Análisis de la correferencia para su anotación en un corpus en euskara

*CEBERIO Klara **ADURIZ Itziar, *DÍAZ DE ILARRAZA Arantza ***GARCÍA AZKOAGA Inés

*Euskal Herriko Unibertsitatea EHU-UPV. Grupo IXA.
Informatika Fakultatea
Manuel Lardizabal 1, 20018 Donostia
{klara.ceberio}{jipdisaa}@ehu.es

**Universitat de Barcelona UB. Departament de Lingüística General.
Universitat de Barcelona, Fac. Filologia
Gran Via de les Corts Catalanes, 585. 08007 Barcelona
itziar.aduriz@ub.edu

***Euskal Herriko Unibertsitatea EHU-UPV. Departamento de Filología Vasca.
Irakasleen Unibertsitate Eskola
Oñati Plaza 2, 20018 Donostia
ines.garciaazkoaga@ehu.es

Resumen:

En este artículo presentamos la segunda fase del estudio que comenzamos hace un par años (Aduriz et al. 2007). Entre las aportaciones que hacemos, están la ampliación del ámbito de estudio, de la anáfora pronominal al estudio de la correferencia, siendo el euskara el idioma en que se basa el estudio (García Azkoaga 2004). Por otra parte, desde el punto de vista computacional, hemos hecho la elección de una aplicación que facilitará el etiquetado de las relaciones correferenciales (Müller & Strube 2003).

Palabras clave: correferencia, anáfora, etiquetado del corpus, Procesamiento del Lenguaje Natural (PLN).

Laburpena:

Hemen aurkezten dugun artikulu honetan jasotzen da orain urte batzuk hasitako (Aduriz et al. 2007) azterketaren bigarren fasea. Ekarpen nagusiek bi alderdi hartzen dituzte: hizkuntzalaritzaren alderdi teorikotik aztergaia zabaldu dugu anafora pronominala eta honen aurrekaria aztertzetik, erreferentziakidetasuna zentzu zabalago batean aztertzera, betiere, euskara abiapuntutzat hartuta (Garcia Azkoaga 2004). Eta bestetik, konputazioaren alderditik, euskarazko corpusa aztergaiko elementuekin markatzeko ala etiketatzeko behar zen tresna egokia lortu dugu, etiketatze hori erraztu eta azkartu aldera (Müller & Strube 2003).

Hitz gakoak: korreferentzia, anafora, corpusaren etiketatzea, Lengoaia Naturalaren Prozesamendu Automatikoa (LNP).

Abstract:

In this paper we present the second stage of the study we began few years ago (Aduriz et al. 2007). With this work, we make two contributions: on the one hand, we have extended the topic from the pronominal anaphora to the study of coreference, working mainly on the Basque language (García Azkoaga 2004). On the other hand, we have chosen an application (Müller & Strube 2003) in order to make easier the annotation process of coreferential chains.

Key words: coreference, anaphora, corpus annotation, Natural Language Processing (NLP).

Tabla de contenidos:

1. Introducción

En estos últimos años la ciencia y la técnica han avanzado a grandes pasos y con ello han surgido nuevos ámbitos de utilización y análisis para el euskara, sobre todo, en las áreas en que el lenguaje y las tecnologías se unen.

Somos usuarios del idioma, pero hay muchas maneras de utilizar el idioma, y gracias a las tecnologías de la comunicación e información, tenemos medios antes inimaginables para que el euskara, tanto hablado como el escrito, pueda utilizarse más allá de sus límites geográficos.

Los magnetófonos y casetes que se utilizaban para grabar la voz en una época han sido sustituidos por instrumentos digitales e informáticos; para hablar por teléfono no es necesario estar conectados a un cable, es más, haciendo una llamada telefónica nos encontramos con que nos responde una máquina que habla como si de un ser humano se tratara; escribimos con la ayuda de un ordenador, incluso se puede hacer trabajar al ordenador mediante la voz; intentamos romper las barreras del idioma mediante la traducción automática, etc. A consecuencia de todo esto, la lingüística ha emprendido varios caminos, y en esta sociedad de la ciencia y de la información, el lenguaje se ha unido con la tecnología de una manera totalmente natural. Al fin y al cabo, ambos no están tan alejados, y se confluyen en ámbitos como el que nos ocupa en este trabajo, el del Procesamiento del Lenguaje Natural (PLN).

La lingüística y la informática se unen aquí para tratar de que la máquina (en este caso el ordenador) comprenda el lenguaje humano. El lingüista reúne, analiza y detalla los datos del idioma a tratar y los describe y organiza de manera que sean útiles para los programas. Es decir, prepara la información lingüística para poder procesarla automáticamente. Los informáticos son los creadores de los programas, que utilizan la información lingüística para procesar textos, con el objetivo de crear analizadores que comprendan el texto o para crear otro tipo de aplicaciones. Por ejemplo, si el lingüista compone una gramática basada en reglas, el informático desarrollaría un analizador sintáctico que procesara esa información, y a partir de ahí crearía una aplicación como por ejemplo, un corrector gramatical.

Por lo tanto, las máquinas y el ser humano pueden interactuar gracias a la información preparada y organizada por expertos. Esta interacción puede darse a cualquier nivel del lenguaje: tanto a nivel morfológico, sintáctico, semántico como textual. En todos estos niveles, si se quiere obtener un análisis completo y sólido, es indispensable partir de la descripción del elemento que es objeto de análisis. Así, por ejemplo, a nivel

morfológico, obtendríamos la información tanto de la categoría y la subcategoría lexical de todos los elementos, como de la descripción de las posposiciones[1]. A nivel sintáctico se reconocerían automáticamente las estructuras sintácticas, desde los sintagmas hasta las oraciones, y dependiendo del objetivo, podrían ser asignadas también las dependencias gramaticales[2]. En cuanto al nivel semántico, a través del tratamiento automático del lenguaje, junto con la desambiguación de significados, podríamos tener conocimiento de los roles semánticos y de los predicados. Finalmente, y en lo que concierne a este trabajo, a nivel textual, a través de las marcas lingüísticas que dejan en el texto, podemos analizar las redes referenciales y anafóricas que entretejen el discurso y que ayudan a que la información progrese de forma comprensible para el receptor.

En el área del Procesamiento del Lenguaje Natural en euskara, durante los últimos años se están desarrollando numerosos trabajos en los niveles mencionados anteriormente, sobre todo en el morfológico y en el sintáctico (Urkia 1997; Gojenola 2000); respecto al análisis semántico, podemos decir que está siendo muy tratado tanto en euskara como en otras lenguas (Pociello 2007). En cuanto a la lengua vasca, el mayor vacío se encuentra, a nuestro entender, en el análisis de las redes referenciales y anáforicas, pues todavía no se ha realizado ningún trabajo en este campo, y no hay ningún corpus que tenga etiquetadas estas redes. Se pretende rellenar en parte ese vacío a partir del trabajo en común que comenzó hace un par de años dentro de la EHU-UPV, entre el Departamento de Filología Vasca y el Grupo IXA.

