LA ANOTACIÓN DE LA REFERENCIA SOBRE UN CORPUS PERIODÍSTICO EN EUSKARA

Itziar Aduriz
Universitat de Barcelona UB
Klara Ceberio - Arantza Díaz De Ilarraza - Inés Garcia Azkoaga
Euskal Herriko Unibertsitatea EHU-UPV

RESUMEN

El discurso se forma mediante lazos que se establecen entre los diferentes componentes del texto para que el lector o el oyente puedan interpretar correctamente el texto. Estos lazos y relaciones son las redes referenciales (correferencias y anáforas) que hemos estudiado para el euskara. En este artículo trataremos de explicar los pasos que hemos llevado a cabo para etiquetar un corpus periodístico a nivel textual. A este nivel es imprescindible que se analicen las redes referenciales para que se pueda automatizar la detección de sus correspondientes antecedentes, y posteriormente, aquellos antecedentes implicados en una relación anafórica. Si tenemos en cuenta las características lingüísticas del euskara y el hecho de que por tratarse de una lengua minoritaria son pocos los estudios que se dedican al tratamiento automático de la misma; esta tarea es importante de cara a trabajos posteriores.

Palabras clave: Referencia, anáfora, anotación.

ABSTRACT

The discourse is formed by chains that are established among the different components of the text sot that the reader or listener can correctly understand the text. This chains and relations are the relational chains (coreferences and anaphors) we have studied in the Basque language. In this paper we will explain the process followed to annotate a news corpus, at text level. At this level it is essential to analyze the referential chain relations as the basis for the future computational treatment of this phenomenon. If we take into account the linguistic features of the Basque language and the fact that it is a minority language, few studies have been carried out in relation to the anaphor's

computational treatment at textual level. This task will be important for future works.
Keywords: Reference, anaphor, annotation.

1. Introducción

El reconocimiento automático de la anáfora y de la referencia ha sido objeto de estudio en distintas lenguas (Reboul 1989; Mitkov 2002) pero no ha sido abordado hasta ahora en lo que se refiere a la lengua vasca. Las peculiaridades lingüísticas de esta lengua, tales como el hecho de que se trate de una lengua pro-drop, es decir, que la marca el sujeto gramatical se refleje en la flexión verbal, y sus diferencias con respecto a otras lenguas indoeuropeas, justifican y hacen necesaria comenzar con la creación y el análisis de un corpus propio. En este corpus nos planteamos trabajar en el etiquetado de los elementos correferenciales, tanto anafóricos como no anafóricos que, gracias a la relación que guardan con un referente dado, permiten la adecuada progresión de la información en un texto.

En el segundo apartado presentamos las características de la referencia en euskara. Seguiremos analizando el comportamiento de las redes referenciales en el corpus periodístico. Explicaremos el proceso de etiquetado que hemos llevado a cabo y terminaremos con las conclusiones y trabajo futuro.

2. Características que presenta la referencia en euskara

Tal y como se ha apuntado en la introducción, el estudio de la referencia en euskara tiene algunas particularidades que la diferencian de las lenguas indoeuropeas. Por ello ha sido necesario un estudio previo de clasificación y detección de estas características antes de comenzar con la anotación. En este punto haremos un repaso de las características más importantes del euskara que tienen relación con el estudio de la referencia.

El euskara carece de género gramatical, lo cual dificulta la resolución o la detección automática de ciertos tipos de anáfora como puede ser la pronominal, que es la que se establece mediante la relación

de un pronombre con su antecedente. A esto hay que añadirle otra característica, en euskara no existen pronombres específicos de tercera persona y los pronombres demostrativos ocupan su lugar como se puede observar en el siguiente ejemplo (Laka 2002).

(1) *Nire lagun Pello* etxera zihoala ikusi dut. *Hark* ez nau ikusi.

