ESTUDIO COMPARATIVO DE DIFERENTES FORMALISMOS SINTACTICOS PARA SU APLICACION AL EUSKARA

Abaitua J., Aduriz I., Agirre E., Alegria I., Arregi X.,
Artola X., Arriola J.M. , Díaz de Ilarraza A., Eceiza N., Gojenola K.,
Maritxalar M., Sarasola K., Urkia M., Zubizarreta J.R.

ZENBAIT FORMALISMO SINTAKTIKOREN AZTERKETA KONPARATIBOA ETA EUSKARARAKO APLIKAZIOA

Abaitua J., Aduriz I., Agirre E., Alegria I., Arregi X.,
Artola X., Arriola J.M. , Díaz de Ilarraza A., Eceiza N., Gojenola K.,
Maritxalar M., Sarasola K., Urkia M., Zubizarreta J.R.

LABURPENA.

Lan honetan sintaxiaren deskripziorako formalismoetatik interesgarriak kontsideratu direnak konparatu dira euskararen tratamendurako egokitasuna aztertzeko. GPSG [Gazdar et al, 85] eta LFG [Kaplan, Bresnan, 82] formalismoak eta beren inplementazioak (ALVEY Natural Language Tools (ANLT) [Carroll et al., 91] eta GFU-LAB[Ruiz, 91]) arreta bereziz aztertu dira.

Lehen balorazio honetarako ondoko ezaugarriak kontsideratu dira: 1) perpaus-mailako osagai sintagmatikoen ordena librea, 2) aditzaren azpikategorizazioa eta 3) aditzaren komunztadura subjektu, objektu zuzena eta objektu ez zuzenarekin.

GFU-LAB sistemak ez du problemarik hiru fenomenoak deskribatzeko. ANLT sistemarekin egindako analisiaren ondorioa hiru fenomenoak tratatzeko GPSG teorian orokortzat hartzen diren printzipio batzuk birplanteiatu egin behar direla da.

HPSG [Pollard et al, 87] eta Murriztapen-Gramatika [Karlsson, 90], non egoera finituzko automatetan oinarritutako formalismo sintaktikoa proposatzen den, ere kontutan hartu dira.

GAIA. Analisi sintaktikoa.

COMPARATIVE STUDY OF SEVERAL SYNTACTIC FORMALISMS AND THEIR APPLICATION TO BASQUE

Abaitua J., Aduriz I., Agirre E., Alegria I., Arregi X.,
Artola X., Arriola J.M. , Díaz de Ilarraza A., Eceiza N., Gojenola K.,
Maritxalar M., Sarasola K., Urkia M., Zubizarreta J.R.

ABSTRACT.

This paper presents a comparative study about the adaptability of the theoretical formalisms for syntactic description that we deemed interesting, to basque. We focused mainly on GPSG [Gazdar et al., 85] and LFG [Kaplan, Bresnan, 82], and their respective implementations: ALVEY Natural Language Tools (ANLT) [Carroll et al., 91] and GFU-LAB [Ruiz, 92].

In this paper, and as a first evaluation, we focused on the following phenomena: 1) free order of syntagmatic sentence constituents, 2) verb subcategorization and 3) agreement of certain features of the verb with subject, direct object and indirect object.

LFG and GFU-LAB describe them nicely. The study of ANLT, on the other hand, suggests that to be able to address the above phenomena satisfactorily, several principles that were proposed as universal in GPSG have to be modified.

HPSG [Pollard et al., 87] and Constraint Grammar [Karlsson, 90] are also examined. The latter presents a syntactic formalism which can be implemented by finite state automata.

SUBJECT: Syntactic Analysis.

ETUDE COMPARATIVE DE DIFFERENTS FORMALISMES SYNTAXIQUES ET DE LEUR APPLICATION AU BASQUE

Abaitua J., Aduriz I., Agirre E., Alegria I., Arregi X.,
Artola X., Arriola J.M. , Díaz de Ilarraza A., Eceiza N., Gojenola K.,
Maritxalar M., Sarasola K., Urkia M., Zubizarreta J.R.

Nous presentons une étude comparative sur l'adaptation au basque des formalismes théoriques pour la description syntaxique que nous avons consideré les plus interessants. Nous avons examiné avec une attention spéciale les formalismes GPSG [Gazdar et al, 85] et LFG [Kaplan & Bresnan, 82] et leurs respectives implementations système ALVEY Natural Language Tools (ANLT) [Carroll et al.,91] et GFU-LAB [Ruiz, 91].

Dans cette première valoration nous avons examiné les caractéristiques suivantes : 1) ordre libre des composants syntagmatiques au niveau de la phrase, 2) sous-categorization verbale et 3) relations de concordance du verbe avec le sujet, objet direct et objet indirect.

GFU-LAB ne présente pas de difficultés pour la description des trois phénomènes. L'analyse realisée pour le basque avec ANLT suggère que, pour pouvoir adapter le traitement des trois phénomènes cités, il faudrait reviser quelques des principes universaux de GPSG.

Nous avons étudié aussi HPSG [Pollard et al., 87] et la Gramatique des Restrictions [Karlsson, 90], laquelle propose un formalisme syntaxique implémenté au moyen d'automates d'états finis.

