Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 83-90 recibido 19-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Reutilización del Treebank de Dependencias del Euskera para la Construcción del Gold Standard de la Sintaxis Superficial de la Gramática de Restricciones (CG)

Reusability of the Basque Dependency Treebank for building the Gold Standard of Constraint Grammar Surface Syntax

José María Arriola, María Jesús Aranzabe, Iakes Goenaga

IXA NLP Group, University of the Basque Country (UPV/EHU)
Manuel Lardizabal 1 48014 Donostia
josemaria.arriola@ehu.es, maxux.aranzabe@ehu.es, iakesg@gmail.com

Resumen: El objetivo del trabajo consiste en reutilizar el *Treebank* de dependencias EPEC-DEP (BDT) para construir el *gold standard* de la sintaxis superficial del euskera. El paso básico consiste en el estudio comparativo de los dos formalismos aplicados sobre el mismo corpus: el formalismo de la Gramática de Restricciones *(Constraint Grammar*, CG) y la Gramática de Dependencias (*Dependency Grammar*, DP). Como resultado de dicho estudio hemos establecido los criterios lingüísticos necesarios para derivar la funciones sintácticas en estilo CG. Dichos criterios han sido implementados y evaluados, así en el 75% de los casos se derivan automáticamente las funciones sintácticas para construir el *gold standard*.
Palabras clave: reutilización recursos lingüísticos, creación *gold standard*, sintaxis superficial

Abstract: The aim of the work is to profit the existing dependency Treebank EPEC-DEP (BDT) in order to build the gold standard for the surface syntax of Basque. As basic step, we make a comparative study of both formalisms, the Constraint Grammar formalism (CG) and the Dependency Grammar (DP) that have been applied on the corpus. As a result, we establish some criteria that will serve us to derive automatically the CG style syntactic function tags. Those criteria were implemented and evaluated; as a result, in the 75 % of the cases we are able to derive the CG style syntactic function tags for building the gold standard.
Keywords: reusability of linguistic resources, gold standard creation, surface syntax

1 Introducción

La principal motivación de este trabajo es la construcción del *gold standard* de la sintaxis superficial del euskera reutilizando el *Treebank* EPEC-DEP (BDT) (Aranzabe, 2008). La premisa fundamental de la que parte el trabajo es la imposibilidad de generar el *gold standard* para evaluar la Gramática de Restricciones (Constraint Grammar, CG) de modo exclusivamente manual (Atro, 2012). La idea es la de agilizar dicho trabajo aprovechando los recursos existentes con el menor coste posible. En esta línea existen trabajos similares, entre los que cabría destacar los siguientes: Gelbukh et al., 2005; Nilson et al., 2008; Aldezabal et al., 2008 y Mille et al., 2009.

La reutilización de este recurso lingüístico nos permitirá obtener el *gold standard* de la sintaxis superficial del euskera correspondiente a las 300.000 palabras que constituyen el corpus EPEC (Corpus de Referencia para el Procesamiento del Euskera) (Aduriz et al., 2006). Este gold standard es un recurso indispensable para evaluar las gramáticas de restricciones del euskera (Aduriz et al., 2000) a nivel de las funciones sintácticas del estilo Constraint Grammar (CG) (Karlsson et al., 1995). Por tanto, al hablar de sintaxis superficial nos referimos al análisis de las funciones sintácticas que guardan las palabras agrupándose entre sí en sintagmas, oraciones simples y compuestas (Aduriz & Ilarraza, 2003). El análisis superficial de la oración de la Figura 1 (*Zure gorputza mapa bat zen non ez*

ISSN 1135-5948

nekien herrialde bakoitza non kokatu (Tu cuerpo era un mapa en el que no sabía dónde ubicar cada país)) muestra un ejemplo del análisis superficial que emplearemos para ilustrar los pasos seguidos en este trabajo. Nos centraremos en el análisis de los sintagmas nominales que aparecen resaltados en negrita en dicha oración, es decir, los sintagmas *Zure gorputza* (Tu cuerpo) y *mapa bat* (un mapa).