(He visto *a mi amigo Pello* irse a casa. *Él (ése)* no (me) ha visto)

Por último mencionar también que al tratarse de una lengua pro-drop, aparte de la marca del sujeto gramatical, en la flexión verbal se reflejan la marca del objeto directo e incluso objeto indirecto, es decir que los casos de objeto directo e indirecto pueden ser elididos. Este mecanismo posibilita que la información que por ejemplo en las lenguas románicas se obtiene mediante los pronombres clíticos no tenga que ser realizada léxicamente en euskara.

(2) Amaiak *Ø* eskutitza idatzi dio.

(Amaia *le* ha escrito una carta a Pello).

En el ejemplo (2) sabemos que *Amaia* ha escrito a una tercera persona singular porque el verbo lo refleja mediante la inflexión de tercera persona singular.

3. La referencia en el corpus periodístico

Gracias a la diversidad de géneros textuales que podemos encontrar en la prensa escrita, el análisis de los textos periodísticos en euskara nos permite acceder fácilmente a los distintos tipos de discurso y al análisis de las distintas relaciones anafóricas y referenciales que se establecen a la hora de construir la cohesión textual de los mismos.

El corpus que hemos tratado en este trabajo se compone de textos extraídos de la prensa escrita en euskara. Se han recogido automáticamente ciertos apartados de periódico (sólo noticias,

excluyendo artículos de opinión) recogidos a lo largo del año 2000, con el fin de proceder al etiquetado de los elementos que constituyen las redes anafóricas y referenciales.

Para que las informaciones que proporciona un texto sean comprendidas e interpretadas adecuadamente es necesario que la cohesión esté bien construida, y es ahí donde tienen un papel primordial las expresiones referenciales, y sobre todo, las relaciones anafóricas que se crean a lo largo del texto. Con frecuencia, cuando en el texto se introduce un objeto de discurso, para interpretarlo adecuadamente, resulta necesario identificar su referente original o antecedente. La relación que se establece entre un elemento y su antecedente, puede expresarse por medio de unidades lingüísticas muy diferentes como podemos ver en el siguiente ejemplo, adaptación de una noticia extraída de un periódico en euskara.

(3) Abuztuaren 17an gorpu bat topatu zuten, deskonposizio egoeran Lodosako (Nafarroa) ubidean. 33 urteko emakume espainiar batena zela jakinarazi zuten ikerketa iturriek. *Emakume hori* Raquel Sanchez Suñen zen eta Calahorrakoa zen, Errioxa erkidegokoa. *Gorpua* topatu zutenean, deskonposizio egoeran zegoen, eta ADN frogak egin behar izan zizkioten. Hilabete lehenago desagertu zen *errioxarra bere* etxetik. *Raquel Sanchez Suñen*-en senideak eta polizia aspaldi ari ziren haren bila. [Berria-tik (2007-09-08) moldatua]

(El 17 de agosto encontraron un cadáver en avanzado estado de descomposición en el canal de Lodosa (Navarra). Fuentes de la investigación informaron de que se trataba de una mujer española de 33 años. *Esa mujer* era Raquel Sánchez Suñen, natural de Calahorra, de la Comunidad Autónoma de la Rioja. Cuando encontraron *el cadáver*, estaba en estado de descomposición y tuvieron que hacerle pruebas de ADN. *La riojana* había desaparecido un mes antes de

su casa. Los familiares de Raquel Sánchez Suñen y la policía llevaban tiempo en *su* busca.) [Adaptación del periódico Berria, (2007-09-08)]

... Emakume hori (esa mujer) → sustantivo con determinante demostrativo

... Gorpua (el cadáver) → sustantivo con determinante

... Errioxarra (riojana) → gentilicio

... bere (su) → pronombre posesivo (en euskara funciona como tal)

... Raquel Sanchez Suñen → nombre propio

... haren (su) → determinante posesivo

Cuando se trata de los pronombres de tercera persona, su relación de identidad con el antecedente no plantea grandes dudas, pero no sucede lo mismo cuando se trata de expresiones nominales. Al tratar de identificar las relaciones anáforicas que aparecen en los textos periodísticos nos encontramos con ejemplos que ponen en evidencia la diversidad y la complejidad de las mismas. Por otro lado, el hecho de que entre dos expresiones haya relación anafórica, no implica que ambas hayan de ser correferentes, ni tampoco, que por el hecho de ser correferentes sean necesariamente anafóricos.

