Procesamiento del Lenguaje Natural, num. 29 (2002), pp.5-11 recibido 27-05-2002; aceptado 01-07-2002

Construcción de un corpus etiquetado sintácticamente para el euskera

I. Aduriz*, I. Aldezabal, M. Aranzabe, B. Arrieta, J.M. Arriola, A. Atutxa, A. Díaz de Ilarraza, K. Gojenola, M. Oronoz, K. Sarasola**

Grupo IXA (http://ixa.si.ehu.es)
Departamento de Lenguajes y Sistemas Informáticos.
Universidad del País Vasco
649 p.k., 20080 Donostia

* Departamento de Lingüística General
Universidad de Barcelona
Gran Via de las Corts Catalans, 585, 08007 Barcelona
itziar@fil.ub.es

**Departamento de Lingüística General
Universidad de Maryland
College Park, Maryland, 20740
jibatsaa@si.ehu.es

Resumen: El objetivo de este trabajo es la construcción de un corpus anotado sintácticamente para el euskera. En esta comunicación presentaremos, en primer lugar, las bases sobre las que se asienta nuestro etiquetado. Tras examinar diversas opciones se optó por el esquema presentado por (Carrol *et al*., 1998). Este esquema sigue los estándares EAGLES y se basa en la idea de añadir a cada frase del corpus una serie de relaciones gramaticales que especifican la dependencia existente entre el núcleo y sus modificadores. Una vez presentado el formalismo de etiquetado, se expondrán los problemas que hemos encontrado en nuestra tarea y las decisiones tomadas. Seguidamente se describirá un ejemplo concreto en el que se muestra la aplicación de dicho esquema sobre un corpus inicial. Finalmente, presentaremos las conclusiones sobre la idoneidad del esquema al euskera y trabajo futuro.
Palabras clave: Lingüística de corpus, anotación sintáctica.

Abstract: The aim of this work is the construction of a syntactically annotated treebank for Basque. In this paper we present first, the basis of the annotation. After examining several options we chose the scheme presented in (Carrol *et al*., 1998). It follows the EAGLES standards and it is based on the idea of adding to each sentence in the corpus a series of grammatical relations specifying the dependencies between modifiers and their nucleus. After the formalism has been presented, we will describe the problems we have found and the decisions we have taken to solve them. Next we present an example showing the application of the scheme to an initial corpus. Finally, we present the main conclusions about the applicability to Basque and future work.
Keywords: corpus linguistics, syntactic annotation

ISSN 1135-5948

1 Introducción

Este artículo describe la creación y elaboración de un corpus[1] anotado sintácticamente para el euskera. Este corpus tendrá varias utilidades, como la evaluación de analizadores sintácticos o el aprendizaje automático. En concreto, tenemos en mente la elaboración de un analizador sintáctico basado en las gramáticas de restricciones, Constraint Grammar (Karlsson *et al*., 1995), (Aduriz *et al*., 1997).

La construcción y obtención de un corpus etiquetado sintácticamente es un paso muy importante dentro de las aplicaciones en el área del procesamiento del lenguaje natural, ya que constituye un recurso indispensable para la construcción de herramientas. En esta línea de investigación, a nivel del estado español, se está desarrollando un proyecto para la elaboración de una base de datos de árboles sintácticos y semánticos, dentro del cual se enmarca el presente trabajo.

Tenemos que situar la necesidad de dicho corpus para el euskera y enmarcarlo dentro de los trabajos en torno al procesamiento del lenguaje natural que lleva a cabo el grupo IXA (http://ixa.si.ehu.es) en la Facultad de Informática de Donostia. El planteamiento de elaboración de recursos robustos para el tratamiento computacional del euskera ha tenido diversos focos de estudio en los últimos años. Así, desde un estudio exhaustivo de la morfología y el léxico, se ha llegado inevitablemente al tratamiento sintáctico, donde se sitúa el presente trabajo.