ESTUDIO COMPARATIVO DE DIFERENTES FORMALISMOS SINTACTICOS PARA SU APLICACION AL EUSKARA

Abaitua J., Aduriz I., Agirre E., Alegria I., Arregi X.,
Artola X., Arriola J.M. , Díaz de Ilarraza A., Eceiza N., Gojenola K.,
Maritxalar M., Sarasola K., Urkia M., Zubizarreta J.R.

RESUMEN.

Presentamos un estudio comparativo sobre la adaptación al euskara de los formalismos teóricos para la descripción sintáctica que hemos considerado más interesantes. Hemos concedido una atención particular a los formalismos GPSG [Gazdar et al, 85] y LFG [Kaplan, Bresnan, 82] y a sus implementaciones sistema ALVEY Natural Language Tools (ANLT) [Carroll et al., 91] y GFU-LAB[Ruiz, 91].

Para esta primera valoración nos hemos limitado a examinar las características siguientes: 1) orden libre de los componentes sintagmáticos a nivel de oración, 2) subcategorización verbal y 3) relaciones de concordancia del verbo con el sujeto, objeto directo y objeto indirecto.

GFU-LAB no presenta dificultades para describir los tres fenómenos. El análisis realizado para el euskara con ANLT sugiere que, para poder adaptarse al tratamiento de los tres fenómenos citados, algunos de los principios propuestos como universales en GPSG deben ser revisados.

También se estudian HPSG [Pollard et al, 87] y la Gramática de Restricciones [Karlsson, 90], en la que se propone un formalismo sintáctico implementado mediante autómatas de estados finitos.

TEMA. Análisis sintáctico.

ESTUDIO COMPARATIVO DE DIFERENTES FORMALISMOS SINTACTICOS PARA SU APLICACION AL EUSKARA

1. INTRODUCCION

Este grupo de investigación ha planteado el procesamiento automático del euskara como un objetivo a medio/largo plazo. A corto plazo perseguimos la implementación de un entorno de herramientas lingüísticas relacionadas con los diferentes tipos de conocimiento necesarios para el procesamiento del lenguaje natural (PLN). Una vez implementado de forma satisfactoria el analizador morfológico [Agirre et al., 89; 92], el siguiente problema que afrontamos es la sintaxis. El grupo interdisciplinario lingüístico/informático intenta conjugar la creación de herramientas prácticas con la adaptación al euskara de diversas teorías y formalismos lingüísticos. En este artículo presentamos un estudio comparativo sobre la adaptación al euskara de los formalismos teóricos que hemos considerado más interesantes. Hemos concedido una atención particular a los formalismos GPSG [Gazdar et al, 85] y LFG [Kaplan, Bresnan, 82] por varias razones: están relativamente extendidos en el entorno académico, existen propuestas de análisis para una gran variedad de lenguas, existen algunos trabajos ya realizados para su adaptación al euskara [Abaitua, 88] [Trask], pero han sido examinados fundamentalmente porque contamos con herramientas para su implementación, como son GFU-LAB[Ruiz et al., 90] y el sistema ALVEY Natural Language Tools (ANLT) [Carroll et al., 91].

Se estudia también HPSG [Pollard, 87], que supone una mejora respecto a GPSG y propone soluciones en muchos casos similares a GFU-LAB. La Gramática de Restricciones [Karlsson 90] supone un nuevo enfoque, implementable mediante autómatas de estados finitos, que está siendo considerado por nuestro grupo. Las redes de transiciones recursivas (ATN) [Bates, 78] han sido descartadas de antemano por su dificultad para ser implementadas con estrategias ascendentes. Esta carencia dificulta el análisis razonable de los componentes sintácticos con recursividad a izquierdas (necesaria para el tratamiento de los sintagmas nominales en euskara, por ejemplo).

Para esta primera valoración nos hemos limitado a examinar las características que creemos más representativas del euskara en contraste con lenguas clásicamente tratadas en PLN, y las más ligadas a las aplicaciones planteadas a corto plazo. Las características seleccionadas son: 1) orden libre de los componentes sintagmáticos a nivel de oración, 2) subcategorización verbal y 3) relaciones de concordancia del verbo con el sujeto, objeto

directo y objeto indirecto. Tratar adecuadamente estos dos últimos fenómenos resulta fundamental para conseguir aplicaciones tales como correctores basados en sintaxis, y sistemas de ayuda en el análisis de corpus para el estudio de los regímenes de subcategorización verbal.

En los apartados 2, 3 y 4 se examinan los tres fenómenos citados desde el punto de vista de GPSG, LFG y HPSG. A continuación se analiza la Gramática de Restricciones y finalmente se presentan algunas conclusiones.

2. EL ORDEN DE CONSTITUYENTES

Es sabido que en euskara el orden de los constituyentes en una oración es relativamente libre. Es decir, dados cuatro componentes como sujeto, objeto, adjunto y verbo, todas sus permutaciones (veinticuatro) son posibles[1]:

```
(1)  Txakurrak egunkaria ahoan zekarren.
       perro  periódico   boca  traía
        suj      obj      adjto    v
     El perro traía el periódico en la boca
```

Hay que decir que esta flexibilidad sólo se da en el plano de la oración. El orden de las palabras dentro de los constituyentes mayores (dentro del sintagma nominal, por ejemplo) está mas restringido. Incluso en las oraciones subordinadas o de relativo el orden es mucho mas rígido (verbo al final).