En este trabajo hablaremos primeramente de la relación existente entre la correferencia y la anáfora para explicar luego, brevemente, el estado actual de los estudios relacionados con el etiquetado de la anáfora y la correferencia en el ámbito del Procesamiento del Lenguaje Natural. A continuación, explicaremos cuáles son las diferentes herramientas de etiquetado que existen y cuál ha sido nuestra elección. Después, hablaremos de las características del corpus que servirá como base a nuestro principal trabajo: el modelo de anotación utilizado para etiquetar ciertas expresiones correferenciales y anafóricas. Terminaremos con las conclusiones y las perspectivas del trabajo futuro.

2. La correferencia y la anáfora

El hablar de correferencia o el querer definir ese término, nos lleva necesariamente a explicar previamente la relación entre la anáfora y la correferencia.

Un referente puede ser recuperado a lo largo del texto por medio de expresiones muy diversas. Como explicaremos seguidamente, entre un referente y la expresión que lo retoma puede haber una relación anafórica, como es el caso de las anáforas asociativas, pero ello no implica que los elementos implicados hayan de ser necesariamente correferentes. Igualmente, el hecho de que dos elementos sean correferentes, como en el caso de los nombres propios por ejemplo, no implica necesariamente que entre ellos haya una relación anafórica. Tenemos además elipsis anafóricas, o incluso referentes evolutivos que pueden estar anafóricamente relacionados sin que ello suponga una relación de identidad con el referente original. Veamos los siguientes ejemplos:

[1] Las posposiciones en euskara son equiparables a las preposiciones en castellano.
[2] Se denomina a la gramática basada en dependencias, no en constituyentes.

(1) *Istripu bat* egon zen... *Anbulantzia* iritsi zenean...

(Hubo un accidente… Cuando llegó la ambulancia…)

(2) Pablok bost lehoi hil ditu eta nik hiru (Kleiber 1994)

(Pablo ha matado *seis leones*, y yo *tres*)

En (1) el sintagma *istripu bat* (un accidente) tiene un referente y *Anbulantzia* (la ambulancia) otro distinto, por lo tanto, las expresiones no son correferentes. En el ejemplo (2) los leones que se mencionan al principio tampoco son los mismos que se mencionan más tarde, aunque la forma de designarlos coincida. Según Milner (1982), a estos casos se les denomina correferencias virtuales, y en el primer caso estaríamos ante una anáfora asociativa y en el segundo caso, ante una anáfora lexical o nominal (Kleiber 1994).

En el caso de la elipsis anafórica (también llamada anáfora cero) es también cuestionable si se trata de correferencia o hay además una relación anáforica.

(3) *Mirenek* sagarrak bildu ditu. Bihar merkatura eramango ditu *Ø Ø*.

(*Miren* ha recogido manzanas. *Ø* Mañana *(las)* llevará al mercado)

Si nuestro punto de partida es considerar la anáfora como una retoma referencial, la elipsis no sería anafórica, pero desde una perspectiva más amplia, es posible hacer una interpretación anafórica de la misma (Charolles 1991) si consideramos que para su interpretación referencial debemos recurrir a un elemento presente en el cotexto.

(4) "Jakingo duzunez, *ura* ez da beti puru-purua izaten; hainbat hondakin eduki ditzake disoluzioan edo partikula solidoak eraman. Batzuetan *Ø* oso zikina izan daiteke (euri zaparrada baten ondorioz, putzu batean, etab.) eta beste batzuetan badirudi *Ø* garden edo garbi dagoela". (Natur Zientziak. "Ostadar" Proiektua, Elkar-G.I.E., DBH-1, 1996, 85. or.; García Azkoagan, 1999)

("Como es sabido, *el agua* no siempre es del todo pura, se pueden encontrar diferentes residuos en la disolución o llevar partículas sólidas. A veces *Ø* puede ser muy sucia (a raíz de una tromba de agua, en un charco, etc.) y otras parece que *Ø* está transparente o limpia".) (Natur Zientziak. "Ostadar" Proiektua, Elkar-G.I.E., DBH-1, 1996, 85. or.; García Azkoagan, 1999)

Asimismo, puede que las expresiones sean correferentes pero que no se dé ningún tipo de relación anafórica. Esto ocurre en el caso de los nombres propios, que al designar directamente, no necesitan ser interpretados a través de un antecedente.

(5) *Mikel* eta *Andoni* Gasteizko jaietara joan dira. *Mikel* goiz itzuli da etxera baina *Andoni* ez da agertu oraindik.

(*Mikel* y *Andoni* han ido a las fiestas de Gasteiz. *Mikel* ha vuelto pronto a casa pero *Andoni*, aún no ha aparecido.)

Por otro lado, guiándonos por las palabras de Kleiber (1988 y 1994), si a la hora de identificar la anáfora tenemos en cuenta el criterio del contexto lingüístico, no se admitirán como anáforicas correferencias de este tipo:

(6) *Ibarretxek* bere agintaldiaren urte bukaerako lehendabiziko diskurtsoa irakurri zuen. *EEAko lehendakariak* esan zuenez ...

(*Ibarretxe* leyó el primer discurso de fin de año de su mandato. Según *el presidente de la Comunidad Autónoma*…)

Las dos expresiones (*Ibarretxe* y *el presidente de la Comunidad Autónoma*) tienen el mismo referente, pero a su vez, son expresiones independientes y las interpretamos directamente. La palabra *lehendakari* (presidente) hace referencia a una persona concreta, y gracias al conocimiento compartido del mundo, no necesitamos añadir ninguna información para relacionarlo con *Ibarretxe*.
Las dudas que pueden surgir en torno a la correferencia, quedan patentes en el caso de los referentes evolutivos:

(7) *Arratoitxoa* begien itxi-ireki batean *adats ilegorridun eta begi distiratsudun neskatxa* bilakatu zen... *Neska gazteak* ibiltzeari ekin zion basotik irteteko asmoz... (Eguzkia baino ahaltsuagoa. Ilargi Erditxoaren ipuinak, 1993)

(*El ratoncito* se convirtió *en una chica pelirroja y con ojos brillantes… la joven mujer* comenzó a andar con la intención de salir del bosque…) (Eguzkia baino ahaltsuagoa. Ilargi Erditxoaren ipuinak, 1993)

Es muy difícil establecer los límites cuando se quieren reflejar los cambios físicos que se dan en el transcurso del tiempo o los cambios que se producen con la transformación de la materia (naturales o inducidos). El *ratoncito* del ejemplo, se convierte en una *chica pelirroja y con ojos brillantes*, pero sin duda, los referentes de ambas expresiones son muy diferentes. Eso no ocurre, sin embargo, cuando se trata de un relato sobre una persona, en estos casos, puede haber una evolución (por ejemplo: *niño* → *joven* → *hombre*), de forma que es un mismo referente el que se retoma mediante diferentes hipónimos. Así pues, correferencia y anáfora no siempre coinciden.