Por lo tanto, los retos que nos planteamos al iniciar dicho tratamiento sintáctico son por una parte, la elección del formalismo adecuado para la anotación sintáctica y por otra parte, continuar con la elaboración del analizador sintáctico, que será evaluado con el corpus previamente etiquetado. Y todo ello con la característica diferencial que supone trabajar sobre una lengua aglutinante que tiene notables peculiaridades morfológicas y sintácticas.

2 Diversas posibilidades en el etiquetado sintáctico

Se han consultado los criterios de anotación sintáctica más significativos en los corpus en lengua inglesa (Rambow *et al*., 2002; Sampson, 1987; Taylor *et al.*, 2001, entre otros) y otras lenguas (Abeillé *et al.*, 2000; Alfonso *et al*., 2002; Bosco *et al*., 2000; Boguslavsky *et al*., 2002; Civit y Martí, 2002; Oflazer *et al*., 2001, entre otros).

De las diversas posibilidades existentes a la hora de elegir el formalismo adecuado para una anotación sintáctica, nosotros veremos dos de los más utilizados en las anotaciones de corpus, siempre dentro de un planteamiento computacional. Por una parte analizaremos brevemente la anotación basada en constituyentes (o parentización) y por otra, la basada en dependencias.

2.1 Anotación sintáctica basada en constituyentes

Uno de los corpus más extensos y más utilizados en inglés, el Penn Treebank (Marcus *et al.*, 1993), está anotado siguiendo la propuesta basada en constituyentes. En dicho corpus están marcados los componentes de cada uno de los constituyentes sintácticos, así como su categoría sintáctica, información a cerca de las funciones sintácticas, como sujeto, objeto, etc.

Veamos un esquema de dicho formalismo:

(O (SN X) (SV (SN Y) (V Z)))[2]

Este método tiene dos características fundamentales:

a) es un método basado en el orden lineal;

b) la información jerárquica queda implícita;

c) la información funcional tiene un carácter secundario.

Teniendo en cuenta dichas características, en las pruebas realizadas sobre nuestros corpus en euskera, nos hemos encontrado con los siguientes problemas:

a) Al ser el euskera una lengua de orden libre en la oración, la parentización deja al descubierto problemas que en otras lenguas de orden rígido no aparecen. Así, la oración *Eguzkiak jende guztia alaitzen du* 'el sol alegra a toda la gente' también nos puede aparecer en este orden: *Jende guztia eguzkiak*

1 El corpus inicial se compone de 50.000 palabras. Es un corpus general y equilibrado en euskera estándar, extraído en su mayor parte del diario *Egunkaria* de los últimos años y del Corpus Estadístico del Euskera del Siglo XX (UZEI).

2 O=oración; SN=sintagma nominal; SV=sintagma verbal; V=verbo.

alaitzen du. La diferencia entre ambas sería debida a la focalización del elemento que va justo delante del verbo. Obviamente, el resultado de análisis de cada caso utilizando la anotación basada en constituyentes sería diferente y por ejemplo el sintagma verbal quedaría constituido por dos sintagmas nominales que no coincidirían en ambos análisis.

b) Por otro lado, tenemos el caso de los elementos discontinuos. Veamos este ejemplo:

Gizona alaitu zen, baita emakumea ere 'el hombre se alegró, la mujer también'

En euskera, el elemento discontinuo compuesto de dos elementos *baita* (...) *ere*, que en castellano significa 'también', admite un sintagma nominal (*emakumea* 'mujer' en este caso) entre los dos componentes.

Es por esto que sería francamente difícil lograr un análisis coherente de un elemento discontinuo como el presentado, por medio de paréntesis.

2.2 Anotación sintáctica basada en dependencias

Basándonos en los mismos corpus de prueba, la anotación basada en dependencias (Carrol *et al*., 1998) ha respondido de una manera más satisfactoria ante los problemas anteriormente señalados. Dicho esquema de anotación basado en dependencias se ha utilizado en el corpus NEGRA para el alemán (Skut *et al*., 1997) y el corpus PDT (Prague Dependency Treebank) para el checo (Hajic, 1998).