En algunos casos, la relación anafórica entre dos elementos surge indirectamente, por inferencia, sin que ambos elementos tengan una identidad referencial. Es el caso de las anáforas asociativas (Kleiber 1994) o virtuales, como las denomina Milner (1982). Por ejemplo:

(4) Misil balistikoek osatzen dute *Txinaren* arsenalaren bizkar-hezurra. *Pekinek*, gutxi gorabehera serieko 170 misil inguru ekoitzi dituela uste da.

(Los misiles balísticos componen la columna arsenal

de China. Se cree que *Pekín* ha producido más o menos 170 misiles de serie.)

En el caso de las anáforas evolutivas, por ejemplo, también resulta cuestionable la correferencia. En un ejemplo como (5), tras una transformación del elemento original, no podemos decir que el referente inicial (*manzana*) y el referente del resultado final (*compota*) sean los mismos:

(5) Har itzazu *lau sagar*. Zuritu eta zatitu. Eduki egosten ordu erdiz. Txiki-txiki egin. Hoztu ondoren, zerbitzatu *konpota* hori gailetatxoekin.

(Se cogen *cuatro manzanas*. Se pelan y parten. Se cuecen durante media hora. Se hacen trozos pequeños. Una vez enfriado, se sirve *la compota* con galletas.)

En el otro extremo, tenemos los ejemplos de expresiones que son correferenciales pero que no implican una relación anafórica, como es el caso de los nombres propios (6), ya que el nombre propio designa directamente al referente, y no necesitamos recurrir al antecedente para interpretar la expresión que lo retoma. También puede resultar dudosa la relación anafórica existente en el ejemplo (7) entre *Ibarretxe* y *lehendakari* (presidente); ambas expresiones tienen un referente común, pero gracias a un conocimiento compartido del mundo podemos identificarlos directamente, en otras palabras, ambos elementos se interpretan de manera autónoma y no uno a través del otro (Kleiber 1988 y 1994):

(6) *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.)

(7) *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*…)

A fin de ir avanzando poco a poco en el reconocimiento automático de las anáforas en euskara, en un trabajo anterior comenzamos a analizar las anáforas pronominales. En esta ocasión vamos a ampliar el análisis, además de tener en cuenta las pronominales, nos centraremos en las anáforas nominales. No obstante, dado que la identificación de este tipo de anáforas puede ser a veces muy compleja, como así lo muestra García Azkoaga (2004), y teniendo en cuenta que con las herramientas informáticas de las que disponemos hoy en día es difícil abordar el etiquetado de todas las anáforas nominales, anotaremos únicamente cierto tipo de anáforas nominales, y relaciones correferenciales que explicaremos más adelante.

4. Creación y análisis del 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 catalán y el castellano (Palomar et al. 2004). El objetivo de este proyecto era la anotación sintáctica y semántica del de un corpus que cuenta con 300.000 palabras.

Antes de proceder a la anotación de distintos tipos de redes referenciales, este apartado explicaremos brevemente el proceso de análisis modular (Aduriz et al. 2006).

Primeramente el corpus es etiquetado automáticamente mediante el analizador morfológico, 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. En este punto el mayor problema es el de la ambigüedad, ya que 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, cuya función principal es reducir la ambigüedad morfosintáctica, eligiendo en cada contexto la mejor opción. Para terminar con el análisis se aplica el *chunker*[1]. Este módulo define los sintagmas, y devuelve el texto dividido en sintagmas.

El nivel morfológico y el sintáctico se han realizado automáticamente. El etiquetado del nivel textual, lo realizaremos manualmente con la ayuda de una herramienta que comentaremos en el siguiente apartado.

