Migración de una gramática sintáctica parcial entre dos formalismos de unificación

Itziar Aduriz
Universidad de Barcelona
Facultad de Filología
jiradagi@si.ehu.es

Itsaso Esparza
Universidad del País Vasco
UPV/EHU
Facultad de Informática de San Sebastián
jibeslei@si.ehu.es

Kepa Bengoetxea, Koldo Gojenola
Universidad del País Vasco
UPV/EHU
Escuela Universitaria de Ingeniería Técnica Industrial de Bilbao
{kepa.bengoetxea, koldo.gojenola}@ehu.es

Resumen: Este trabajo presenta el proceso de migración de una gramática sintáctica del euskera de un formalismo a otro. Debido a diferencias en los formalismos y también en el tipo de gramáticas, la transición directa de una gramática a otra no es posible. Esto lleva a que la construcción de la nueva gramática por parte de un lingüista parta prácticamente de cero. Por ello se ha planteado, de manera paralela a la construcción manual de la gramática, un experimento consistente en derivar una gramática de manera semiautomática generando reglas partiendo de la gramática antigua y un corpus analizado con ésta. Este experimento ha servido por un lado para comprobar la viabilidad de obtener una nueva gramática de manera prácticamente automática, y a la vez ha valido para ayudar en el proceso de construcción manual de la gramática, sirviendo de punto de comparación y para detección de errores u omisiones.
Palabras clave: Análisis sintáctico, unificación, procesamiento de corpus.

Abstract: This work presents the migration process of a syntactic grammar of Basque from one formalism to another. Due to differences in the formalisms and the kind of grammars, it is not possible to make a direct translation. As a consequence, the construction of a new grammar by a linguist must start almost from scratch. For this reason we devised an experiment in parallel with the manual construction of the grammar, consisting in deriving several grammars semi automatically using the old grammar and a corpus analyzed with it. This experiment was useful to test the viability of obtaining a grammar automatically and at the same time also helped in the process of the manual construction of the new grammar, as the automatically obtained grammars could be compared with the manual one, and could also help to detect errors or omissions.
Keywords: Syntactic analysis, unification, corpus processing.

1 Introducción

Actualmente existen numerosos formalismos sintácticos para la descripción de gramáticas y la obtención de sus correspondientes analizadores sintácticos. Estos formalismos basan su diseño e implementación en diferentes técnicas, desde autómatas, transductores, gramáticas independientes del contexto hasta los más completos formalismos de unificación. Debido a razones de eficiencia, búsqueda de nueva funcionalidad u otros motivos, es relativamente frecuente migrar una gramática de un formalismo/analizador a otro. Este proceso es costoso, al requerir un gran esfuerzo, con la intervención de personal lingüístico e informático.

En este trabajo se va a estudiar la posibilidad de realizar la migración de un formalismo a otro de manera semiautomática, de forma que se minimice el trabajo lingüístico a realizar, y se comparará el resultado del analizador resultante con otro obtenido mediante la codificación manual de una gramática por un lingüista. Se aplicará a una gramática parcial del euskera ya existente. Con objeto de delimitar más precisamente el experimento a realizar, se ha escogido como objeto de este trabajo la parte de la gramática correspondiente a los sintagmas nominales simples, formados por combinaciones de nombres, adjetivos,

1. *Zenbait etxe handi-ekin* (algunas) (casa) (grandes)-(con) Con algunas casas grandes
2. *Atzerri-ko hiri handi bat-eko etxe hori-ei* (extranjero)-(de) (ciudad) (grande) (una)-(de) (casa) (amarilla)-(a) A las casas amarillas de una ciudad grande del extranjero
3. *15 urte-ko ikasle-en ezagutza-k* (años)-(de) (estudiantes)-(de) (conocimiento)-(s) Los conocimientos de estudiantes de 15 años
4. *EAE-ko Eskola Kontseilu-ko presidente-arentzat* (CAV)-(de) (escuela) (Consejo)-(de) (presidente)-(para) Para el presidente del Consejo Escolar de la CAV

Tabla 1. Ejemplos de sintagmas a analizar.

determinantes, y pronombres. No se han tenido en cuenta sintagmas nominales con complementos oracionales, coordinación o postposiciones complejas (estructuras formadas por la combinación de un morfema de caso con una palabra[1]). La tabla 1 muestra ejemplos de sintagmas a analizar, junto con su correspondiente traducción. Los sintagmas nominales del euskera corresponden a los sintagmas nominales y preposicionales de lenguas como el español o inglés. Estos sintagmas tienen una marca de caso al final del sintagma, precedida de nombres, adjetivos, determinantes, pronombres o adverbios. Un caso a destacar son los complementos del nombre formados con uno de los casos genitivos (sufijos *–en* y *–ko*), y que pueden ser aplicados de manera recursiva (ver ejemplos 2, 3 y 4 de la tabla 1).