Entre las anáforas que son a su vez correferenciales nos encontramos con distintos tipos de pronombres: personales, demostrativos, posesivos. A esta lista podemos añadir otro tipo de unidades lingüísticas tales como algunos adverbios de lugar que necesitan ser interpretados por medio de un antecedente:

(8) *Koba-zuloan* sartu ginen. *Barruan* oso ilun zegoen dena eta hango isiltasuna beldurgarria zen.

(Entramos *en la cueva*. *Dentro* estaba todo muy oscuro y el silencio que reinaba allí era aterrador).

Retomando el caso de los pronombres, en el siguiente ejemplo la referencia del pronombre es dudosa, como subraya Zabala (1996), ya que puede hacer referencia tanto a *Andoni*, como a otra persona. Si le hiciera referencia al elemento dentro de la oración, en este caso a *Andoni*, se trataría de una utilización reflexiva del pronombre. El funcionamiento de este tipo de pronombres se explica mediante la gramática; es el caso de las anáforas ligadas. El elemento B necesita un antecedente (elemento A) en la oración. A y B son correferenciales y están ligados. El funcionamiento de la anáfora se limita en la oración y la relación entre los elementos es gramatical porque está condicionada por la sintaxis y la semántica.

(9) Andoni *bere* etxera eraman nuen.

(Le llevé a Andoni a *su* casa)

En el ejemplo que viene a continuación, en cambio, el funcionamiento de los pronombres es diferente, el pronombre y el antecedente son correferenciales pero no se encuentran en la misma oración.

(10) *Ozono geruza* oso garrantzitsua da lurraren bizitzarako. *Bera/hura* da erradiazio kaltegarrietatik babesten gaituen filtro naturala.

(*La capa de ozono* es muy importante para la vida en la tierra. (*Ella*) Es el filtro natural que nos protege de las radiaciones peligrosas.)

El pronombre anafórico, siendo un elemento semánticamente incompleto necesita como referente algún elemento de la oración y la relación entre ambos es lingüística. A pesar de todo, se tienen en cuenta criterios pragmáticos a la hora de elegir el antecedente correcto.

En cualquier caso, en opinión de algunos autores esta distinción entre ligada/libre no tiene gran valor práctico, pues según Kempson (1986) la característica que pone en evidencia el carácter pragmático de la anáfora libre, está también presente en las anáforas ligadas.

Por otro lado, en euskara la utilización de los pronombres no es tan necesaria, al contrario que en francés y en inglés, ya que la flexión del verbo puede marcar el sujeto, objeto directo e indirecto, como podemos observar en el siguiente ejemplo:

(11) *Ura* oso garrantzitsua da gure bizitzan. *Ø* Gure gorputzaren osagairik nagusiena da.

(*El agua* es muy importante en nuestra vida. *Ø* Es el elemento más importante de nuestro cuerpo).

Si nos centramos en las composiciones libres no sólo tenemos las anáforas pronominales, sino también las nominales, y entre ellas, podemos hacer distinción entre anáforas fieles y no fieles. La anáfora fiel, en el estricto sentido de la palabra se basa en la recuperación léxico-sintáctica del antecedente, y en este caso el anaforizante puede tomar las siguientes formas:

- Repetición del mismo término (puede cambiar la marca de la declinación):

  (12) ...*haurtzaindegiak* hazkuntzarako gune dira... *haurtzaindegietako* profesionalen lana ez da erraza...

  (… *las guarderías* son un centro de educación…el trabajo de los profesionales *de la guardería* no es fácil…)

- Recuperación del sintagma nominal indefinido (sustantivo + determ. indefinido) mediante un sintagma nominal con un determinante definido e incluso con la aparición o la omisión de un elemento atributivo (‘sustantivo+ (elemento atributivo)+ (determinante indefinido)’ → ‘sustantivo+ (elemento atributivo) + determ. definido –artículo, demostrativo-)

  (13) ...*hiztegi berri bat* argitaratu dute... *hiztegiak / hiztegi honek / hiztegi berri hauek*...

  (… han publicado un diccionario nuevo… los diccionarios… / este diccionario / estos nuevos diccionarios…)

  Cuando la retoma se realiza mediante el demostrativo, estaríamos, ante una referenciación deíctica.

En el caso de las anáforas no fieles, el lexema del elemento anafórico y el del antecedente son diferentes. En este caso, nos encontramos con anáforas conceptuales que pueden resumir el contenido del antecedente o incluso hacer valoraciones sobre el mismo:

(14) [...] Baina benetako damua gero etorri zitzaigun gogora, *bat-batean atea itxi zenean haize bolada handi bat* eragin zuen. *Egoera beldurgarri hura* artean sinestezina zen!! [...] (Garcia Azkoaga, 2004: NE-DBH2-5)

([…] Pero el arrepentimiento verdadero vino más tarde, *cuando se cerró la puerta repentinamente y se produjo un gran viento*. ¡*Esa terrible situación* era increíble! […])(Garcia Azkoaga, 2004: NE-DBH2-5)

Una vez analizados los tipos de correferencia y anáfora hemos observado que las herramientas informáticas de las que disponemos hoy en día no nos permiten aún profundizar en el análisis y etiquetado de las expresiones que no son correferenciales, por lo que hemos limitado nuestro trabajo al ámbito de la correferencia, sea o no sea anafórica. Así, las expresiones a etiquetar serán las siguientes: nombres propios; pronombres personales, demostrativos y posesivos, anáforas fieles y adverbios anafóricos (v. 6.2).

3. Corpus etiquetados correferencialmente

Hace ya unos cuantos años en los que el análisis de la anáfora y la correferencialidad son objeto de estudio en el área del Procesamiento del Lenguaje Natural (PLN). Para poder desarrollar una herramienta sólida es imprescindible la existencia de un corpus etiquetado (Mitkov 2002) como primer paso del proceso.