El euskara se define tipológicamente como lengua con núcleo a la derecha. A consecuencia de esta propiedad el orden más natural de los constituyentes es con verbo al final (condición necesaria en las oraciones de relativo y otras subordinadas). El resto de los ordenamientos se puede explicar por alteraciones de tipo pragmático. También se explica por esta cualidad la recursividad a la izquierda de los modificadores nominales. Estas propiedades del euskara han sido debatidas en la bibliografía especializada.

Desde el punto de vista del análisis lingüístico, el problema que se plantea es si la gramática parte de unas reglas que generan una configuración básica que luego se altera mediante reglas específicas de permutación, o bien si parte de unas reglas de orden libre que

[1] Las oraciones con verbo inicial necesitan del prefijo `ba-`.

se restringen después mediante condiciones que delimiten las particularidades de cada ordenamiento. En resumen, la disyuntiva está entre:

```
(2.a)     O --> SN SV
(2.b)     O --> X* V X*
```

Las reglas de tipo (2.a) se suelen denominar reglas de bipartición, y generan una estructura configuracional. Las reglas de tipo (2.b) contienen la clásica categoría polivalente X con la estrella de Kleene, y generan estructuras no configuracionales. (2.a) necesita otra regla como:

```
(3)       SV --> SN V
```

El problema de una gramática como (2.a) es que debe añadir un complicado cuerpo de reglas de reordenación, pero además existe el problema añadido de que cierto orden de constituyentes rompe la estructura del SV:

```
(4)     Ahoan   zekarren   txakurrak    egunkaria.
     boca-adjto  traía     perro-suj  periódico-obj
```

Las reglas de corte configuracional como (2.a) y (3) se corresponden con la bipartición histórica de sujeto y predicado. En los análisis de Chomsky y la teoría de los Principios y los Parámetros las reglas de bipartición son necesarias, pues en ellas basa la teoría una serie de principios. Por ejemplo, el sujeto se reconoce como un elemento externo al SV.

2.1. GPSG (ANLT).

GPSG se caracteriza, entre otras cosas, por hacer uso de dos tipos de reglas para describir una categoría no terminal: por un lado existen reglas de dependencia inmediata (ID) que indican la estructura de una categoría sin especificar el orden de sus elementos constituyentes y por otro lado las reglas que indican las precedencias entre constituyentes (LP). Este mecanismo parecía idóneo para la descripción del orden libre en el euskara. Así, una regla ID típica de GPSG sería :

```
(5)       SV --> SN[NOR],       H[SUBCAT NORK-NOR]
                 egunkaria      zekarren
```

Esta regla no restringe el orden de los constituyentes y, por lo tanto, todas las permutaciones posibles (cuatro en este caso) estarían permitidas. Pero esta libertad en el orden tiene sus limitaciones, ya que el orden es libre sólo para los constituyentes de la regla, no pudiendo permutar entre sí constituyentes de reglas diferentes. Por ejemplo, considerando la siguiente regla :

```
(6)       O --> SN, SV
```

la oración siguiente no es generable[2]:

```
(7)       egunkaria txakurrak zekarren
             obj      suj          v
```

Una posibilidad sería proponer una regla de topicalización, pero esto nos complica antes de tiempo la gramática. Incluso de esa forma, no se pueden explicar oraciones como

```
(7')      bazekarren txakurrak egunkaria
              v        suj         obj
```

a no ser que queramos decir que el verbo está topicalizado. De cualquier manera, las reglas configuracionales se muestran poco flexibles para dar cuenta de órdenes libres. Su motivación original, identificar el sujeto de la oración, no parece suficientemente justificada para el euskara.

La solución parece pasar por un análisis con una estructura no configuracional o "plana" para los constituyentes de la oración. Así podríamos escribir, como para el inglés, una serie de reglas que reflejan el régimen de subcategorización verbal, p. ej. la regla (5) para verbos transitivos sin objeto indirecto, y mediante el uso de metarreglas propias de GPSG expandir las reglas para que incluyan todos los demás constituyentes, p. ej. sujeto y adjuntos[3]:

```
(8)  MR1: SV --> U, H   ==>  O --> SN, U, H, Adjto*
```

De esta manera no necesitaríamos la regla (6), y las reglas del sintagma verbal se convertirían, mediante la metarregla (8), en reglas del estilo de:

```
(9)  O -->   SN,       SN[NOR],   H[SUBCAT NORK-NOR],Adjto*
           txakurrak   egunkaria        zekarren      ahoan
```

Así sí es posible generar oraciones como:

```
(10) egunkaria txakurrak ahoan zekarren
        obj       suj     adjto     v
```

[2] El sujeto *txakurrak* se encuentra entre los constituyentes del predicado.

[3] U es una metavariable que se instancia con cualquier número de categorías.