Las características principales de este método son las siguientes:

a) el orden lineal tiene menor relevancia;

b) es un método fuertemente basado en la jerarquía;

c) la información funcional sí tiene relevancia.

Así, siguiendo este formalismo, la característica del euskera de orden libre en la oración no presenta ningún problema en este sentido; véase el ejemplo anterior analizado con este método:

a. ***Eguzkiak*** *jende guztia alaitzen du*

b. Jende guztia ***eguzkiak*** *alaitzen du*

ncsubj (erg, alaitzen, ***eguzkiak****)*

ncobj (abs, alaitzen, jende)

Como vemos, la anotación de dependencias únicamente especifica que *eguzkiak* es el sujeto (ncsubj[3]) y que *jende* es el objeto (ncobj[4]). Es decir, a diferencia del resultado de análisis del constituyentes, la misma anotación es válida para las dos oraciones, lográndose así la abstracción respecto al orden lineal.

De la misma manera, tal y como presentamos seguidamente, se soluciona la característica de los elementos discontinuos. Utilizaremos el ejemplo presentado anteriormente:

Gizona alaitu zen, ***baita*** *emakumea* ***ere***

ncsubj (abs, alaitu, Gizona)

conj[5] *(****baita****, alaitu, e,* ***ere****)*

ncsubj (abs, e, emakumea)

Basándonos en algunos ejemplos, hemos comprobado que la anotación basada en dependencias se acopla con mayor acierto tipo a una lengua como el euskera. Además de esto, podríamos añadir en favor de dicha aproximación, que es un método sencillo e intuitivo. Así pues éste será el esquema que aplicaremos en la primera aproximación de anotación sintáctica de un corpus en euskera.

3 Anotación sintáctica basada en el esquema de dependencias

3.1 Estado de la cuestión

Como se ha mencionado al comienzo, hay que tener en cuenta que presentamos la primera aproximación al trabajo de anotación sintáctica basada en el esquema de dependencias. Por tanto, se esbozarán las primeras decisiones

[3] ncsubj=non-clausal subject
[4] ncobj=non-clausal object
[5] conj=conjunción

tomadas y conclusiones de dicho trabajo y los problemas que hemos encontrado al analizar una parte del corpus.

Así, siguiendo la misma línea de trabajo y el planteamiento inicial de (Carrol *et al.*, 1998), hemos elaborado un esquema de la jerarquía de las relaciones gramaticales, tanto de elementos con realización léxica como de los elementos *pro-drop*. Esta jerarquía aparece en el siguiente gráfico (figura 1), y en ella se especifican todas las relaciones gramaticales que hemos encontrado al etiquetar el corpus. La jerarquía define unas clases generales que se van especificando por niveles. Así, por ejemplo, las clases generales contempladas son los complementos, los roles temáticos (arg_mod), modificadores, elementos conectores, etc. Estas clases, a su vez, se subdividen en sentencias y núcleos no sentenciales (nc). De nuevo, cada subclase se especifica a nivel más detallado, teniendo en cuenta su función gramatical (p.e.: ncsubj, ncobj y nczobj).

