MARIA JESUS ARANZABE URRUZOLA
ARANTZA DIAZ DE ILARRAZA SANCHEZ

Grupo de investigación IXA
Universidad Del País Vasco
Facultad de Informática
Paseo Manuel Lardizabal 1
20018 Donostia/San Sebastián
{maxux.aranzabe, a.diazdeilarraza}@ehu.es

Análisis sintáctico computacional del euskera mediante una Gramática de Dependencias

Resumen
El objetivo básico del trabajo que aquí se presenta es llevar a cabo un análisis sintáctico total del euskera mediante una Gramática de Dependencias (EDGK). La idea principal es relacionar los elementos explícitos de la oración a partir de las unidades resultantes (tradicionalmente llamados *chunks*) en el módulo anterior, y dejar para una fase posterior el análisis sintáctico profundo porque requiere información semántica. El sistema de representación que se ha adoptado es el de las dependencias por las propias características del euskera y de la secuencia de Procesamiento del Lenguaje Natural anterior.

1. Introducción

Este artículo presenta la Gramática de Dependencias (EDGK) desarrollada para la obtención de un análisis sintáctico total del euskera. La gramática se basa en el esquema de anotación sintáctica utilizado para la creación del Treebank Eus3LB (Palomar et al., 2004) y en el resultado proporcionado por el analizador sintáctico IXATI (Aduriz et al., 2004). Este esquema de anotación basado en dependencias (Carroll et al., 1998) ha sido usado tanto para el etiquetado manual del corpus, como para el desarrollo del analizador sintáctico. Es un esquema que pretende ser básico o neutro en el sentido de no seguir ninguna teoría concreta, pero proporcionando información apta para el estudio del euskera desde cualquier perspectiva, sea ésta estrictamente lingüística o computacional. La aplicación de dicha gramática se realiza en dos fases mediante el formalismo *Contraint Grammar* (Karlsson et al., 1995) y el programa Burubil (Aranzabe, 2008).

Los pasos seguidos en la descripción de la Gramática se presentan en el tercer apartado. Previamente, se hace una muy breve descripción de las principales características del euskera en base a la elección del formalismo basado en dependencias. En el cuarto punto se muestran los datos que corresponden a la evaluación. Finalmente, se recogen las conclusiones.

2. Principales características del euskera en base a la elección del formalismo adoptado

El euskera presenta una serie de características que la diferencian de otras lenguas. Es una lengua flexiva en la que las relaciones gramaticales entre los diferentes elementos de la oración se marcan por medio de sufijos el final de las palabras. Esta característica es distintiva ya que la cantidad de información morfológica presente en la palabra es mucho mayor con respecto a otras lenguas. Siendo una lengua de núcleo final a nivel sintáctico, las marcas morfológicas del sintagma (número, caso, etc.), consideradas como núcleo, las lleva el elemento final del mismo.

A nivel de oración, el verbo aparece como ultimo elemento en un orden neutro. Esto es, dada la tipología propuesta por Greenberg, se asume como regla que el euskera es un tipo de lenguaje Sujeto-Objeto-Verbo (SOV) (Laka, 1998). No obstante, esto corresponde al orden neutro, porque el orden de las palabras en la oración puede variar; por lo tanto, el euskera es conocido como un lenguaje de 'orden de palabras libre'.

Estas características aconsejaron realizar una anotación mediante dependencias, de manera similar a la realizada para idiomas como el checo (Hajic, 1999), aunque también planteada para idiomas de orden menos libre como el inglés (Järvinen y Tapanainen, 1998). Además de esto, podríamos añadir a favor de dicha aproximación, que es un método sencillo e intuitivo.

Las dependencias representan las relaciones de núcleo-modificador entre los elementos terminales de las oraciones. Así, en las dependencias todos los nodos de la representación arbórea son terminales, puesto que las relaciones núcleo-modificador se establecen directamente entre las palabras.

Otras características fundamentales de este método son las siguientes:

a) el orden lineal tiene menos relevancia que en la representación de constituyentes;
b) es un método fuertemente basado en la jerarquía;
c) la información funcional si tiene relevancia.

El análisis de la oración (1) es un claro ejemplo de anotación sintáctica basada en dependencias. Básicamente, la anotación indica el tipo de dependencia (*ncmod*, *ncsubj*) seguida de cinco atributos que representan: i)

información morfosintáctica útil como es el caso, ii) núcleo de la dependencia, iii) elemento dependiente, iv) palabra que lleva el caso dentro del SN, y v) función sintáctica.