Sin embargo este esquema de análisis resulta inadecuado cuando tratamos de incluir en nuestro análisis las formas verbales perifrásticas y un tratamiento del foco. El lugar del *focus* en euskara es inmediatamente anterior a la forma perifrástica del verbo (verbo principal Vp más auxiliar conjugado Vaux). Esto plantea un problema para el esquema anterior, ya que en una regla plana como la siguiente:

```
(11) O --> ... SN, Vp, Vaux ...
```

el otro componente del formalismo (reglas LP), que especifica un orden parcial entre categorías sintagmáticas, es incapaz de especificar que dos componentes sean consecutivos. Es importante señalar que, en una regla como (11), el mecanismo ID/LP no puede impedir que entre Vp y Vaux aparezca otro constituyente, ni tampoco puede especificar que el constituyente que precede a Vp sea el foco.

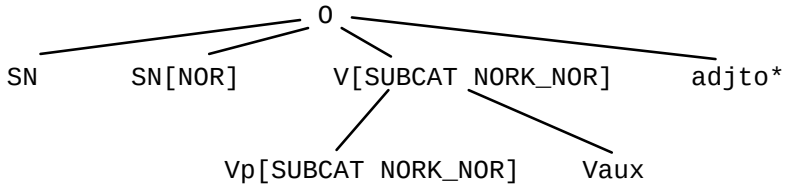
Una posible solución a este problema sería usar una regla ID separada para la forma perifrástica, junto con una regla LP para especificar el orden de los dos componentes verbales:

```
(12)      V -->Vp, Vaux
(13)      LP1[4]: Vp < Vaux
```

Esta solución sería válida si no entrara en conflicto con un principio importante de GPSG: sólo las categorías de nivel léxico son susceptibles de poseer un patrón de subcategorización. De seguir este principio, el régimen del verbo principal, especificado mediante el rasgo SUBCAT, no podría propagarse al nodo V, que es el único que tiene acceso a reglas del estilo de (5), que es donde se realiza la subcategorización.

El sistema ANLT, a pesar de ser una implementación basada en GPSG, permite una mayor flexibilidad, ya que el usuario tiene acceso a principios que en GPSG no se podrían alterar. El análisis del tratamiento del orden que ha sido presentado para GPSG es aplicable a ANLT, dado que este sistema también se adhiere a las reglas ID/LP. Lo que sí puede alterarse en ANLT es el principio que obliga al rasgo SUBCAT a permanecer en el nivel léxico. Utilizando los mecanismos de propagación a medida que ofrece ANLT se puede tratar el rasgo SUBCAT tal y como deseamos:

[4] Las categorías Vp preceden a las categorías Vaux.



A pesar de ello, y dado que ANLT limita el uso de las metarreglas al nivel léxico (al igual que GPSG), la metarregla (8) no podría activarse, y habría que escribir a mano las reglas de tipo (9), lo que nos llevaría al tedioso trabajo de duplicar todas las reglas de subcategorización verbal una por una. También se perderían los tratamientos clásicos en GPSG para la pasivización, dependencias a larga distancia, etc. que se basan en el uso de metarreglas.

Todo esto da idea de la importancia que se da en GPSG a que ciertos fenómenos se den sólo a nivel léxico, lo que va en contra de nuestro análisis. ANLT sí puede implementar ese análisis, pero a costa de perder algunas generalizaciones deseables, dado que ANLT está fuertemente influenciado por las restricciones de GPSG.

2.2. LFG (GFU-LAB).

El estudio de lenguas con orden flexible y marcas de caso desarrolladas ha motivado que en la formulación estándar de LFG [Kaplan y Bresnan, 82] se hable de la codificación no configuracional de funciones, optando por reglas de tipo (2.b) y utilizando un mecanismo de codificación que tenga en cuenta las marcas de caso.

En el formalismo GFU-LAB [Ruiz et al., 90], inspirado en LFG, una macro sintáctica se encargaría de especificar el valor funcional deseado:

```
(14) @función =    (   D/case=c erg   U/subj=D |
                       D/case=c dat   U/obj2=D |
                       D/case=c abs   U/(obj|subj)=D )
```

Esta macro declara que un constituyente con marca de caso ergativo desempeña la función de sujeto; un constituyente con marca de dativo la función de objeto indirecto y un constituyente con marca de absolutivo la función de sujeto u objeto (dependiendo de lo que rija el verbo). Esta macro aparece, por ejemplo, en la resolución de la categoría polivalente X:

```
(15) O -->      X* V X*

     VP -->     SN    (@función |
                       U/{D/pcase}=D )
```

Estas reglas, de gran sencillez, permiten dar cuenta de las veinticuatro combinaciones de (1). Para explicar las alteraciones en el orden canónico Abaitua [Abaitua, 88] propone un modelo de reglas que se inspira en las de Uszkoreit [Uszkoreit, 86] para el alemán.

Para verbos perifrásticos Abaitua [Abaitua, 85] propone una regla de esta forma:

```
X -->          &X            Vn          Vaux
           (U/foco=D )     (U=D)        (U=D)
```

donde el foco genera una dependencia de larga distancia.