En la Figura 1 se presenta el análisis morfológico en formato CG. Básicamente, se puede observar que para cada palabra (representada entre los símbolos "< >") de la oración[1] el analizador morfológico ofrece un análisis por línea. El conjunto de los distintos análisis de cada palabra constituye la cohorte donde se recoge la siguiente información y en este orden: el lema, la categoría, la subcategoría, el caso, el número y por último la función sintáctica. Ciñéndonos a las palabras que tomamos como base para ilustrar el proceso, tenemos los siguientes análisis: *zure* (@IZLG>: complemento del nombre; *gorputza* y *mapa* que presentan las siguientes funciones sintácticas: (@SUBJ: sujeto; @OBJ: objeto; @PRED: predicado y @KM>: elemento modificador de la palabra portadora del caso).

```
"<$.>"<PUNT_PUNT>"
"<Zure>"<HAS_MAI>"
 "zu"  PRON 2ª PERSON S GEN  @IZLG>
"<gorputza>"
   "gorputz" N C ABS S @SUBJ
   "gorputz" N C ABS S @OBJ
   "gorputz" N C ABS S @PRED
   "gorputz" N C @KM>
"<herrialde>"
"<bakoitza>"
"<non>"
"<kokatu>"
"<ez>"
"<nekien>"
"<mapa>"
  "mapa"  N C @KM>
  "mapa"  N C ABS S @SUBJ
  "mapa"  N C ABS S @OBJ
  "mapa"  N C ABS S @PRED
"<bat>"
  "bat"  DET INDET  ABS S @SUBJ
  "bat"  DET INDET  ABS S @OBJ
  "bat"  DET INDET  ABS S @PRED
"<zen>"
"<$.>"<PUNT_PUNT>"
```

Figura 1: Ejemplo análisis superficial.

La idea principal es derivar las funciones sintácticas en estilo CG (etiquetas precedidas del símbolo @, ver Figura 1) partiendo del esquema de anotación basado en la Gramática de Dependencias que ha sido utilizado para el etiquetado manual de EPEC-DEP (BDT). El esquema de dependencias está constituido por 29 etiquetas y se basa fundamentalmente en el trabajo de Carroll *et al.*, (1998). Dichas etiquetas de dependencias representan las relaciones, siempre binarias, que se establecen entre los elementos terminales de las oraciones donde una palabra es el núcleo y la otra el dependiente. Así, en las dependencias se destacan las relaciones que se establecen entre las palabras de la oración (Figura 2) en oposición a otras aproximaciones que se basan en constituyentes o estructuras de frase. Por ejemplo, las relaciones de dependencia que corresponden a las palabras que sirven de ejemplo ilustrativo de la Figura 2 se definen de esta manera:

- El adjetivo posesivo *Zure* (tu) en función de complemento del nombre depende del nombre *gorputza* (cuerpo)
- El nombre *gorputza* (cuerpo) en función de sujeto depende del verbo *zen* (era).
- El nombre *mapa* (mapa) en función de predicado depende del verbo *zen* (era).
- El determinante *bat* (un) en función de modificador depende del nombre *mapa* (mapa).

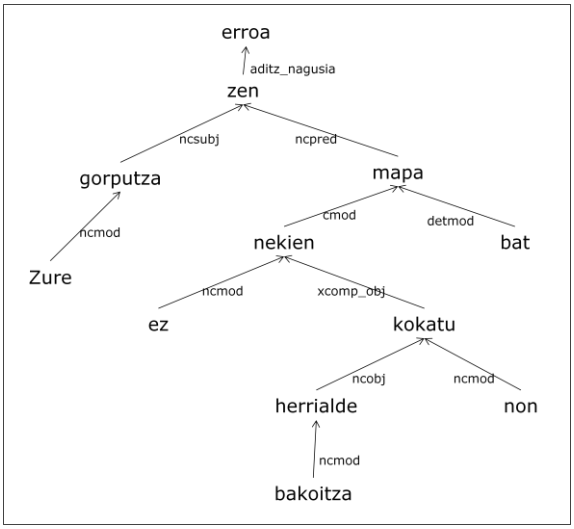

Figura 2: Análisis de dependencias.