(1) Dima Arratiako bailaran dago.
('Dima se sitúa en el valle de Arratia')

ncsubj (abs, dago, Dima, Dima, subj)
ncmod (gel, bailaran, Arratiako, Arratiako izlg)
ncmod (ine, dago, bailaran, bailaran, adlg)

3. Gramática de Dependencias Computacional: EDGK

En esta sección se presenta la Gramática de Dependencias definida para el desarrollo del analizador sintáctico. Para definir esta gramática que se aplica después del análisis sintáctico parcial, se han tenido en cuenta las unidades sintácticas o *chunks* reconocidas por el analizador sintáctico IXATI y los principios seguidos en la construcción del Treebank Eus3LB, concretamente se ha prestado atención a estos puntos:

a) Contexto, esto es, se ha analizado el contexto en el que se da la relación entre el núcleo y modificador. En el contexto se describen las características que ha de presentar cada uno de los elementos (categoría gramatical de la palabra, función sintáctica y tipo de sintagma nominal o grupo verbal, junto con la posición que ocupa en dicha unidad sintáctica) para que se aplique la regla correspondiente.
b) La posición que presenta cada una de las palabras en la oración.
c) Principio lingüístico, esto es, las condiciones que ha de cumplir una palabra para que se le asigne una determinada etiqueta.

Basándonos en estos puntos y las características de las estructuras oracionales del euskera, hemos deducido unos principios lingüísticos y expresado a modo de reglas mediante el formalismo *Constraint Grammar.* Así, esta gramática recoge las reglas que corresponden a cada una de las etiquetas de dependencia descritas en el esquema de anotación utilizado en la construcción del corpus sintáctico (Aldezabal et al., 2007)

Esta gramática consta de 505 reglas distribuidas de la siguiente manera (véase tabla 1.):

Oración simple	**Oración compuesta**		**Etiquetas auxiliares**
	O. subordinadas	**O. coordinadas**	
327	121	3	54

Tabla 1. Número de reglas de la Gramática de Dependencias Computacional.

Este número de reglas no es definitivo, puesto que no se da por terminada la redacción de ellas.

El tipo de información utilizado, por ejemplo, en la definición de las reglas que tienen como finalidad asignar las etiquetas de dependencia a los núcleos de los sintagmas es el siguiente:

a) Sintagma constituido por un única palabra o más de una
b) Categoría gramatical del núcleo del sintagma
c) Caso de declinación
d) Palabra que aparece al final del sintagma
e) Posición que presentan en la oración el núcleo y el modificador
f) Signos de puntuación

En concreto, las reglas definidas para realizar el análisis sintáctico de la oración *Dima Arratiako bailaran dago* ('Dima se sitúa en el valle de Arratia') son:

1. Regla que define la dependencia del sujeto con respecto al verbo

Mediante esta regla se describe la relación de dependencia que existe entre el sujeto (*Dima*) y el verbo principal de la oración (*dago* 'se sitúa').

Así, para que se asigne (MAP) la dependencia (NCSUBJ>) a una palabra, las condiciones (IF) que ha de cumplir ésta son: palabra de categoría nombre (IZE), en caso absolutito (ABS) y único constituyente del sintagma (%SINT). Además es necesario que se halle antes de un signo de puntuación (PUNTUAZIOA) un grupo verbal cuyo núcleo sea un verbo finito (ADI + @-JADNAG) y sea el primer elemento verbal de dicho sintagma (%ADIKATHAS).

```
MAP (NCSUBJ>) TARGET (@SUBJ) IF (0(IZE) + (ABS) + (%SINT)
                                 )
                                (1* (ADI) + (@-JADNAG) + (%ADIKATHAS) BARRIER
                                  PUNTUAZIOA)
                                  );
```

2. Regla que define la dependencia del complemento adnominal con respecto a otro nombre

Mediante esta regla se describe la relación de dependencia que se da entre las palabras que constituyen el sintagma nominal *Arratiako bailaran* 'en el valle de Arratia'.