2.3. HPSG.

En HPSG se sustituyen las reglas específicas para una categoría por esquemas generales. Para el euskara se puede proponer la siguiente regla para la oración[5]:

```
(16) [subcat <>] --> H [LEX+], C*
```

En esta regla no se indica ningún orden específico, es decir, se permite un ordenamiento completamente libre, siendo en la práctica equivalente a (2.b). Cada lengua tendrá sus propios principios de ordenamiento. La teoría defiende una serie de restricciones de precedencia lineal, similares a las de GPSG, y que también están basadas en la propuesta de Uszkoreit [Uszkoreit, 86]. Lo que propone es reunir en conjuntos disjuntivos las restricciones de precedencia lineal. Es decir, el orden de constituyentes se puede determinar por la interacción de varios órdenes parciales:

```
(17) [TR:AGENT] < [TR:THEME]
     [TR:AGENT] < [TR:GOAL]
     [TR:GOAL] < [TR:THEME]
     [PRON:+] < PRON:-]
```

Estas reglas se construyen sobre la información de los rasgos que componen un constituyente. En el análisis de una oración los constituyentes se van reconociendo según sus rasgos y al final se comprueba que éstos concuerden con los indicados en el verbo. La regla de caracterización de funciones de (14) se sustituye en HPSG por un algoritmo similar

[5] `subcat <>` indica que se han consumido todos los argumentos del núcleo léxico.

de unificación de estructuras parciales con la estructura del núcleo verbal, donde el verbo indica los regímenes que le son propios.

3. SUBCATEGORIZACION

El concepto de subcategorización es central a todas las teorías de sintaxis actuales. Las que se fundamentan en el mecanismo de unificación emplean recursos similares, siendo las diferencias existentes entre ellas más de notación que de contenido. En la teoría HPSG el concepto de subcategorización es la esencia de su gramática, ya que esta teoría se fundamenta sobre el concepto de núcleo (head) y toda la oración se analiza respecto a ese núcleo y la manera en que rige o subcategoriza sus complementos.

LFG, y concretamente el formalismo GFU-LAB, no es muy distinta. La estructura funcional de una oración tiene como núcleo la forma semántica que se resuelve a partir de la entrada léxica del verbo principal. Todo el complejo de ecuaciones funcionales lo único que hace es completar ese germen de estructura nuclear con la información parcial de los constituyentes que acompañan al verbo:

```
(18) zekarren =   V :   U/pred=ekarri<subj obj>
                        U/subj/caso=erg
                        U/obj/caso=abs
                        U/tmp=pasado
```

La forma conjugada que aparece en el ejemplo rige dos argumentos, que asociamos con las funciones de sujeto y objeto y cuyos casos sintácticos deben ser ergativo y absolutivo, respectivamente. Otros predicados verbales regirán complementos distintos.

```
(19.a)    ohartu = VN:  U/pred=ohartu<subj mdl>
                        U/subj/caso=abs
                        U/mdl/caso=mdl
                        U/aux=izan.

(19.b)    uste = VN:    U/pred=uste<subj comp>
                        U/subj/caso=abs
                        U/comp/tipo=ela
                        U/aux=ukan
```

En todos estos ejemplos los argumentos señalados aparecen como obligatorios. En el formalismo GFU-LAB es posible indicar funciones optativas mediante el signo *&*. Por ejemplo, la especificación de la característica facultativa del objeto indirecto, marcado por el caso dativo sería: *ekarri<subj obj &obj2>*. Los adjuntos como *ahoan* ("*en la boca*") no son regidos por el verbo en consideración y, por tanto, son tratadas como adjuntos de libre

aparición. En la entrada léxica del verbo se pueden detallar otras peculiaridades, como el tipo de rol temático o de propiedad selectiva:

```
(20) U/subj/th=agent
     U/subj/selct=anim
     U/obj/th=theme
```

Lo que de esta manera se está creando es una estructura de rasgos con información parcial. Estas estructuras definen las propiedades de subcategorización del verbo, siendo prácticamente idénticas en LFG (21.a) y en HPSG(21.b), como se muestra a continuación:

(21.a)

pred ekarri<subj obj>		
subj	caso erg	
	th agent	
	selct anim	
obj	caso abs	
	th theme	

(21.b)

PHON zekarren					
SYN|LOC	HEAD	MAJ V			
		VFORM FIN			
	SUBCAT<	MAJ SN(1)		MAJ SN(2)	>
		CASE ERG		CASE ABS	
SEM|CONT	RELN EKARRI				
	ARG1 (1)				
	ARG2 (2)				

El tratamiento dado en GPSG para la subcategorización consiste en utilizar un rasgo de nombre SUBCAT con un valor que sirve para indexar las reglas de un determinado esquema de subcategorización. Este rasgo permite enlazar la información relativa a la subcategorización (información léxica) con las reglas ID apropiadas. Por ello se propone que el rasgo SUBCAT sólo puede aparecer en elementos preterminales. Como consecuencia, los elementos subcategorizados deben aparecer como hijos de un mismo constituyente. Por ejemplo, para describir la subcategorización de un objeto por un verbo (no se va a entrar en la discusión sobre si el sujeto es subcategorizable por el verbo en euskara):

```
(22) SV --> SN[NOR]  H[SUBCAT NORK_NOR]
```

La entrada léxica para un verbo de este tipo podría ser:

```
(23) zekarren: [ CAT V
                 SUBCAT NORK_NOR ]
```

Esta solución no parece satisfactoria en el caso del euskara, al entrar en conflicto, como se ha explicado anteriormente, con las reglas que determinan el orden de constituyentes. También parece que esta información sobre subcategorización está relacionada con la concordancia entre constituyentes, tal y como se explica más adelante. Pero en GPSG ambos fenómenos se tratan de forma totalmente independiente.