El hecho de que el formalismo de dependencias esté basado en palabras al igual

[1] Sólo se ofrece el análisis de aquellas palabras de la oración utilizadas a modo de ejemplo.

que el formalismo de la gramática de restricciones hace factible el poder aprovechar el trabajo de etiquetado no sólo para las gramáticas de dependencias sino también para las gramáticas de restricciones.

En este artículo presentamos los resultados obtenidos a través del estudio comparativo de ambos formalismos de cara a la construcción del *gold standard*. Tras explicar los motivos del trabajo, en la segunda sección presentamos brevemente los recursos lingüísticos en los que se basa el mismo. En la tercera sección, describimos los pasos seguidos en nuestro estudio, centrándonos en los puntos principales del mismo. En la siguiente sección explicamos los criterios para la derivación automática de las funciones sintácticas. En la quinta sección mostramos los datos que corresponden a la evaluación. Finalmente, resumimos las conclusiones principales.

2 Características de los recursos sintácticos existentes

El corpus lingüístico etiquetado manualmente a nivel sintáctico siguiendo el análisis de la Gramática de Dependencias se denomina EPEC-DEP (BDT) y está constituido de 300.000 palabras del euskera estándar.

Con respecto a este recurso básico cabe destacar dos características:

- El trabajo previo llevado a cabo para la formalización del esquema de dependencias (Aranzabe 2008); (Aldezabal et al., 2009).
- La evaluación empírica de la calidad del etiquetado del corpus (Uria et al., 2009).

Ambas características redundan en la calidad y fiabilidad lingüística de nuestro punto de partida.

Por otro lado, disponemos del corpus EPEC etiquetado manualmente a nivel morfosintáctico siguiendo el formalismo de las gramáticas de restricciones (CG). Este corpus se compone de las mismas palabras que componen el Treebank EPEC-DEP (BDT), pero etiquetadas en este caso a nivel morfosintáctico y en formato CG (Karlsson et al., 1995) Para su etiquetado se parte del análisis automático de las gramáticas de restricciones y una vez analizado automáticamente los lingüistas han verificado y corregido en su caso el resultado de la desambiguación morfosintáctica automática. Estos análisis están pendientes de ser desambiguados a nivel sintáctico, es decir, si bien la categoría y el caso han sido resueltos, la desambiguación de las funciones sintácticas tales como por ejemplo la de sujeto u objeto están aún por resolver. Nuestro objetivo es por tanto asignar una sola función sintáctica a cada palabra partiendo del Treebank EPEC-DEP (BDT) construido manualmente.

3 Metodología para el estudio

El punto de partida lo constituye el conjunto de etiquetas fijadas para el análisis de dependencias (Aranzabe 2008); (Aldezabal et al., 2009) que han sido aplicadas manualmente por los lingüistas para la construcción de EPEC-DEP (BDT). Para reutilizar dicho esquema, se establecieron los siguientes pasos:

- Estudiar para cada relación de dependencia fijada en el esquema de anotación de dependencias las equivalencias con las funciones sintácticas siguiendo el formalismo CG.
- Una vez establecidas las equivalencias, definir los criterios por medio de los cuales se derivarán las funciones sintácticas siguiendo el formalismo CG.
- Implementar dichos criterios y estudiar los resultados de la aplicación de dicho criterios, para su posterior refinamiento si así procede.
- Evaluar la aplicación de dichos criterios por medio de dos lingüistas. Dos lingüistas se ocuparán de examinar por separado los resultados obtenidos automáticamente y determinarán la validez de los mismos.

En el siguiente apartado (4) explicaremos de modo somero el proceso relativo a la equiparación y definición de los criterios.

4 Criterios para la derivación automática de las funciones sintácticas

Los criterios para derivar automáticamente[2] las funciones sintácticas en estilo CG a partir del Treebank de dependencias se basan fundamentalmente en la especificación de las etiquetas de dependencias. A continuación se

[2] Basándonos en dichos criterios se implementó el programa en C++. Dicho programa examina un fichero de configuración y en virtud de los criterios lingüísticos mencionados, asigna la función sintáctica correspondiente.

muestra el esquema utilizado para describir las relaciones de dependencia correspondientes a los núcleos del sintagma nominal:

Etiqueta_dependencia (Caso del sintagma, núcleo, núcleo del sintagma, elemento portador del caso, función sintáctica)

Por ejemplo las relaciones de dependencia de las palabras que constituyen el sintagma nominal *Zure gorputza* en función de sujeto de la oración de la Figura 1 se etiquetan de la siguiente manera:

ncmod
(1. caso: genitivo,
2. núcleo del SN: *gorputza*,
3. modificador del SN: *zure*)

ncsubj
(1. caso: absolutivo,
2. núcleo de la oración: *zen*,
3. núcleo del SN: *gorputza*,
4. palabra que lleva el caso dentro del SN: *gorputza*,
5. Función: sujeto)

Teniendo en cuenta la relación de dependencia expresada por la etiqueta *ncsubj* (*non-clausal subject*) observamos que de dicha relación se puede derivar la función de sujeto (@SUBJ) en CG. De este modo y una vez determinada la equivalencia se establecen las condiciones que permiten derivar automáticamente una determinada función sintáctica para cada etiqueta de dependencias.

Por ejemplo, a partir de la etiqueta *ncsubj* se especifican las siguientes condiciones generales para derivar la función sintáctica de sujeto en CG (@SUBJ):

a. Asignar la función @SUBJ a la palabra del cuarto campo en la especificación de la etiqueta de dependencia.
b. Si la palabra del tercer campo no es la misma que la del cuarto, a esa palabra del tercer campo se le asignará la función @KM> (modificador del elemento portador del caso).
c. Si la palabra del cuarto y tercer campo son la misma, prevalecerá el criterio (a.).

Aplicando estos criterios al sintagma *Zure gorputza*, a la palabra *gorputza* se le asignará la función de sujeto (@SUBJ). En la Figura 3, esto se refleja por medio de la etiqueta *Correct* que asignamos automáticamente al análisis que contiene la función sintáctica obtenida mediante la aplicación de los criterios anteriormente explicados.

En relación al sintagma *mapa bat* (un mapa) la función de predicado (@PRED) desempeñada por *mapa* la derivaremos de la siguiente relación de dependencia:

ncpred
(1. caso: absolutivo,
2. núcleo de la oración: *zen*,
3. núcleo del SN: *mapa*,
4. palabra que lleva el caso dentro del SN: *bat*,
5. Función: predicado)

detmod
(1. caso: null,
2. núcleo del SN: *mapa*,
3. palabra que lleva el caso dentro del SN: *bat*)

En este caso partiendo de la etiqueta *ncpred* (*non-clausal predicate o sintagma en función de predicado*) establecemos los siguientes criterios generales:

a. Asignar la función @PRED a la palabra del cuarto campo o slot en la especificación de la etiqueta.
b. Asignar la función de verbo principal a la palabra del segundo campo (@+JADNAG).
d. Si la palabra del tercer campo no es la misma que la del cuarto, asignar a esa palabra la función @KM> (modificador del elemento portador del caso). Salvo que a dicha palabra se le haya aplicado anteriormente una regla para los elementos conjuntivos.
e. Si la palabra del tercer y cuarto campo son la misma, prevalecerá el criterio (a.).

Aplicando estos criterios al sintagma *mapa bat*, a la palabra *mapa* que se encuentra en el tercer campo o slot de la etiqueta de dependencias se le asignará la función de modificador del elemento portador del caso @KM>). Y a *bat* palabra que aparece en el cuarto campo se le asignará la función de predicado (@PRED). Ello viene reflejado en la Figura 3 por medio de la etiqueta *Correct* que asignamos automáticamente al análisis que contiene la función sintáctica obtenida mediante la aplicación de los criterios anteriormente explicados.