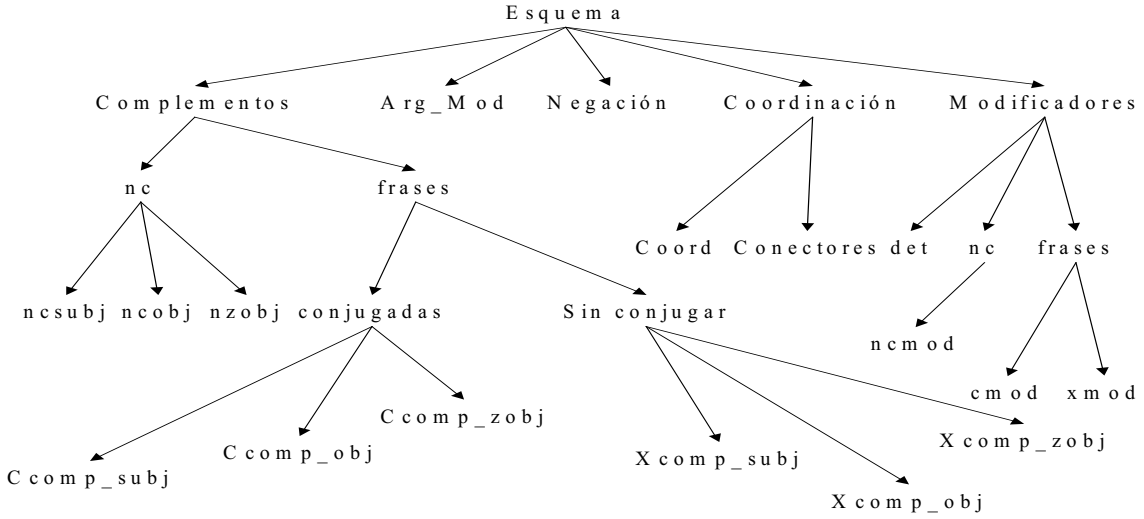


Figura 1.- Vista parcial de la jerarquía de relaciones gramaticales

Posteriormente, y como continuación de la jerarquía presentada, hemos analizado y descrito todas y cada una de las relaciones arriba mencionadas con el objetivo de saber con detalle el tipo y número de etiquetas que requiere cada relación (número de campos a describir, característica que describe en cada caso, etc.). Este trabajo de descripción es muy importante, ya que de cara a tratamientos posteriores, como puede ser el volcado de información a SGML, quedarán descritas las plantillas para cada tipo de relación.

En el siguiente esquema se muestra una de las relaciones gramaticales mencionadas. En este caso se muestran las plantillas de relación gramatical correspondiente a sintagma nominal que tienen función de sujeto, objeto directo o indirecto. Se añade un ejemplo para la mejor comprensión del esquema.

ncsubj (1. caso,

2. núcleo de la oración,

3. nucleo del SN,

4. palabra que lleva el caso dentro del SN,

5. Función: sujeto)

ncobj (1. caso,

2. núcleo de la oración,

3. nucleo del SN,

4. palabra que lleva el caso dentro del SN,

5. Función: objeto directo)

nczobj (1. caso,

2. núcleo de la oración,

3. nucleo del SN,

4. palabra que lleva el caso dentro del SN,

5. Función: objeto indirecto)

Ejemplos:

a) **Ikasle batek / greba / egin du**

b) **Ikasleek / greba / egin dute**[6]

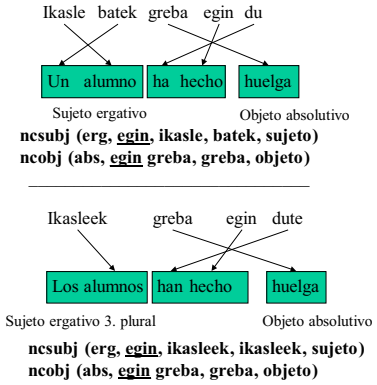


Figura2. Relaciones gramaticales entre los componentes de las frases a y b.

El análisis de una parte del corpus nos ha permitido estudiar los problemas existentes y tomar estas decisiones sobre el diseño del etiquetado. Nos hemos centrado sobre todo en analizar las oraciones de relativo, causales, coordinadas, oraciones en las que aparezcan elementos discontinuos y elipsis, etc.

Paralelamente al trabajo de definición de etiquetas que queremos utilizar en el etiquetado sintáctico, se está trabajando en la elaboración de reglas de cara a la elaboración de etiquetado automático del corpus. Para ello se está utilizando Constraint Grammar, formalismo ya utilizado por este grupo de investigación para la desanbiguación morfosintáctica y análisis sintáctico superficial.

3.2 Problemas en la anotación

Muchos de los problemas que han surgido en el análisis de una parte del corpus, han sido debidos al hecho de que el euskera es una lengua aglutinante. Esta característica es distintiva con respecto al inglés y fundamental, ya que en las lenguas aglutinantes no existen preposiciones sino posposiciones, es decir, sufijos o elementos morfológicos dependientes.