4. CONCORDANCIA

En euskara existe concordancia entre el verbo y los constituyentes cuyos casos son ergativo, absolutivo y dativo. La concordancia en número y persona se resuelve como una extensión de la noción de subcategorización. Esta información se añade a la estructura de rasgos de la forma conjugada en su entrada léxica. Así la forma vasca *zekarzkigun*, literalmente "*él/ella nos los traía*"

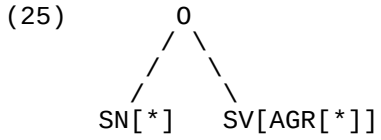
```
(24) zekarzkigun = V:   U/pred=ekarri<subj obj obj2>
                        U/subj/caso=erg
                        U/subj/per=3
                        U/subj/num=sg

                        U/obj/caso=abs
                        U/obj/per=3
                        U/obj/num=pl

                        U/obj2/caso=dat
                        U/obj2/per=1
                        U/obj2/num=pl
                        U/tmp=pasado
```

Es decir, la estructura parcial del verbo es en sí misma una estructura con gran información respecto a las funciones que rige.

En GPSG se describe un principio universal, el CAP (Control Agreement Principle), que se encarga de asegurar la concordancia entre constituyentes de una oración. Este principio se define en función de los tipos semánticos de las categorías que intervienen en un árbol local (árbol correspondiente a la aplicación de una regla de dominación inmediata o ID). A grandes rasgos, este principio dice que las categorías nominales (controladoras) en un árbol local deben unificarse con el valor del rasgo AGR (agreement) en caso de que haya una categoría controlada (verbos, adjetivos o determinantes) tal que su forma semántica sea un funtor cuyos argumentos sean las categorías controladoras. Por ejemplo:



De esta forma se asegura la concordancia entre sujeto y verbo, entre determinante y nombre, etc. Esta idea es útil para lenguajes como el inglés, que tienen una concordancia entre constituyentes muy limitada. En euskara el verbo puede concordar en número y persona con tres constituyentes (sujeto, objeto y objeto indirecto), con lo que el tener un único rasgo AGR para este cometido no es suficiente. En [Gazdar et al., 85:107] se propone una extensión de este mecanismo para que se permita al rasgo AGR tomar una secuencia de categorías como valor en el caso de lenguas como el euskara. De esta forma la estructura resultante podría ser, por ejemplo:

```
(26) zekarren: V  [ AGR [[CASE ERG, NUM SING, PER 3]
                         [CASE NOM, NUM SING, PER 3]]    ]
```

Las implicaciones que esta modificación conlleva no son triviales porque, como se señala en [Gazdar et al., 85: 107], esto lleva a que se duplique información codificada en el rasgo que controla la subcategorización SUBCAT. La solución que se sugiere consiste en juntar la información de SUBCAT y AGR en un solo rasgo. Esta idea será recogida posteriormente en el formalismo HPSG, donde toda la información sobre subcategorización y concordancia (así como otros tipos de información) de los verbos se agrupa en el rasgo de nombre SUBCAT.

5. GRAMATICA DE RESTRICCIONES.

Un aspecto interesante de este enfoque [Karlsson, 90] es la forma de abordar el problema del análisis sintáctico que puede dividirse en cuatro fases integradas: análisis morfológico, desambiguación morfológica, determinación de los límites intrasentenciales y asignación de funciones sintácticas superficiales. Los fundamentos sobre los que se asienta el "parser" que aquí explicamos son los siguientes: a) no va a distinguir entre frases gramaticalmente correctas e incorrectas y b) va a ser capaz de analizar un amplio conjunto de estructuras para lo que también contará con un amplio diccionario.

El "parser" contiene reglas de tres tipos diferentes correspondientes a las fases mencionadas: reglas dependientes del contexto para la desambiguación morfológica, reglas

para determinar los límites entre proposiciones intrasentenciales y reglas para la asignación de funciones sintácticas superficiales.

Las reglas correspondientes al primer grupo pueden hacer referencia a una interpretación simple de una palabra o a una clase de interpretaciones definida mediante una característica gramatical compleja. Hablar de la interpretación morfológica *wi* de una palabra significa que nos referimos a la tripleta <palabra, lema, características>. Bajo el nombre de características se engloban todos aquellos símbolos utilizados en la descripción gramatical y obtenidos por el analizador morfológico así como funciones sintácticas y otros aspectos gramaticales. Por ejemplo: categorias morfológicas (N, A, ADV, CONJ, PAST, SG, PL), características morfosintácticas (regido por la preposición p, etc.), funciones sintácticas (objeto directo, sujeto, etc.). La estructura general de las reglas es la siguiente:

```
wip Op-id / lc_rc
```

wip, *lc* y *rc* pueden ser bien átomos o bien objetos compuestos; *wip* hace referencia a una interpretación de palabra y *lc* y *rc* especifican el contexto situado a la izquierda y a la derecha, respectivamente, de la palabra cuya interpretación se está tratando. *Op-id* corresponde a la identificación de la operación a aplicar sobre la interpretación *wip*.