Encontramos en la literatura numerosas referencias bibliográficas de corpus que han sido etiquetados con las correspondientes relaciones correferenciales y anafóricas. Algunos de los que han sido anotados en habla inglesa son: The Lancaster Anaphoric Treebank (UCREL) (Garside et al. 1997), the MUC Coreference Task (MUC-7) (Hirschman 1997), el corpus de la Universidad de Wolverhampton (Mitkov, 2000), parte del Penn Treebank Corpus (Ge 1998), DRAMA scheme (Passoneau and Litman 1997) y el MATE/GNOME scheme (Poesio 2004).

Asimismo hemos consultado los recursos utilizados para otros idiomas, tales como TIGER Project (Kunz & Hansen-Schirra 2003) para el alemán, o el trabajo llevado a cabo en la Universidad de Praga (Hajič & Urešová 2004), donde el corpus ha sido anotado a nivel pragmático, incluyendo elementos correferenciales.

Por último, hemos de mencionar el estudio realizado para el idioma castellano (Navarro et al. 2003). En este corpus también se ha trabajado a nivel pragmático con la ayuda de una herramienta para la anotación. Con esta aplicación se marcan las relaciones anafóricas y correferenciales (incluyendo la elipsis), así como sus correspondientes referentes.

4. Herramientas para el etiquetado

Este trabajo es un paso más en el proceso que habíamos comenzado en el ámbito de la detección de la anáfora. En una primera fase, nos centramos en un único tipo de anáfora, realizamos la anotación de la anáfora pronominal. Esta vez hemos querido ampliar el objeto de estudio, pasando de la anáfora pronominal al ámbito de la correferencia, marcando así redes referenciales y anafóricas más amplias. De esta manera pretendemos alcanzar una comprensión más amplia del texto; reconocer más elementos del sistema referencial del texto para así poder entender mejor el texto automáticamente.

En el etiquetado realizado hasta ahora observamos que nos sería de gran utilidad alguna aplicación que se haya desarrollado para este fin. De este modo hemos elegido una herramienta que nos facilitará esta anotación.

A la hora de elegir la aplicación más adecuada para llevar a cabo nuestro trabajo hemos tenido en cuenta algunas características tales como: el nivel de etiquetado que ofrecen, la adecuación a nuestras herramientas, formatos utilizados, etc.

Los más destacables entre los que hemos estudiado son los que mencionaremos en este apartado.
La herramienta desarrollada en la universidad de Wolverhampton, ClinkA (Orasan 2000). Se presenta como una aplicación robusta, siguen el modelo de etiquetado que se propone MUC-7 Coreference Task Definition (Hirschman & Chinchor 1997) que se aleja de nuestro modelo.

Otra aplicación muy interesante es el anotador general Alembic Workbench. Entre otros niveles, se presenta la posibilidad de marcar la correferencialidad. Una de las ventajas que tenía era la posibilidad de ampliar las etiquetas, pero, estas etiquetas se incluyen directamente en el texto, resultando un poco confuso.

En el marco del MATE Workbench (Dybkjær and Bernsen 2000) han creado una herramienta de gran capacidad. En teoría hubiera sido una plataforma ideal a para nuestra labor, pero después de haber leído experiencias negativas al tratar con textos extensos, decidimos no trabajar con esta aplicación.

Terminaremos esta enumeración de herramientas nombrando la aplicación MMAX (Müller and Strube 2003). Además de ser ligera y de fácil manejo, presenta la ventaja de poder adecuarlo a las necesidades del usuario. Por otro lado, utiliza el sistema 'stand-off' para acumular información, es decir, los datos fundamentales (el texto mismo) se guardan en un archivo y la información de segundo nivel (la información gramatical y textual) en otro archivo diferente. Además, pueden participar más de uno en el proceso de etiquetado y hay la posibilidad de analizar el acuerdo entre los anotadores. Por todas estas características es esta última aplicación la que hemos elegido para nuestra anotación.

5. El corpus

El corpus EPEC (Aduriz et al. 2006) es el corpus en el que nos hemos basado para este trabajo. Este corpus surgió dentro del proyecto 3LB, junto con el del catalán y el del castellano (Palomar et al. 2004). El objetivo de este proyecto era la anotación sintáctica y semántica del corpus. En cuanto a la parte del euskara, se etiquetaron 50.000 palabras sintácticamente, utilizando el sistema de anotación basado en dependencias (Aranzabe et al. 2003).

En este apartado explicaremos el proceso de análisis modular (Aduriz et al. 2006), para ver qué información se le añade en cada paso, antes de anotar las relaciones correferenciales con sus antecedentes correspondientes.

Primeramente el corpus es etiquetado automáticamente mediante el analizador morfológico *Morfeus* (Aduriz et al. 1998), el cual analiza todas las palabras por separado sin tener en cuenta el contexto. Después de este primer proceso, todas las palabras tendrán la información morfosintáctica que les corresponde: la categoría gramatical, la subcategoría, información del número y si son definidos o indefinidos, el caso de declinación, y la mayoría de las veces la información sobre su función sintáctica. El siguiente ejemplo nos muestra un ejemplo del analizador morfosintáctico: "*Udaberrian hegazti ugari pasatzen da gure mendien eta herrien gainetik."* (En primavera pasan muchas aves sobre nuestros montes y pueblos):

(15)

```
/<Udaberrian>/<HAS_MAI>/                                    (En primavera)
("udaberri" IZE ARR DEK NUMS MUGM DEK INE @ADLG)
/<hegazti>/                                                 (aves)
("hegazti" IZE ARR DEK ABS MG @OBJ @SUBJ @PRED)
("hegazti" IZE ARR @KM>)
/<ugari>/                                                   (muchas)
("ugari" ADJ IZO DEK ABS MG @OBJ @SUBJ @PRED)
("ugari" ADJ IZO @<IA)
("ugari" DET DZG MG DEK ABS MG @SUBJ)
("ugari" DET DZG MG @ID>)
("ugaritu" ADI SIN AMM ADOIN @-JADNAG)
/<pasatzen>/                                                (pasar vb. ppal.)
("pasatu" ADI SIN AMM ADIZE DEK INE @OBJ @-JADNAG_MP_OBJ)
("pasatu" ADI SIN AMM ADOIN ASP EZBU @-JADNAG)
/<da>/                                                      (pasar vb. auxiliar)
("izan" ADL A1 NR_HU @+JADLAG)
("izan" ADT A1 NR_HU @+JADNAG)
/<gure>/                                                    (nuestros)
("gu" IOR PERARR NUMP GU DEK GEN DEK ABS MG @IZLG>
@<IZLG @OBJ @SUBJ @PRED)
("gu" IOR PERARR NUMP GU DEK GEN @IZLG>)
("guretu" ADI SIN AMM ADOIN @-JADNAG)
/<mendien>/                                                 (montes)
("mendi" IZE ARR DEK GEN NUMP MUGM DEK ABS MG @IZLG>
@<IZLG @OBJ @SUBJ @PRED)
("mendi" IZE ARR DEK GEN NUMP MUGM @IZLG>)
/<eta>/                                                     (y)
("eta" LOT JNT EMEN @PJ)
("eta" LOT MEN KAUS @PJ)
/<herrien>/                                                 (pueblos)
("herri" IZE ARR DEK GEN NUMP MUGM DEK ABS MG @IZLG>
@<IZLG @OBJ @SUBJ @PRED)
("herri" IZE ARR DEK GEN NUMP MUGM @IZLG>)
/<gainetik>/                                                (sobre)
("gain" IZE ARR DEK NUMS MUGM DEK ABL @ADLG)
/<.>/<PUNT_PUNT>/
```

En esta primera fase el mayor problema es el de la ambigüedad, ya que el análisis está fuera de contexto muchas palabras pueden resultar ambiguas, bien sea por el léxico, bien sea por la función sintáctica.