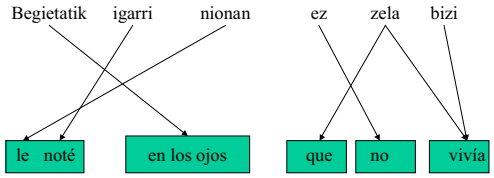
Siendo el euskera una lengua de núcleo final a nivel sintáctico, las marcas morfológicas del sintagma, consideradas como núcleo, las lleva el elemento final del mismo. Y estas posposiciones no van siempre ligadas al mismo nucleo: pueden ir ligadas al nucleo nominal (*ikasleek greba egin dute*), pero también pueden ir ligadas al determinante cuando el sintagma está compuesto de sustantivo + determinante (*ikasle batek greba egin du*).

Esto trae consigo un cambio en el punto de vista de las relaciones de dependencia. Si bien tradicionalmente (siguiendo la morfología del inglés) las relaciones de dependencia vienen introducidas por una preposición, la relación entre el verbo y el sustantivo se lleva a cabo por dicha preposición (*I am talking to a boy*). En euskera, en cambio, la relación con el verbo se lleva a cabo, o por un elemento nominal o por un determinante o adjetivo, que lleva consigo en cada caso la marca posposicional (*Ni mutil batekin hitz egiten ari naiz* '*I am talking to a boy*').

El problema que se deriva de esta característica es que en euskera, a veces la relación sintáctica se crea a través del determinante y el verbo, pero otras, se crea a traves del elemento nominal y el verbo. Hemos solucionado este problema dotando de cuatro etiquetas a la clase referente al modificador (ncmod) y en la última de ellas se especificará siempre el elemento que lleva el caso en el sintagma nominal: ncmod (caso, núcleo de la oración, núcleo del sintagma, determinante o nombre que lleve el caso). Veamos por ejemplo el esquema de estas dos oraciones centrándonos en los modificadores (*begietatik* y *gauza guztiengatik*):

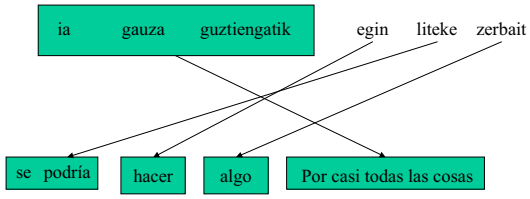
[6] a) 'Un alumno ha hecho huelga'
b) 'Los alumnos han hecho huelga'

a) Begietatik / igarri nionan / ez zela bizi[7]



ncmod (abl, igarri, begietatik, begietatik)

b) Ia/ gauza guztiengatik/ egin liteke/ zerbait[8]



ncmod (mot, egin, gauza, guztiengatik)

4 Conclusiones y trabajo futuro

Hemos presentado los primeros datos de este trabajo sobre la anotación sintáctica de un corpus en euskera. Son las primeras conclusiones que podemos sacar sobre un trabajo que está todavía en sus inicios. Nuestro objetivo es el etiquetado manual sintáctico de un corpus de 50.000 palabras, previamente etiquetado morfológicamente además de la identificación de los elementos anafóricos y correferentes.

Después de examinar diferentes opciones en el campo de la anotación sintáctica, hemos decidido seguir el esquema planteado por (Carroll *et al*., 1998), que parece ser el más adecuado para una lengua como el euskera, de orden libre en la oración.

El modelo de anotación ha favorecido el desarrollo de la evaluación de esquemas basados en dependencias ya que proporcionan una mejor medida para la evaluación de resultados de análisis en general (Lin, 1998).

Gracias a la flexibilidad del modelo de dependencias, nos va a ser posible incluir otros tipos de etiquetas como, por ejemplo, las correspondientes a los papeles temáticos que, aunque en principio no constituyen información puramente sintáctica, son un paso importante de cara a la interpretación semántica que pretendemos abordar en un futuro.