Las condiciones que ha de cumplir la palabra dependiente *Arratiako* 'de Arratia' para que se le asigne (MAP) la dependencia ncmod> son las siguientes: que cumpla la función de complemento del nombre (@IZLG>), sea una palabra de categoría nombre (IZE) con caso genitivo de lugar (GEL) e inicie el sintagma nominal (%SIH). A su vez, su gobernante ha de ser un nombre (IZE) que cierra el sintagma nominal (%SIB) y se encuentre a su derecha (1).

```
MAP (NCMOD>) TARGET (@IZLG>) IF (0(IZE) + (GEL) + (%SIH)
                                )
                                (1(IZE) + (%SIB)
                                  );
```

3. Regla que define la dependencia del complemento circunstancial con respecto al verbo

Mediante esta regla se le asigna la etiqueta de dependencia a la palabra *bailaran* 'en el valle', núcleo del sintagma nominal que depende del verbo *dago* 'se sitúa'.

Para que se de esta relación de dependencia son éstas las características que han de presentar ambas: la palabra dependiente ha de ser un nombre (IZE) en inesivo (INE) que se encuentra al final del sintagma nominal (%SIB). A su vez, la palabra gobernante ha de situarse a su derecha (1) y ha de ser un verbo compuesto (ADI + @-JADNAG) que inicie el grupo verbal (%ADIKATHAS) o un verbo sintético (ADT + @+JADNAG) que constituya el grupo verbal.

```
MAP (NCMOD>) TARGET (@ADLG) IF (0(IZE) + (INE) + (%SIB)
                                )
                                (1(ADI) + (@-JADNAG) + (%ADIKATHAS) OR
                                  (ADT) + (@+JADNAG) + (%ADIKAT)
                                   );
```

4. Regla que define la dependencia del verbo principal de la oración

La regla definida para la anotación del verbo principal es la siguiente:

```
MAP (ADITZ_NAGUSI) TARGET (@-JADNAG)
                          IF (0 (ADI) + (ASP) + (%ADIKATHAS)
                              )
                              (1(ADL) + (@+JADLAG) + (%ADIKATBU)
                                );
```

Mediante esta regla se identifican todos los verbos de las oraciones simples que cumplen dichas condiciones. Este tipo verbos constituye un grupo verbal de dos elementos: (%ADIKATHAS) y (%ADIKATBU). Así, al verbo (ADI) que presenta la característica del aspecto (ASP) se le asignará dicha dependencia si a su derecha y cerrando el grupo verbal se encuentre un verbo auxiliar (ADL) que cumpla la siguiente función sintáctica: @+JADLAG.

Una vez explicado cómo se definen las reglas, la tabla 2 muestra el tipo de análisis que resulta de la aplicación de la gramática:

Posición	Forma	Lema	Categoría + subcat.	Núcleo	Dependencia
1	Dima	Dima	IZE_LIB	4	ncsubj
2	Arratiako	Arratia	IZE_LIB	3	ncmod
3	bailaran	bailara	IZE_ARR	4	ncmod
4	dago	egon	ADI_SIN	0	root

Tabla 2. Análisis sintáctico de la oración *Dima Arratiako bailaran dago* ('Dima se sitúa en el valle de Arratia').

A su vez, la definición y posterior aplicación de las reglas han ayudado en la concreción del análisis sintáctico parcial previo (Aranzabe, 2008).

En una segunda fase y mediante el programa *Burubil*, se hacen explícitas las relaciones de dependencia que se representan de esta manera:

D-NCSUBJ (w4, w1)
D-NCMOD (w3, w2)
D-NCMOD (w4, w3)

La lectura de este análisis es la siguiente: la letra D representa la dependencia, a continuación se describe la dependencia asignada al modificador o dependiente, y por último entre paréntesis se reflejan las palabras que se encuentran en relación de dependencia , escribiendo los identificadores que corresponden al núcleo o gobernante y modificador o dependiente sucesivamente, por ejemplo (w4, w1). El resultado de la unión de estas palabras es un árbol de dependencias.

4. Evaluación

En esta sección se presenta la evaluación de la Gramática Computacional desarrollada siguiendo los formalismos Constraint Grammar y Gramática de Dependencias. Para medir la idoneidad de la gramática se ha utilizado una muestra del corpus EPEC (Corpus de Referencia para el procesamiento del Euskera) (Aduriz et al., 2006).

En total se han utilizado 1.639 oraciones. Oraciones que corresponden a los 432 verbos que aparecen en dicha muestra. Una vez elegido el corpus a evaluar, se han comparado los dos análisis, esto es los árboles de dependencia obtenidos en la anotación sintáctica manual y automática.

En la evaluación efectuada se ha medido la precisión y cobertura del analizador sintáctico (véase tabla 3). El porcentaje que corresponde a la cobertura 69%) muestra las relaciones de núcleo-modificador que ha reconocido el analizador sintáctico; a su vez, el porcentaje que corresponde a la precisión enseña cuántas de esas relaciones son correctas, un 62%.