Hay distintos problemas no resueltos con este formalismo. En primer lugar el establecer estas reglas requiere un cuidadoso estudio estadístico de un corpus grande; en algunos casos se podrán determinar los límites entre proposición anidadas, pero no siempre.

La asignación de funciones sintácticas se basa en el principio siguiente: "En una oración simple no coordinada puede haber como máximo un verbo conjugado, un sujeto y un objeto".

Es interesante resaltar el hecho de que la estructura sintáctica que aquí se propone es horizontal y superficial, y está basada en las palabras que aparecen en la frase. Para cada palabra de la frase es necesario indicar si es núcleo o modificador y en este último caso dónde se encuentra su núcleo. En el sistema tratado se definen distintas etiquetas para los núcleos de los sintagmas presentes en la frase (SUBJ, OBJ, MAINPRED, ADVL, etc. son algunos de ellos y se les considera en el mismo nivel en la estructura). Las etiquetas de los modificadores (N/, N\, GEN-N/, A/, A\, etc) indican si el elemento que modifican se encuentra a su izquierda, para lo que se utilizará el símbolo "\", o a su derecha, "/". Para

poder realizar la asignación de funciones sintácticas se define el operador "=:@". Mediante este operador se asigna la función sintáctica determinada por "@" a la palabra cuya interpretación se está examinando. Para finalizar, interesa recalcar que lo que aquí se presenta es una teoría que ha de ser formalizada e implementada y que ha de ser probada para lenguas diferentes. Hasta el momento ha sido experimentada con lenguas tales como el finlandés, sueco e inglés. Para poder expresar los diferentes fenómenos linguísticos son necesarias un gran número de reglas, pero hay que tener en cuenta que se pretende una amplia cobertura de la lengua tratada y que las reglas son obtenidas con métodos estadísticos.

Mostraremos a continuación algunos ejemplos sobre los diferentes tipos de reglas diseñadas para el tratamiento del euskara. En primer lugar veremos dos reglas de desambiguación morfológica:

```
(27.a)    $ERG =0/ $ERG * __
(27.a)    $ERG =0/ __ * $ERG
```

Las reglas (27.a) y (27.b) indican que nunca se encontrará más de un nombre en caso ergativo. En caso de encontrarnos ante una palabra con más de un análisis morfológico posible, correspondiendo uno de ellos al caso ergativo, esa posibilidad queda rechazada si se encuentra antes o después de dicha palabra alguna otra con una interpretación morfológica en caso ergativo. Por ejemplo:

```
(27') Txakurrek    egunkariak    dakartzate.
      perro-erg  periódico-abs     traer

       Los perros traen los periódicos.

       Resultado del analisis morfológico:
             txakurrek  (ERG PL)
             egunkariak (ABS PL)
                        (ERG SG)
```

Donde *egunkariak* puede ser absolutivo o ergativo, pero esta última posibilidad es eliminada porque *txakurrek* sólo puede ser ergativo.

La regla (28) asegura la concordancia entre sujeto y verbo para el caso de los verbos transitivos. Suponiendo que ERG1 representa las palabras en caso ergativo correspondientes a la primera persona y que VNORK1 denota un verbo transitivo en primera persona.

```
(28)      $ERG1 =!!/ __ * VNORK1
```

Para la asignación de funciones sintácticas utilizamos las siguientes:

```
(29.a)    ZENB-MUGAG =:I/
(29.b)    A :=I\ / I<mugag-dekgabe> __
(29.c)    I<erg> :=SUBJnork
```

La regla (29.a) establece que cualquier ordinal no declinado (ZENB-MUGAG) que aparezca en una frase es un modificador, forma parte de un sintagma y su núcleo es el nombre que se encuentra a su derecha. La regla (29.b) establece que cualquier adjetivo que tenga a su izquierda un nombre sin declinar es modificador de éste. La regla (29.c) establece que cualquier nombre en caso ergativo hace la función de núcleo del sintagma nominal sujeto. Es interesante resaltar que se puede encontrar en cualquier punto de la frase.

6. CONCLUSIONES.

El objetivo del estudio realizado ha sido la evaluación de la idoneidad para la descripción del euskara de las implementaciones de los formalismos GFU-LAB (inspirado en la teoría LFG) y ANLT (basado en GPSG), así como del formalismo HPSG y de la Gramática de Restricciones. Para los dos primeros se ha realizado un estudio profundo de tres fenómenos lingüísticos como son el orden de constituyentes en una oración, la subcategorización verbal y la concordancia entre constituyentes.

El estudio sugiere que el resultado del análisis de oraciones ha de ser una estructura relativamente "plana" o no configuracional, en la que se admite un orden libre de constituyentes. Un formalismo que trate de describir el euskara deberá adecuarse a estas características.

GFU-LAB no presenta dificultades para describir los tres fenómenos. Esto no es sorprendente, ya que tanto el sistema como el formalismo subyacente fueron diseñados para tratar lenguas como el euskara. Por otro lado el sistema, basado en una gramática libre de contexto a la que se asocian ecuaciones de unificación, presenta la ventaja de la flexibilidad y la facilidad de utilización para usuarios no familiarizados con el formalismo.