El proceso de desambiguación se realiza mediante otro módulo llamado EUSTAGGER. La función principal de este analizador es reducir la ambigüedad morfosintáctica por lo tanto, desambiguar. Así, en cada contexto elegirá la mejor opción. Lo observamos con el ejemplo anterior:

(16)

```
/<Udaberrian>/<HAS_MAI>/                    (En primavera)
("udaberri"  IZE ARR DEK NUMS MUGM DEK INE  @ADLG)
/<hegazti>/                                 (aves)
("hegazti"  IZE ARR  @KM>)
/<ugari>/                                   (muchas)
("ugari"  DET DZG MG DEK ABS MG  @SUBJ)
/<pasatzen>/                                (pasar vb. ppal.)
("pasatu"  ADI SIN AMM ADOIN ASP EZBU  @-JADNAG)
/<da>/                                      (pasar vb. auxiliar)
("izan"  ADL A1 NR_HU  @+JADLAG)
/<gure>/                                    (nuestros)
("gu"  IOR PERARR NUMP GU DEK GEN  @IZLG>)
/<mendien>/                                 (montes)
("mendi"  IZE ARR DEK GEN NUMP MUGM DEK ABS MG  @IZLG>
@<IZLG @OBJ @SUBJ @PRED)
("mendi"  IZE ARR DEK GEN NUMP MUGM  @IZLG>)
/<eta>/                                     (y)
("eta"  LOT JNT EMEN  @PJ)
/<herrien>/                                 (pueblos)
("herri"  IZE ARR DEK GEN NUMP MUGM  @IZLG>)
/<gainetik>/                                (sobre)
("gain"  IZE ARR DEK NUMS MUGM DEK ABL  @ADLG)
/<.>/<PUNT_PUNT>/
```

En este punto del análisis tenemos el corpus analizado morfológicamente, asignados las principales funciones lingüísticas, y casi totalmente desambiguado. Para terminar con el análisis aplica el *chunker*[3]. Este módulo define los sintagmas, entre ellos las estructuras sintácticas básicas, tales como entidades (nombres propios), posposiciones complejas y los sintagmas nominales o verbales:

(17)

```
"<Udaberrian>"<HAS_MAI>"                    (En primavera)
"udaberri"  IZE ARR DEK NUMS MUGM DEK INE  @ADLG  HAS_MAI
%SINT
"<hegazti>"                                 (ave)
"hegazti"  IZE ARR  @KM>  %SIH
"<ugari>"                                   (muchas)
"ugari"  DET DZG MG DEK ABS MG  @SUBJ %SIB
"<pasatzen>"                                (pasar vb. ppal.)
"pasatu"  ADI SIN AMM ADOIN ASP EZBU  @-JADNAG  NOTDEK
%ADIKATHAS
"<da>"                                      (pasar vb auxiliar)
"izan"  ADL A1 NOR NR_HU  @+JADLAG %ADIKATBU
"<gure>"                                    (nuestro)
"gu"  IOR PERARR NUMP GU DEK GEN  @IZLG> %SIH
"<mendien>"                                 (montes)
"mendi"  IZE ARR DEK GEN NUMP MUGM  @IZLG>
"<eta>"                                     (y)
"eta"  LOT JNT EMEN  @PJ  AORG
"<herrien>"                                 (pueblos)
"herri"  IZE ARR DEK GEN NUMP MUGM  @IZLG>
"<gainetik>"                                (sobre)
"gain"  IZE ARR DEK NUMS MUGM DEK ABL  @ADLG %POS %SIB
```

[3] *Chunker:* término utilizado para denominar el fragmentador de sintagmas o el divisor sintagmático.

"<$.>"<PUNT_PUNT>"
PUNT_PUNT

Hemos observado que tanto el nivel morfológico como el sintáctico se han realizado automáticamente. El etiquetado del nivel textual, lo realizaremos manualmente con la ayuda de la aplicación que ya hemos mencionado.

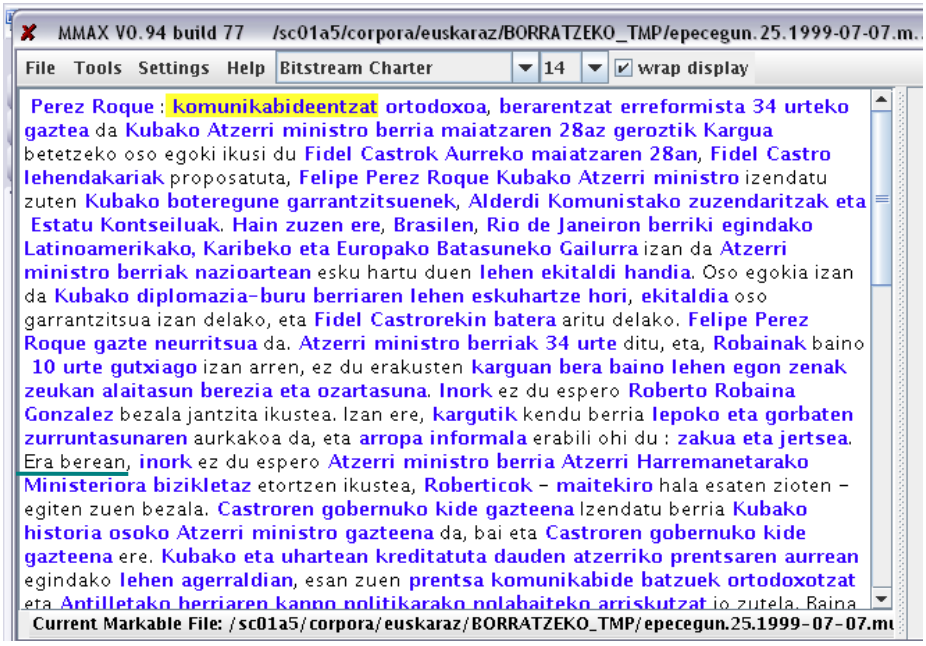
6. La anotación de la correferencia

Partiremos del corpus analizado, estructurado, que aparte de facilitarnos la anotación, nos proporcionará la información lingüística que nos servirá para sacar conclusiones pertinentes sobre el tema de la correferencia.