En la Figura 3 se muestra el resultado de la aplicación de los criterios para los sintagmas *Zure gorputza* y *mapa bat* en los cuales se han marcado las funciones sintácticas derivadas con la marca *Correct*.

```
"<$.>"<PUNT_PUNT>"
"<Zure>"<HAS_MAI>"
 Correct "zu"  PRON 2ª PERSON S GEN  @IZLG>
"<gorputza>"
 Correct  "gorputz" N C ABS S @SUBJ
   "gorputz" N C ABS S @OBJ
   "gorputz" N C ABS S @PRED
   "gorputz" N C @KM>
"<herrialde>"
"<bakoitza>"
"<non>"
"<kokatu>"
"<ez>"
"<nekien>"
"<mapa>"
 Correct "mapa"  N C @KM>
  "mapa"  N C ABS S @SUBJ
  "mapa"  N C ABS S @OBJ
 "mapa"  N C ABS S @PRED
"<bat>"
  "bat"  DET INDET  ABS S @SUBJ
  "bat"  DET INDET  ABS S @OBJ
  Correct"bat"  DET INDET  ABS S @PRED
"<zen>"
"<$.>"<PUNT_PUNT>"
```

Figura 3: Resultado de la derivación.

A continuación, explicaremos una serie de características básicas que se extraen del estudio comparativo.

En el análisis del sintagma nominal *mapa bat* (un mapa) se refleja una de las diferencias fundamentales entre el análisis de dependencias y el de las gramáticas de restricciones. Así, mientras que en las etiquetas de dependencias la función sintáctica principal recae en el elemento léxico (*mapa*), en la gramática de restricciones se asigna a la palabra portadora del caso en posición final del sintagma nominal (*bat*).

En relación al determinante *bat* (un) vemos que aparece en el análisis de dos etiquetas de relaciones de dependencias: *ncpred* y *detmod*.

Pero a la hora de derivar la función sintáctica de *bat* sólo tendremos en cuenta la etiqueta de *detmod* de la que derivaremos la función de modificador del sustantivo.

Las estructuras coordinadas son tratadas de manera distinta en ambos formalismos. En CG la conjunción se analiza con la etiqueta correspondiente al tipo de conexión que realiza, mientras que en GD es la conjunción la que se etiqueta con la función sintáctica correspondiente a la estructura coordinada. Así la relación de dependencia se expresa tomando como elemento gobernante la conjunción y los elementos coordinados como dependientes que se encuentran al mismo nivel. El análisis de las estructuras coordinadas resulta aún más complejo cuando además de las conjunciones los signos de puntuación, como por ejemplo la coma, funcionan como elementos de coordinación. Es éste por tanto uno de los fenómenos lingüísticos a tratar más profundamente.

Los criterios para derivar las funciones sintácticas constan de 10 reglas para las funciones sintácticas de los núcleos y 11 reglas para las funciones sintácticas de los dependientes. Hay a su vez un grupo de tres reglas que se encargan de las conjunciones y otra serie de categorías sintácticas.

Siguiendo este proceso a través del estudio del esquema general de cada etiqueta hemos establecido 41 criterios correspondientes a las principales etiquetas de dependencias, puesto que quedan fuera de este proceso de equiparación aquellas etiquetas denominadas como auxiliares. Estas etiquetas son las empleadas para etiquetar unidades multipalabra, posposiciones y partículas subordinantes independientes. En la Tabla 1 se muestran de modo simplificado las condiciones necesarias para derivar la función sintáctica correspondiente a los sintagmas, es decir, las etiquetas sintácticas en estilo CG que se pueden derivar de la correspondiente función sintáctica de dependencias (GD). Del mismo modo se han derivado las funciones sintácticas correspondientes a las oraciones.

Significado de la etiqueta	Etiqueta GD	Condiciones	Nº slot	Etiqueta CG
Sujeto	ncsubj	3 y 4 NO IGUAL; 3: @KM>	4	@SUBJ
Objeto	ncobj	3 y 4 NO IGUAL; 3: @KM>	4	@OBJ
Objeto indirecto	nczobj	3 y 4 NO IGUAL; 3: @KM>	4	@ZOBJ
Predicado	ncpred	2: @+JADNAG; 3 y 4 NO IGUAL; 3: @KM>	4	@PRED
Modificador	ncmod	2: CAT= V	4	@ADLG
Modificador	<ncmod	1: -	4	@<IA
Modificador	ncmod>	1: -	4	@IA>
Complemento del nombre	<ncmod	1: GEN	4	@<IZLG
Complemento del nombre	<ncmod	1: GEL	4	@<IZLG
Complemento del nombre	ncmod>	1: GEN	4	@IZLG>
Complemento del nombre	ncmod>	1: GEL	4	@IZLG>
Determinante	detmod>	---	3	@ID>
Determinante	<detmod	---	2	@<ID
Graduador	gradmod>	---	3	@GRAD>
Graduador	<gradmod	---	3	@<GRAD
Sujeto en aposición	aponcmod_subj	---	4	@SUBJ
Objeto en aposición	aponcmod_obj	---	4	@OBJ
Objeto indirecto en aposición	aponcmod_zobj	---	4	@ZOBJ
Adverbial en aposición	aponcmod_adlg	---	4	@ADLG
Complemento del nombre en aposición	aponcmod_izlg>	---	4	@IZLG>
Complemento del nombre en aposición	<aponcmod_izlg	---	4	@<IZLG