Precisión	Cobertura
62%	69%

Tabla 3. Resultados de la evaluación del analizador sintáctico.

La diferencia fundamental que se da entre los dos tipos de análisis sintácticos comparados, el manual y automático, es debida a las oraciones coordinadas y unidades léxicas complejas.

De esta primera evaluación del analizador sintáctico basado en dependencias se deduce que han de refinarse las reglas que constituyen la gramática, y describir unas nuevas a medida que surjan distintas estructuras oracionales o sintagmáticas.

5. Conclusiones

En este artículo se ha presentado la Gramática Computacional de dependencias desarrollada para el análisis sintáctico total del euskera. Teniendo en cuenta que para realizar el análisis total se requiere información tanto sintáctica como semántica, el objetivo marcado en esta primera fase ha sido hacer explícita la relación entre las palabras o sintagmas de la oración.

Son las primeras conclusiones que podemos sacar sobre un trabajo que todavía está en sus inicios. Nuestro objetivo es el tratamiento automático del Corpus de Referencia para el Procesamiento del Euskera (EPEC) que consta de 300.000 palabras.

Al ser el euskera una lengua de orden libre en la oración, optamos por el formalismo de las dependencias (Tesnière, 1959) y decidimos seguir el esquema planteado por (Carroll et al., 1998). Este modelo de anotación sintáctica 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 son un paso importante de cara a la interpretación semántica que pretendemos abordar en un futuro.

Bibliografía

(Aduriz et al., 2004) Aduirz I., Aranzabe M.J., Arriola J.M., Díaz de Ilarraza A., Gojenola K., Oronoz M. y Uria L. A Cascaded Syntactic Analyser for Basque. *Computational Linguistics and Intelligent Text Processing*, pp. 124-135. LNCS Series . Springer Verlag. Berlín. 2004.

(Aduriz et al., 2006) Aduriz I., Aranzabe M.J., Arriola J.M., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., Urizar R. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing. In Wilson A., Rayson P. and Archer D. (eds.), *Corpus Linguistics Around the World*, pp. 1-15. Rodopi (Netherland). 2006.

(Aldezabal et al., 2007) Aldezabal I., Aranzabe M.J., Arriola J.M., Díaz de Ilarraza A., Estarrona A., Fernández E, Iruskieta M. y Uria L. EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa)dependentziekin etiketatzeko eskuliburua. UPV/EHU / LSI / TR 12-2007

(Aranzabe, 2008) Aranzabe A. Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala. PhD, Universidad del País Vasco (UPV/EHU).

(Carroll et al., 1998) Carroll J., Briscoe T. y Sanfilippro A. Parser evaluation: a survey and a new proposal. *Proceedings of the First International Conference on Language Resources and Evaluation*, pp. 447-454. Granada, España. 1998.

(Hajic, 1999) Hajic J. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In Hajicová (ed.): *Issues of Valency and Meaning. Studies in Honour of Jamila Panevová*, Karolinum, Charles University Press, Prague, pp. 106-132. 1999.

(Järvinen y Tapanainen, 1998) Järvinen T. y Tapanainen P. Towards an implementable dependency grammar. Colina-ACL´98. *Processing od Dependency-Based Grammars*, Kahane and Polguere (eds.), pp. 1-10, Montreal, Canadá. 1998.

(Karlsson et al., 1995) Karlsson F., Voutilainen A., Heikkila J. y Anttila A. *Constraint Grammar: a Language-Independent System for Parking Unrestricted Text*. Mouton de Gruyter. 1995.

(Laka, 1998) Laka I. *A Brief Grammar of Euskara, the Basuqe Language*. Documento HTML. http://www.ehu.es/grammar. Office of the Vice-Dean for the Basque Language. Universidad del País Vasco (UPV/EHU). 1998.

(Lin, 1998) Lin D. Dependency-based Evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems*. Granada, España. 1998.

(Palomar et al., 2004) Palomar M.; Civil M., Díaz de Ilarraza A., Moreno L., Bisbal E., Aranzabe M.J., Ageno A., Martí M.A. y Navarro B. 3LB: Construcción de una base de árboles sintáctico-semánticos para el catalán, euskera y castellano. *XX Congreso de la SEPLN*. Barcelona. 2004.

(Tesnière, 1959) Tesnière L. *Eléments de Syntaxe Structurale*. Librairie Klincksieck, París, 1959.