Al igual que cuando etiquetamos las anáforas pronominales (Aduriz eta al. 2007) hemos puesto especial atención en los sintagmas no verbales. Hemos ampliado el campo de estudio, aparte de marcar los determinantes demostrativos que en euskara cumplen a veces la función de pronombre anafórico, también nos fijaremos en los sintagmas nominales que expresen correferencia (v. 6.1.1). Para esta anotación contaremos con la aplicación antes mencionada MMAX (Müller & Strube 2003).

6.1. La aplicación MMAX

En este punto trataremos de explicar el funcionamiento de la aplicación MMAX. Ya hemos comentado que es la herramienta que nos facilitará la anotación, ya que obtendremos los textos ya etiquetados, con los sintagmas nominales marcados como observamos en el siguiente ejemplo:



(Fig.1: Ejemplo de un sintagma, en sombreado)

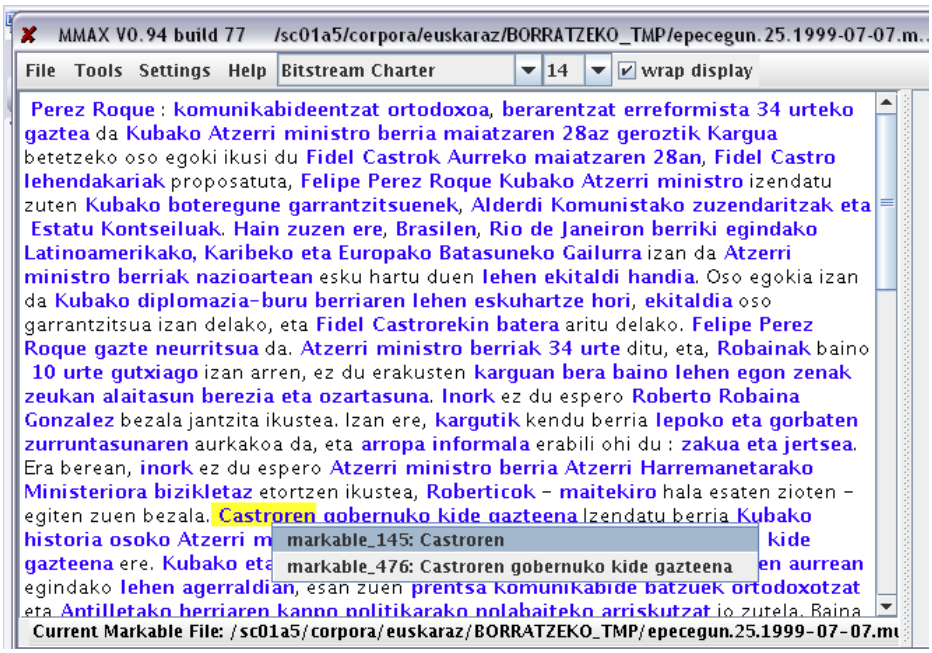
La aplicación hace uso de los colores para marcar las características del texto (puede que no se aprecie debidamente), por ejemplo, el texto marcado aparecerá en negrita y color (azul en este caso), son los que pueden ser correferenciales. El texto que esté sombreado (en color amarillo) es el que pertenece a un mismo sintagma.

6.1.1. El concepto de ‘marcable’

En la primera fase de la anotación hemos tenido que verificar la corrección de todos los sintagmas. Cuando nos referimos a ‘marcable’, estamos hablando de cualquier sintagma nominal, que tenga la posibilidad de ser correferencial, anafórico o antecedente.

El corpus estará marcado ya con sus correspondientes sintagmas nominales, pero no siempre correctamente. Es por ello que como fase anterior al etiquetado de las redes referenciales, hemos puesto especial hincapié en este punto. Como ejemplo tenemos la locución *Era berean* (asimismo, v. Fig. 1), que estaría marcada anteriormente automáticamente, los anotadores tendrán que desecharlo como posible ‘marcable’, ya que en este texto no pertenece a ninguno de los tipos de correferencia o anáfora que hemos decidido etiquetar.

Por otro lado, nuestras herramientas tienen la capacidad de marcar el sintagma nominal completo, pero puede darse el caso en el que el sintagma nominal esté compuesto por más de un componente, como por ejemplo un genitivo, o una oración subordinada, que podría ser antecedente de algún elemento anafórico o correferencial. La aplicación facilita el etiquetado de este tipo de componentes.



(Fig. 2.: Componente genitivo de un sintagma)

Los anotadores se encargarán de marcar estos componentes, porque puede que más adelante sean antecedentes de algún elemento correferencial.

(18) [[*Castroren*] gobernuko kide gazteena] da [Perez Roque]. [*Hark*] [gobernuan] dirauen artean [[Atzerri Ministro] lanetan] arituko da.

([Perez Roque] era [el miembro más joven del gobierno [*de Castro*]]. Mientras [*él*] permanezca en el Gobierno seguirá como [Ministro de Asuntos Exteriores]).

6.2. Criterios de etiquetado

Una vez explicados algunos conceptos de la aplicación, comentaremos los criterios de etiquetado de este estudio. Estos criterios han podido establecerse gracias a la anotación hecha anteriormente (Aduriz et al. 2007), y también el análisis teórico-descriptivo que hemos realizado en este trabajo.

Es imprescindible que los criterios de etiquetado sean claros y exactos, teniendo en cuenta que para esta anotación se requerirá el trabajo de al menos dos personas, a fin de garantizar la calidad del trabajo. De todos modos, en este tipo de labores suele ser difícil llegar al acuerdo total entre los anotadores.

La anotación que se realizará en dos fases, primeramente verificando los 'marcables', es decir, los sintagmas nominales, que puedan ser tanto anafóricos como antecedentes de los mismos y en una segunda fase etiquetaremos las expresiones correferenciales con su antecedentes correspondientes. Las anáforas fieles' (v. 2) también serán consideradas en esta anotación.
En esta primera fase, consideraremos los sintagmas que tienen las siguientes características gramaticales:

1. Pronominales
    a. Pronombres (de 3ª persona)

    (19) *Anaia Nartxi* falta da eta *berak* asko zekien, asko laguntzen zuen.