Tabla 1: Equivalencias sintácticas de los sintagmas.

A través del estudio comparativo hemos conseguido establecer las equivalencias a nivel sintáctico entre ambos formalismos. En el siguiente apartado mostraremos los resultados de la evaluación.

5 Evaluación

Para la evaluación se tomaron al azar 100 oraciones que cubrían todas las relaciones de dependencia. La evaluación consistió en examinar manualmente dichas oraciones. Y se observó que las reglas derivaban de forma correcta la etiqueta de CG en todos los casos.
Por tanto se consideró que no era preciso el examinar manualmente todo el corpus obtenido automáticamente. Antes de llevar a cabo dicha evaluación que se consideró como definitiva, se llevaron a cabo otra serie de evaluaciones que permitieron subsanar o completar la gramática.

El 25 % del corpus que ha quedado sin etiquetar automáticamente responde

fundamentalmente al hecho de que existen dos puntos de partida de análisis que son diferentes: en el análisis de CG los elementos multipalabra (ya sean posposiciones complejas, locuciones, partículas subordinantes, etc.) no son analizados como una sola unidad; por otro lado, el BDT que ha sido etiquetado manualmente por los lingüistas presenta estos elementos como una sola unidad, por tanto no hay correspondencia entre el número de tokens. Esta es la razón principal por la cual el proceso no se ha realizado totalmente de modo automático.

También cabría señalar otra serie de peculiaridades que presenta el corpus y que dificultan el proceso automático: títulos o encabezamientos, referencias bibliográficas, fórmulas matemáticas, estructuras parentizadas, vocativos, fechas entre corchetes, etc.

Como resultado de la aplicación de los criterios anteriormente señalados, hemos obtenido automáticamente las funciones sintácticas en el estilo de las gramáticas de restricciones en el 75 % de los casos partiendo del Treebank EPEC-DEP (BDT). Es decir, de las 304.751 palabras de las que consta el corpus, 228.982 han sido desambiguadas automáticamente de modo correcto. El 25 % restante (75.769) no se ha podido derivar automáticamente. Pero ello no significa que ambos formalismos no sean equiparables en todos los casos restantes. Así, cabe destacar que en la mayoría de los casos las diferencias radican a nivel de tratamiento de las unidades multipalabra, o de construcciones posposicionales o de ciertos elementos de

subordinación independientes. En todos ellos el denominador común es el de que nos encontramos con distintos estadios de análisis, en el análisis lingüístico llevado a cabo manualmente por los lingüistas se reconocen como una sola unidad las unidades multipalabra o se recogen como una unidad las construcciones posposicionales, por ejemplo.

En cambio, en el corpus analizado siguiendo el formalismo de las gramáticas de restricciones dichas estructuras aún no han sido procesadas en la mayoría de los casos, de ahí que no podamos equiparar propiamente dichos análisis.

6 Conclusión

Los recursos existentes se han reutilizado por medio de las reglas para derivar automáticamente las etiquetas en formato CG. Hemos establecido las bases metodológicas para la constitución del *gold standard* y hemos obtenido un 75% automáticamente. Por tanto, se trata de un trabajo en curso: queda por etiquetar manualmente el 25% del corpus que no se ha conseguido obtener automáticamente. Sobre este corpus un lingüista ha llevado a cabo un trabajo de evaluación del coste en número de horas. Se estima que serán necesarias 450 horas para llevar a cabo el etiquetado del 25% restante.