5. Anotación de la referencia

Partiendo de un corpus anotado morfosintácticamente se nos ha facilitado la labor de fijarnos en unas estructuras gramaticales específicas que han sido seleccionadas pensando en su posterior detección automática.

Al igual que cuando etiquetamos las anáforas pronominales (Aduriz et al. 2007), hemos ampliado la anotación de las estructuras pronominales, a las anáforas nominales. Sin embargo, 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. Para esta anotación contaremos con la aplicación que mencionaremos después, MMAX (Müller & Strube 2001).

En esta anotación han participado tres personas; dos personas han anotado el mismo corpus y la tercera actuado como juez en los casos dudosos a fin de garantizar la coherencia del etiquetado. El proceso se ha llevado a cabo con la ayuda de la aplicación mencionada, que primeramente hemos adecuado a las necesidades específicas del idioma tratado, el euskara.

5.1. La aplicación MMAX

Ésta es la herramienta que ha facilitado la anotación, ya que hemos obtenido textos ya etiquetados a nivel morfosintático, con los sintagmas marcados como observamos en el siguiente ejemplo:

(8) [Aurreneko bozketan] [inork] ez du [gehiengo osoa] lortu, eta [Gallastegi] [lehiatik kanpo] geratu da. [Bigarren itzulian] hautatuko dute [errektorea], [Montero eta Perezen artean]. [Monterok] lortu zituen [boto gehien] [atzoko bozketan], eta [Perez] izan zen [bigarrena], (…)

([En la primera votación] [nadie] ha conseguido [mayoría absoluta], y [Gallastegi] se ha quedado [fuera de la votación]. [En la segunda vuelta] elegirán [al rector] [entre Montero y Perez]. [Montero] consiguió [más votos] [en la votación de ayer], y [Perez] fue el segundo (…))

En esta anotación se han verificado los sintagmas que vienen marcados automáticamente, ya que éstos pueden ser anafóricos y también antecedentes de las relaciones correferenciales. Por una parte se han corregido algunos casos y por otra se ha considerado necesario marcar partes de los sintagmas, es decir, los componentes del sintagma que puedan funcionar como antecedentes de una correferencia, como se muestra en el siguiente ejemplo.

(9) Bigarren itzulian hautatuko dute errektorea, [[*Montero*] eta [*Perezen*] artean]. *Monterok* lortu zituen boto gehien atzoko bozketan eta *Perez* izan zen bigarrena, hamar boto gutxiagorekin.

(En la segunda vuelta se elegirá al rector, [entre [*Montero*] y [*Perez*]]. [*Montero*] fue quien logró más votos en la votación de ayer y [*Perez*] fue el segundo, con 10 votos menos)

Otro tipo de elemento que puede ser componente del sintagma es el genitivo. En estos casos tendremos en cuenta los nombres propios de persona y de lugar, que pueden tratarse de antecedentes.

(10) Hala ere, egoitza inguratu zuen [[*Indonesiako*] Poliziak] eta inori ez zion hara hurreratzen utzi. [[*Indonesiako* herritarrak]] beldur ziren.

(Sin embargo, [la policía [de *Indonesia*]], rodeó el edificio y no dejaron acercarse a nadie. Los [habitantes [*de Indonesia*]] tenían miedo)

Una vez definidos los componentes que pueden ser parte de la relación referencial hemos etiquetado las expresiones referenciales: las anáforas y las estructuras correferenciales. En este momento se está cuantificando el resultado de las anotaciones que se está realizando por dos lingüistas en paralelo tal como se ha mencionado antes y esperamos disponer de resultados en un plazo breve.

5.2. Tipología de la referencia

Siguiendo los criterios que hemos mencionado en los anteriores puntos hemos procedido a identificar y a etiquetar los siguientes elementos:

Nombres propios: como el nombre indica marcaremos los nombres propios, de persona, lugar u organización.