    (Falta *mi hermano Nartxi*, y *él* sabía mucho, ayudaba mucho).

    b. Determinantes demostrativos, en función de pronombre.

    (20) Tentsioa handitu zen, eta gordeta zeuden *herritarrengan* izua nagusitu zen. *Horiek* UNAMETen egoitzan babesten saiatu ziren

    (La tensión aumentó, y se sembró el miedo entre los *habitantes* que estaban escondidos. *Ellos* intentaron guarecersen en la residencia UNAMET)

2. Repeticiones del mismo término:
    a. Sustantivo + (elem. atrib.) + artículo

    (21) *Errektorea* aukeratzeko behin betiko bozketa bihar egingo da, goizeko hamarretan hasita. Bihar arte ez da jakingo EHUko *errektore berria* nor den.

    *(*Mañana, desde las diez de la mañana se celebrará la votación para elegir *al rector*. No se podrá saber quién es el nuevo rector de la UPV hasta mañana*).*

    b. Sustantivo+ (elem. atrib.) + determinante demostrativo

    (22) 1993an *24 urpeko misil jaurtitzaile* zituen Txinak JL-I deritze *urpeko horiei*.

    (En 1993, China tenía *24 misiles lanzadores submarinos*, a *esos submarinos* les denominan JL-I)

    c. Nombre propio + (marca de declinación)

    (23) Kofi Annan Nazio Batuetako idazkari nagusiak gaitzetsi egin zituen atzo gertatutako bortizkeria ekintzak. Annanek berehala neurriak (…)

    (*El secretario general de las Naciones Unidas Kofi Annan* condenó los sucesos violentos sucedidos ayer. *Annan*, medidas urgentes…)

3. Adverbios (de lugar)

(24) Milaka herritar ihes egiten saiatu ziren eta *Diliko portuan* pilatu ziren. *Han* egon ziren zain (…)

(Miles de ciudadanos intentaron huir y se amontonaron en *el puerto de Dili*. *Allí* estuvieron esperando…)

Todos estos elementos pueden hacer referenciar algún componente del texto anterior, surgiendo así diferentes tipos de relación entre ellos. En el siguiente apartado explicaremos los criterios que hemos seguido para marcarlos.

6.2.1. Tipología de relaciones correferenciales

Una vez etiquetados los elementos como posibles componentes de relaciones correferenciales, hemos establecido una tipología de relaciones. La herramienta que hemos elegido para esta tarea nos facilita este proceso, ya que nos posibilita asociar los componentes de la correferencia y la anáfora de un modo fácil y claro. Toda esta información se guarda en el formato estándar 'xml', de manera que lo podremos utilizar en otras aplicaciones.

Entre las correferencias anáforicas etiquetaremos las expresiones pronominales y nominales, de éstas últimas nos centraremos sólo en las anáforas fieles. Por otro lado tendremos en cuenta los adverbios de lugar y los nombres propios que tengan valor anafórico. Conscientes de la variedad de relaciones anáforicas existentes, hemos decidido limitarnos a los antes mencionados, ya que con los otros tipos de anáforas nominales puede haber más dificultades a la hora de marcarlos automáticamente.

Explicamos brevemente las relaciones que ha de establecer el etiquetador:

- Pronominales: los pronombres personales, los determinantes demostrativos que cumplen función de pronombre, *elkar* (pronombre recíproco), *X-en burua* (pronombre reflexivo).

  (25) *Ura* oso garrantzitsua da gure bizitzan. *Hura* da gure gorputzaren osagairik nagusiena.

  (*El agua* es muy importante en nuestra vida. *Ø* Es el elemento más importante de nuestro cuerpo).

- Anáforas fieles: cuando se repite la palabra que tiene un mismo lexema, aunque vaya acompañada de un atributo u otro caso de declinación.

  (26) Igandean mendira igo zirenean, hango iturriko *ur freskoa* edan zuten. *Ur horrek* egingo zien kalte nonbait.

  (El domingo, cuando subieron al cielo, bebieron *agua fresca* de la fuente. Por lo visto *esa agua* les sentó mal).

- Adverbios de lugar: cuando hacen referencia a algún lugar que se menciona anteriormente.

  (27) *Koba-zuloan* sartu ginen. *Barruan* oso ilun zegoen dena eta hango isiltasuna beldurgarria zen.

  (Entramos *en la cueva*. *Dentro* estaba todo muy oscuro y el silencio era aterrador)

- Nombres propios: cuando aparece algún elemento que haga referencia al nombre propio.

(28) *Perez Roque* ... Kubako diplomazia-buru berria.... *Felipe Perez Roque*
(*Pérez Roque*… el nuevo jefe de diplomacia de Cuba… *Felipe Pérez Roque*)

7. Conclusiones y perspectivas futuras

En este artículo hemos analizado cómo se realizan las relaciones correferenciales, asimismo, con la intención de tratar estas relaciones de manera más productiva, hemos procedido a la segunda fase de la anotación de la correferencia en el corpus EPEC.

La aportación más significativa se da desde el punto de vista teórico, ya que hemos ampliado el ámbito de estudio, de la anáfora pronominal a otros casos de correferencia. Al ampliar el área de estudio teórico ha tenido su influencia en el lado práctico, que nos ha llevado a cambiar y adecuar los criterios de etiquetado.

Por otra parte, en lo que se refiere a la anotación del corpus, ha sido importante la elección de la aplicación, ya que una vez elegida ha habido una labor de integración de los textos a esta aplicación concreta. La información morfológica y sintáctica que se ha asignado de manera automática en los textos que hemos anotado, ha sido de gran ayuda en esta anotación, que servirá como base para un futuro etiquetado automático de este tipo de relaciones correferenciales.

Ciertamente, el trabajo presentado en este artículo se enmarca en un proyecto general que comprende la anotación lingüística del corpus en todos sus niveles: morfológico, sintáctico, sintáctico y textual (discursivo). Todo ello con el objetivo de propiciar un banco de pruebas adecuado que servirá para desarrollar herramientas idóneas con el fin de que una herramienta informática pueda llegar a comprender el texto plenamente.

Finalmente, la anotación de relaciones correferenciales servirá para crear una herramienta que efectúe la misma labor de manera automática y ésta podrá ser utilizada a su vez en otro tipo de aplicaciones tales como los sistemas de búsqueda de respuestas, sistemas de resumen automático incluso, sistemas de traducción automática.

8. Bibliografía

Aduriz I., Aranzabe M. J., Arriola J.M., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A. & Urizar R. (2006). "Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing". *Corpus Linguistics Around the World. Book series: Language and Computers. Vol 56 (1- 15 or.).* Ed. Andrew Wilson, Paul Rayson, and Dawn Archer. Rodopi. Netherlands.