El sistema ANLT, a pesar de ser una implementación basada en GPSG, permite una mayor flexibilidad definitoria que ésta, ya que el usuario tiene posibilidad de modificar los principios intrínsecos de GPSG. El análisis realizado con el euskara sugiere que, para poder adaptarse al tratamiento de los tres fenómenos citados, algunos de los principios propuestos como universales en GPSG deben ser revisados : el que el rasgo SUBCAT sólo pueda aparecer a nivel léxico, la cardinalidad del rasgo AGR y los principios que gobiernan las propagaciones de ambos rasgos. ANLT sí puede implementar estas modificaciones, pero a costa de perder algunas generalizaciones deseables, al estar fuertemente influenciado por las restricciones de GPSG. Por otro lado, el complejo entramado de reglas ID/LP, metarreglas, principios universales y restricciones de rasgos hace que el uso del sistema exija una comprensión mínima de la teoría GPSG, tarea ésta no trivial incluso para usuarios con formación lingüística. Sin embargo, el hecho de haberse descrito un extenso subconjunto de la lengua inglesa con este sistema sugiere que su potencia es suficiente para una implementación de un sistema real (aún cuando haya problemas relativos a su ineficiencia).

Dado que las gramáticas HPSG son un desarrollo posterior a GPSG y LFG, y puesto que integran aspectos de ambas, en teoría parece que los fenómenos considerados podrán ser descritos, pero no hemos podido corroborarlo en la práctica.

La Gramática de Restricciones parte de un planteamiento diferente al no basarse en gramáticas libres de contexto, sino en reglas codificables como autómatas de estados finitos. La información de origen morfológico juega un papel importante en el proceso de análisis y desambiguación. Las reglas se obtienen mediante procesos de análisis de corpora teniendo como objetivo el tratamiento de textos reales. La implementación mediante autómatas le confiere una enorme eficiencia.

Como próximo paso se intentará extender el análisis lingüístico al tratamiento del sintagma nominal (complementos del nombre, oraciones de relativo), y a la tipología oracional (oraciones interrogativas, negativas, ...). Se estudiarán también oraciones complejas, con fenómenos de coordinación y subordinación. El propósito es escribir una gramática para un tratamiento amplio de la sintaxis vasca.

REFERENCIAS

[Abaitua, J., 1985]. "*An LFG parser for Basque*", Univ. de Manchester, tesis de Msc.

[Abaitua, J., 1988]. "*Complex predicates in Basque: from lexical forms to functional structures*", Universidad de Manchester, tesis doctoral.

[Agirre E., Alegría I., Arregi X., Artola X., Díaz de Ilarraza A., Sarasola K., Urkia M., 1989]. *"Aplicación de la morfología de dos niveles al Euskara",* En Procesamiento del Lenguaje Natural. (Sociedad Española para el Procesamiento del Lenguaje Natural), Boletín nº 8, pp 87 - 102, Barcelona.

[Agirre E., Alegría I., Arregi X., Artola X., Díaz de Ilarraza A., Maritxalar M., Sarasola K., Urkia M., 1992].*"XUXEN: A spelling checker/corrector for Basque based on two-level morphology"*. Proceedings of the 3rd Conf. on Applied Natural Language Processing (ANLP'92) Trento 119-125.

[Bates, M., 1978]. "*The theory and practice of ATN grammars*", en L. Bolc (ed.) "Natural Language Communication with Computers", Springer Verlag, Berlin.

[Carroll, J, T. Briscoe, C. Grover, 1991]. "*A development environment for large natural language grammars*", Technical Report No. 233, Computer Laboratory, University of Cambridge.

[Gazdar, G., E.Klein, G.Pullum, I. Sag, 85, 1985]. "*Generalized Phrase Structure Grammars*", Blackwell, Oxford.

[Goenaga, P., 1980] "*Gramatika bideetan*", 2. argitalpena, EREIN, Donostia.

[Kaplan, R., J.Bresnan, 1982]. "*Lexical-Functional Grammar: a formal system for grammatical representation*", en J. Bresnan (ed.) "The mental representation of grammatical relations".

[Karlsson, F. 90]. "*Constraint Grammar as a framework for parsing running text*", Proccedings of the 13th International Conference on Computational Linguistics, Vol .3, pp. 168-173, Helsinki.

[Pollard, C., I. Sag, 1987]. "*An Information-Based Approach to Syntax and Semantics: Volume 1 Fundamentals*", CSLI Lecture Notes No 13, Chicago University Press: Chicago.

[Ruiz, J. C., 1991]. "*Gramática Funcional de Unificación: un formalismo para el tratamiento computacional de la sintaxis y la semántica*".

[Ruiz, J.C., J.R. Zubizarreta, J. Abaitua, 1990]. "*Un compilador de LFG y su aplicación al euskara*", en actas del VI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, 1990.

[Trask, L., 1983]. "*Euskal izen sintagmaren egituraz*", en "Iker-2: Piarres Lafiteri Omenaldia", Euskaltzaindia, Bilbo.

[Uszkoreit 86]. "*Constraints on order*", Linguistics, 25, pp 883:906