurra(#2)] , [adiera_ikur(2.)],[],[], [kategoriak ([azpikategoria(da)],[],[kategoria(ad.)],[])],[], [definizioa(Jaio, amaren sabeletik irten.)],[adibidea(Baxenabarren sortu, Lapurdin da hazten, itsas haizeak ez du batere izutzen. Sortua ...[])],[],[],[],[],[])])],[])],[]))

•entrada: [3] sortu, sor, sortzen

•número_homografo: 1

•subcategoría: da-du

•categoría: ad.

•data: 1545

•acepción:

número_acepción:1

definición: Izatea, bizia hartu edo eman, ...

ejemplo: Ezerezetik sortu. Lurreko animaliak ...

subacepción: Amaren sabelean gorpuztu.

ejemplo: Salomonek sortu zuen Booz, ...

•acepción:

número_acepción:2

subcategoría: da

categoría: ad.

definición: Jaio, amaren sabeletik irten.

ejemplo: Baxenabarren sortu, Lapurdin da hazten ...

[2] estrechar

[3] crear

La labor de fijar la estructura de la futura DBL y los campos de los que ha de componerse ésta es una labor que queda por culminar.

3. Algunas consideraciones y conclusiones sobre los resultados del análisis.

La gramática se ha aplicado a la totalidad del MRD del que disponemos por el momento[4], habiéndose analizado 23.157 artículos, de los que hemos sido capaces de reconocer la estructura del 85% de los mismos. El proceso de análisis se aplicó inicialmente sobre una muestra que correspondía a una etapa más reciente en la confección del diccionario, lo cual incidió en la obtención de unos mejores resultados. Pero al tratarse del fruto de una labor lexicográfica llevada a cabo durante diez años, se ven reflejadas las diferencias en los resultados obtenidos muestra a muestra. Así, en las primeras los resultados son un tanto inferiores, debido a que en un principio se observa una cierta carencia en cuanto a la fijación de los criterios lexicográficos (códigos tipográficos diversos, mayor cantidad de errores tipográficos, etc.).

Hemos examinado los factores que inciden en los resultados obtenidos. A través del estudio se ha observado que los errores son en su mayoría tipográficos, siendo muy frecuente la aparición de un determinado tipo de letra fuera del campo para el que está destinado, dando lugar a fragmentaciones de texto erróneas.

Otro factor a tener en cuenta es la aparición de códigos lexicográficos que no han sido incluidos por el autor en el prefacio de la obra, o la aparición de los mismos pero en distinto formato. Todo ello conlleva el que se hayan tenido que añadir dichas variantes o formas nuevas no previstas dentro de la gramática.

Además de los errores anteriormente citados, nos topamos con una serie de artículos de naturaleza muy especial, que ofrecen una información muy detallada sobre el comportamiento gramatical de la entrada en cuestión. Dichas entradas habrán de ser tratadas por una gramática específica para las mismas.

Tras un examen de la gramática que nos permita obtener mejores resultados, nuestro siguiente paso es construir las herramientas necesarias para lograr el análisis correcto de las entradas del diccionario que han quedado por analizar de forma semiautomática. Esta herramienta será la base para el diseño e implementación de un sistema más general de ayuda al lexicógrafo, pudiendo servir tanto en la construcción de diccionarios como en la actualización de diccionarios similares al que hemos estudiado.

Agradecimientos

Damos las gracias al autor de la obra Ibon Sarasola por su consentimineto y receptibilidad para con el trabajo aquí descrito.

Así mismo, no se puede olvidar a UZEI, entidad que ha facilitado la utilización del diccionario en soporte magnético.

Este trabajo se ha podido llevar a cabo gracias a una beca de investigación concedida por el Departamento de Educación, Universidades e Investigación del Gobierno Vasco.

[4] Todavía no disponemos del último tomo.

Bibliografia

Ageno, A. et al. 1992,. From LDB to LKB. ESPRIT BRA-3030 Acquilex wp nº 039.

Agirre, E. et al. 1992. XUXEN: A Spelling Checker/Corrector for Basque Based on Two-Level Morphology. in Proceedings of the third conference on Applied Natural Language Processing, Trento, Italy.

Agirre E., Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., Insausti J.M., Sarasola K.. *"Different issues in the design of a general-purpose Lexical Database for Basque"* First Workshop on Applications of Natural Language to Databases.(NLDB´95). Versailles. 1995.

Agirre, E. et al. Euskararen Datu-Base Lexikala (EDBL). UPV/ EHU / LSI / TR8-94.

Amsler, R. A Taxonomy for english Nouns and Verbs, in Proceedings of the 19th annual Meeting of the Association for Computational Linguistics, (ACL´81), pages 133-138, Stanford, California, 1981.

Boguraev, B. and Briscoe, T. (eds.). Computational Lexicography for Natural Language Processing. New York: Logman, 1989.

Boguraev, B. et al. (1991). Database models for Computational lexicography. Research Report RC 17120, IBM Research Center, Yorktown Heights, NY. Byrd, R. J. et al.. Tools and Methods for Computational Lexicography,Computational Linguistics, vol. 13, ns.3-4, 219-240. 1987.

Calzolari, N. and Picchi , E. (1986). A project for Bilingual Lexical Database System, Proceedings of the Second annual Conf. of the Centre for the New OED, University of Waterloo, Waterloo, Ontario, pp.79-82 Waterloo.

Castellón, Irene. Lexicografía Computacional: Adquisición automática de conocimiento léxico. Tesis Doctoral, Universidad de Barcelona, 1993.

Martí , M.A. and Castellón, I. Gramática para el análisis del diccionario VOX. Boletín SEPLN, 10, 123-143. 190.

Sarasola, I. Hauta-Lanerako Euskal Hiztegia. G.K. Donostia,1984 / 1995.