Aduriz, I., Ceberio, K., Díaz de Ilarraza, A. (2007). "Pronominal Anaphora in Basque: Annotation issues for later computational treatment". *6th Discourse Anaphora and Anaphor Resolution Colloquium. DAARC2007,* Lagos Portugal.

Aldezabal I. (2004). *Aditz-azpikategorizazioaren azterketa sintaxi partzialetik sintaxi osorako bidean. 100 aditzen azterketa, Levin-en (1993) lana oinarri hartuta eta metodo automatikoak baliatuz*. Euskal Filologia Saila, Euskal Herriko Unibertsitatea.

Aranzabe M. J., Arriola J. M., Atutxa A., Balza I., Uria L. (2003). "Guía para la anotación sintáctica manual de Eus3LB (corpus del euskara anotado a nivel sintáctico, semántico y pragmático)". *UPV/EHU/LSI/TR-13*.

Charolles M. (1991). "L'Anaphore. Definition et classification des formes anaphoriques", *Verbum,* Tome XIV, 2-3-4, 203-216.

Corblin F. (1983). "Défini et démonstratif dans la reprise immédiate", *Le Français moderne, 51,* 118-133.

Dybkjær L. and Bernsen N. O. (2000). "The MATE Workbench", *Proceedings of the LREC'2000 workshop on Data Architectures and Software Support for Large Corpora*, Athens, 33-37 (a).

Garcia Azkoaga I.M. (1999). "Elementu anaforikoak eskolako testuetan", Fontes Linguae Vasconum, 82, 393-417.

Garcia Azkoaga I.M. (2004). *Kohesio anaforikoa hiru testu generotan. Adinaren araberako azterketa.* Bilbao, Euskal Herriko Unibertsitatea.

Garside R., Leech G. & McEnery A. (eds.) (1997). *Corpus Annotation. Linguistic Information from Computer Text Corpora.* London: Longman.

Gojenola K. (2000). *Euskararen sintaxi konputazionalerantz. Oinarrizko baliabideak eta beren aplikazioa aditzen azpikategorizazio-informazioaren erauzketan eta erroreen tratamenduan.* Donostia, Informatika Fakultatea, Euskal Herriko Unibertsitatea.

Hajič J. & Urešová Z., (2004). "The Prague Dependency Treebank", Presentación ante el Grupo IXA, Donostia.

Hirschman L. and Chinchor N. (1997) "MUC-7 coreference task definition". *In MUC-7 Proceedings, Science Applications International Corporation.*

Kempson R. (1986) "Definite NPs and Context-Dependence: a Unified Theory of Anaphora". In T. Hyers et alii (aerg.), *Reasoning and Discourse Processes, Academic Press*, London, 209-239.

Kleiber G. (1988). "Peut-on définir une catégorie générale de l'anaphore?". *Vox Romanica*, 47, 1-13.

Kleiber G. (1994). *Anaphores et pronoms.* Louvain-la-Neuve, Duculot.

Kunz K. & Hansen-Schirra S. (2003) "Coreference Annotation of the TIGER Treebank" In *Proceedings of the Workshop Treebanks and Linguistic Theories*. Växjö, Sweden.

Milner J.C. (1982). *Ordres et raisons de la langue.* Paris, Seuil.

Mitkov R. (2002). *Anaphora resolution.* London: Longman.

Müller C., Strube M. (2003). "Multi-Level Annotation in MMAX" In *Proc. of the 4th SIGDIAL.*, Sapporo, Japan. 4-5 July 2003, pp.198-207.

Navarro B., Civit M., Martí M. A., Marcos R., Fernández B. (2003). "Syntactic, semantic and pragmatic annotation in Cast3LB" In *Proceedings of the Shallow Processing of Large Corpora. A Corpus Linguistics Workshop.* Lancaster, UK.

Pociello E. (2007). *Euskararen ezgutza-base lexikala: Euskal WordNet.* Donostia, Euskal Herriko Unibertsitatea.

Orasan C. (2000). "CLinkA a Coreferential Links Annotator" in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000).* Athens, Greece.

Urkia M. (1997). *Euskal morfologiaren tratamendu informatikorantz.* Filologia eta Historia-Geografia Fakultatea, Euskal Herriko Unibertsitatea.

Zabala I. (1996). "Testu-lotura: lotura tematikoa eta erreferentzia-sareak testu teknikoetan". In *Zabala I. (koord.). Testu-loturarako baliabideak. Euskara Teknikoa*. Bilbo, Euskal Herriko Unibertsitatea, 15-44.

9. Glosario

%ADIKATBU: final del sintagma verbal
%ADIKATHAS: comienzo del sintagma verbal
%SIB: final del sintagma;
%SIH: comienzo del sintagma;
%SINT: sintagma de una sola palabra;
@+JADLAG: verbo auxiliar;
@+JADLAG_MP: verbo auxiliar subordinado
@+JADNAG: verbo principal conjugado;
@+JADNAG_MP: verbo principal conjugado subordinado;
@<IA: adjetivo a la derecha;
@<IZLG: adjetivo que va a la derecha;
@ADLG: complemento verbal;
@ID>: determinante a la izquierda;
@IZLG>: adjetivo a la izquierda;
@-JADNAG: verbo principal no conjugado;
@-JADNAG_MP_OBJ: verbo principal no conjugado subordinado con función de objeto directo
@KM>: modificador de la forma que lleva el caso
@OBJ: objeto directo;
@PJ: coordinación coordinativa;
@PRED: predicativo;
@SUBJ: sujeto;
ABL: ablativo;
ABS: absolutivo;
ABZ: adlativo de cercanía;
ADI: verbo;
ADIZE: verbo nominal;
ADJ: adjetivo;
ADL: verbo auxiliar;
ADOIN: raíz del verbo;
ADT: verbo sintético;
AMM: morfema de forma verbal;
AORG: marca del -a orgánico;
ARR: (nombre) común;
ASP: morfema de aspecto;
DEK: declinación;
DET: determinante;
DZG: determinante no definido;
EMEN: conjunción copulativa;
EZBU: marca de aspecto no perfectivo;
GEN: genitivo posesivo;
HAS_MAI: letra que comienza con mayúscula;

INE: inesivo;
IOR: pronombre personal;
IZE: nombre;
IZO: adjetivo que va después del nombre;
JNT: conjunción;
KAUS: causal;
LOT: elemento que se atribuye a las conjunciones y conectores.
MEN: subordinado;
MG: indefinido;
MUGM = M: definido;
NOTDEK: verbo que no tiene declinación;
NUMP = PL: P: número plural;
NUMS = S: número singular;
PERARR: pronombre personal común
PUNT_PUNT: punto;
SIN: adjetivo simple.