[7] 'Le noté en los ojos que no vivía'

[8] 'Se podría hacer algo por casi todas las cosas'

5 Agradecimientos

Este trabajo se ha realizado dentro del proyecto "Construcción de una base de datos de árboles sintácticos y semánticos", subvencionado por el Ministerio de Educación y Ciencia (PROFIT: FIT-150500-2002-244).

6 Referencias bibliográficas

Abeillé A., L. Clément, y A. Kinyon. 2000. Building a treebank for French. En *Proceedings of the Second Conference on Language Resources and Evaluation* (LREC00), páginas 87-94, Athens, Greece.

Aduriz I., J. M. Arriola, X. Artola, A. Díaz de Ilarraza, K. Gojenola, y M. Maritxalar. 1997. Morphosyntactic disambiguation for Basque based on the Constraint Grammar Formalism. En *Proceedings of Recent Advances in NLP* (RANLP97), páginas 282-288. Tzigov Chark, Bulgary.

Alfonso S., E. Bick, R. Haber, y D. Santos. 2002. "Floresta Sintá(c)tica": A treebank for Portuguese. En *Proceedings of the Third International Conference on Language Resources and Evaluation* (LREC02), páginas 1698-1703, Las Palmas de Gran Canaria, Spain.

Boguslavsky I., I. Chardin, S. Grigorieva, N. Grigoriev, L. Iomdin, L. Kreidlin, y N. Frid. 2002. Development of a Dependency Treebank for Russian and its possible Applications in NLP. En *Proceedings of the Third International Conference on Language Resources and Evaluation* (LREC02), páginas 852-856, Las Palmas de Gran Canaria, Spain.

Bosco C., V. Lombardo, D. Vasallo, y L. Lesmo. 2000. Building a treebank for Italian: a Data-driven Annotation Schema. En *Proceedings of the Second International Conference on Language Resources and*

Evaluation (LREC00), páginas 99-105, Athens, Greece.

Carroll J., E. Briscoe, y A. Sanfilippo. 1998. Parser evaluation: a survey and a new proposal. En *Proceedings of the 1st International Conference on Language Resources and Evaluation* (LREC98), páginas 447-454, Granada, Spain.

Civit, M., y M. Martí. 2002. Design Principles for a Spanish Treebank. En *Proceedings of the Treebanks and Linguistic Theories*. Sozopol, Bulgaria (forthcoming).

Hajic J. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. En *Issues of Valency and Meaning*, páginas 106-132, Karolinum, Praha.

Karlsson F., A. Voutilainen, J. Heikkila, y A. Anttila. 1995. Constraint Grammar: A Language-independent System for Parsing Unrestricted Text. *Mouton de Gruyter*.

Lin D. 1998. A dependency-based method for evaluating broad-coverage parsers. *Natural Language Engineering* 4(2): 97-114, Cambridge University Press.

Marcus M. P., B. Santorini, y M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19(2): 313-330.

Oflazer K., B. Say, D. Z. Hakkani-Tür, y G. Tür. 2001. Building a Turkish Treebank. En *Building and using syntactically annotated corpora*. Kluwer, Dordrecht.

Rambow O., C. Crecwell, R. Szekely, H. Taber, y M. Walker. 2002. A Dependency Treebank for English. En *Proceedings of the Third International Conference on Language Resources and Evaluation* (LREC02), páginas 857-863, Las Palmas de Gran Canaria, Spain.

Sampson G. 1987. Probabilistics models of analysis. En R. Garside, G. Leech, G. Sampson, editors, *The Computational Analysis of English*, capítulo 1, páginas 16-29. Longman.

Skut W., B. Krenn, T. Brants, y H. Uszkoreit. 1997. An Annotation Scheme for Free Word Order Languages. En *Proceedings of the Fifth Conference on Applied Natural Language Processing* (ANLP-97). Washington, DC, USA.

Taylor A., M. Marcus, y B. Santorini. 2001. The Penn Treebank: an overview. En *Building and using syntactically annotated corpora*. Kluwer, Dordrecht.