Por otro lado, si bien nuestro punto de partida ha sido el Treebank EPEC-DEP (BDT) observamos que los criterios establecidos también se podrían utilizar en sentido inverso, es decir, partiendo de un corpus etiquetado en estilo CG para obtener el etiquetado en dependencias. Con ello la reutilización se incrementaría. A su vez, la información formalizada para la derivación de etiquetas en uno u otro sentido, se podría reutilizar en técnicas de aprendizaje (*machine learning*), como por ejemplo, para entrenar un *parser* estadístico.

En un futuro se prevé el estudio detallado de las estructuras más complejas como la coordinación, aposición… con el objetivo de derivar automáticamente un mayor número de funciones sintácticas.

Agradecimientos

Este trabajo ha sido financiado por el Gobierno Vasco (IT344-10).

Bibliografía

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

Aduriz I. y Díaz de Ilarraza A. Morphosyntactic disambiguation and shallow parsing in Computational Processing of Basque. 2003. *Inquiries into the lexicon-syntax relations in Basque*. Bernarrd Oyharçabal (Ed.). University of the Basque Country.

Aduriz I., Arriola J. M., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M. y Urkia M. 2000. *Euskararako murriztapen-gramatika: mapaketak, erregela morfosintaktikoak eta sintaktikoak*. UPV/EHU/LSI/TR12-2000

Aduriz I., Aranzabe M. J., Arriola J. M., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A. y 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. Andrew Wilson, Paul Rayson, and Dawn Archer. *Corpus Linguistics Around the World. Book series: Language and Computers.* Vol 56 (pag 1- 15). Rodopi Netherlands.

Aldezabal I., Aranzabe M.J., Díaz de Ilarraza A. y Fernández K. 2008. From Dependencies to Constituents in the Reference Corpus for the Processing of Basque. *Procesamiento del Lenguaje Natural*, nº 41 (2008), pp.147-154.

Aldezabal I., Aranzabe M. J., Arriola J. M. y Díaz de Ilarraza A. 2009. Syntactic annotation in the Reference Corpus for the Processing of Basque (EPEC): Theoretical and practical issues *Corpus Linguistics and Linguistic Theory 5-2,* 241-269. Mouton de Gruyter. Berlin-New York.

Aranzabe, M. J. 2008. Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala. [Recursos sintácticos basados en la

Gramática de Dependencias: Treebank y Gramática Computacionsl]. PhD Thesis, Euskal Filologia Saila (UPV/EHU).

Carroll J., Briscoe T. y Sanfilippo A. 1998. Parser evaluation: A survey and a new proposal. International Conference on Language Resources and Evaluations, University of Granada (Spain).

Gelbukh A., Torres S. y Calvo H. 2005. Transforming a Constituency Treebank into a Dependency Treebank. *Procesamiento del Lenguaje Natural*, (35), 145-152.

Karlsson F., Voutilainen A., Heikkilä J. y Anttila A. 1995. *Constraint grammar: A language-independent system for parsing unrestricted text.* Berlin & NewYork: Mouton de Gruyter.

Mille, S., Burga, A.,Vidal,V. y Wanner, L. 2009. Towards a Rich Dependency Annotation of Spanish Corpora. In Proceedings of SEPLN, SanSebastian.

Nilsson, J. y Hall J. 2005. Reconstruction of the Swedish Treebank Talbanken. MSI report 05067, Växjö University: School of Mathematics and Systems Engineering.

Uria L., Estarrona A., Aldezabal I., Aranzabe M. J., Díaz de ILarraza A. y Iruskieta M. 2009. Evaluation of the Syntactic Annotation in EPEC, the Reference Corpus for the Processing of Basque Lecture Notes in Computer Science (LNCS) nº 5449, Alexander Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing*. pp 72-85. Mexico City, Mexico.

Voutilainen A., Purtonen T. K. y Muhonen K. 2012. Outsourcing Parsebanking: The FinnTreeBank Project. Diana Sousa, Krister Lindén, Wanjiku Nganga (Ed.), *Shall we Play the Festschrift Game? : Essays on the Occasion of Lauri Carlson's 60th Birthday*. pp 117-131. Springer Verlag.