(11) Bigarren itzulian hautatuko dute errektorea, *Montero* eta *Perezen* artean. *Monterok* lortu zituen boto gehien atzoko bozketan eta *Perez* izan zen bigarrena, hamar boto gutxiagorekin.

(En la segunda vuelta se elegirá al rector, entre *Montero* y *Perez*. *Montero* fue quien logró más votos en la votación de ayer y *Perez* fue el segundo, con 10 votos menos)

Pronombres y demostrativos (que aunque ya han sido etiquetados anteriormente incluimos también aquí).

(12) Adituek uste dute *lau DF5* baino ez daudela zabalduta, baina *haietako* bakoitzak lau megatoi ditu eta EEBBak, Errusia edo Europa jotzeko gaitasuna.

(Los expertos creen que sólo hay *cuatro DF5s* extendidos, pero que cada uno *de ellos* contiene cuatro megatones, suficientes para atacar a EEUU, Rusia y Europa)

La anáforas fieles consistentes en la repetición de parte o de la totalidad del antecedente, de forma que pueden retomar el mismo lexema pero con distinto determinante o con la aparición o la omisión de un elemento atributivo [sustantivo + (elem. atrib.) + (artículo / determinante)].

(13) Txinak *150 bonba nuklear* zituen 1993an, gehienak hidrogenozkoak edo termonuklearrak. Hain zuzen ere, hidrogenoaren kateko erreakzioaren energia baliatzen da *bonba horietan.*

(China tenía *150 bombas nucleares* en 1993, la mayoría de hidrógeno y termonucleares. Precisamente *esas bombas* utilizan la energía de la cadena de reacción del hidrógeno).

Adverbios de lugar, en los que es necesaria la interpretación del antecedente para su pleno entendimiento.

(14) UNAMET Nazio Batuetako misioaren egoitzaren bi eraikini eraso zieten, timortarrak *han* babestuta zeudela.

(Atacaron *dos edificios de la residencia de la misión de las Naciones Unidas UNAMET*, mientras los timorenses se protegían *allí*.)

6. Conclusiones y trabajo futuro

En este trabajo hemos presentado las características de ciertos tipos de referencia que se dan en los textos periodísticos en euskara. Partiendo de estas particularidades hemos llevado a cabo el etiquetado de la correferencia y la anáfora en este tipo de textos estableciendo una tipología y unas bases para el etiquetado de mayor volumen de corpus. Esta labor se ha realizado mediante una aplicación específicamente preparada y adecuada para el euskara, que ha sido de gran ayuda.

Esta anotación servirá también para facilitar la detección automática de la anáfora y la correferencia, pudiéndola utilizar no sólo en herramientas destinadas a la resolución automática de la referencia, sino también en otro tipo de aplicaciones tales como sistemas de búsqueda de respuesta, sistemas de resumen automático o en la traducción automática.

Notas

1. Fragmentador, divisor de sintagmas

Referencias bibliográficas

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.

Aduriz, I., Aranzabe, M., 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.

Garcia Azkoaga, I.M. 2004. *Kohesio anaforikoa hiru testu generotan.*

Adinaren araberako azterketa, Bilbao, Euskal Herriko Unibertsitatea.

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

_____ 1994. *Anaphores et pronoms*, Louvain-la-Neuve, Duculot.

Laka, I. 2000. A Brief Grammar of Euskara, the Basque Language. Documento HTML. Euskarako errektoreordetza, Euskal Herriko Unibertsitatea. http://www.ehu.es/grammar

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

Mitkov, R. 2002. *Anaphora resolution*. London: Longman.

Müller, C., Strube, M. 2001. "MMAX: A Tool for the Annotation of Multi-modal Corpora". *In Proc. of the 4th SIGDIAL*, Sapporo, Japan.

Reboul, A. 1989. "Résolution authomatique de l'anaphora pronominal". in RUBATEL (ed.) *Modèles du discours. Recherches actuelles en Suisse romande*. *Actes des rencontres de Linguistique française,* Berne, Peter Lang, 173-192.