La migración se realizará entre dos formalismos de unificación, PATR (Shieber, 1986) y RASP (*Robust Accurate Statistical Parsing*, Briscoe y Carrol, 2002). Ambos son formalismos relativamente cercanos, y comparten ciertos rasgos, al estar basados en gramáticas independientes del contexto y en unificación.

PATR es uno de los formalismos basados en unificación más sencillos para la definición de gramáticas. Existen varias implementaciones de este formalismo. En el caso del euskera, se dispone de una gramática parcial para el euskera, comprendiendo sintagmas nominales, cadenas verbales, oraciones simples y ciertos casos de subordinación. Esta gramática ha sido usada previamente como *chunker*, aplicándola a tareas como la obtención de patrones de subcategorización verbal (Aldezabal et al., 2000, 2003). El principal problema de este analizador es que funciona satisfactoriamente como *chunker*, al estar la ambigüedad delimitada a nivel del sintagma nominal, pero a la hora de abordar el análisis de oraciones completas aparecen problemas de ineficiencia (tiempo y espacio) y de ambigüedad (cientos o miles de análisis para una oración). Esto lleva a tener que ampliar el analizador actual incluyendo mejoras de eficiencia y añadiendo un módulo de tratamiento probabilístico, o bien a cambiar de formalismo, que ha sido en este caso la opción elegida.

El sistema RASP estuvo inicialmente basado en una implementación del formalismo GPSG (*Generalized Phrase Structure Grammar*, Gazdar et al., 1985) plasmada en el sistema *Alvey Natural Language Tools* (Briscoe y Carroll, 1993). Esta primera versión ha sufrido varias modificaciones, consistentes básicamente en disminuir la complejidad original del modelo (eliminando apartados como meta reglas o reglas de precedencia lineal), aumentar la eficiencia y añadir nuevas funcionalidades, como un mecanismo de entrenamiento estadístico.

Aunque los dos formalismos comparten varias características, los sistemas tienen importantes diferencias:

- La gramática PATR está construida a nivel de morfema, basándose en los análisis "clásicos" del euskera (Goenaga, 1980). Por el contrario, el formalismo RASP se ha definido a nivel de palabra. Como ejemplo, la figura 1 muestra el análisis de un mismo sintagma nominal basándose en morfemas (parte superior de la figura) o en palabras (parte inferior). Se puede apreciar que hay diferencias considerables en los análisis obtenidos. Mientras que el análisis a nivel de

[1] El equivalente en español de las postposiciones podrían ser las preposiciones compuestas (p. ej. 'a través de').

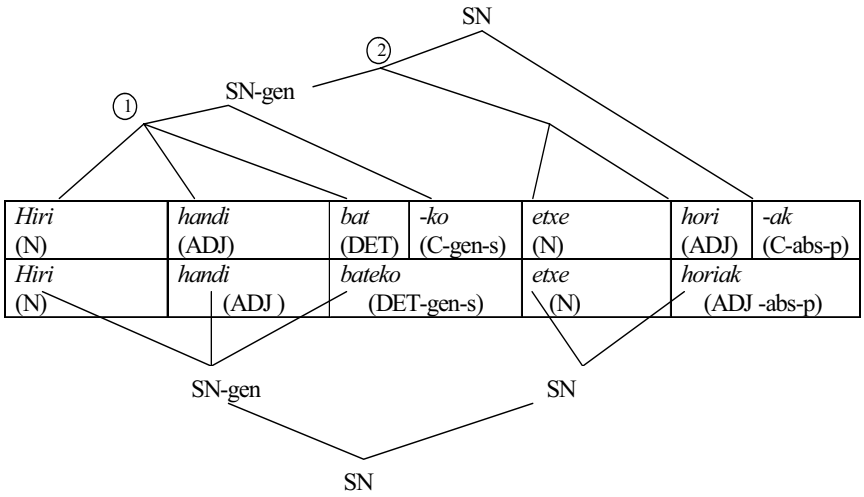

Figura 1. Análisis basado en morfemas (arriba) frente a análisis basado en palabras (debajo). C = marca de caso, gen = genitivo, abs = absolutivo, s = singular, p = plural

morfema agrupa un morfema con un conjunto de palabras (ver nodos del árbol etiquetados con 1 y 2), el análisis a nivel de palabra plantea la combinación de palabras completas. Aunque las gramáticas tradicionales del euskera se definen tomando el morfema como unidad básica de análisis, la gran mayoría de los sistemas de análisis existentes están basados en palabras. Este hecho plantea la opción de modificar los sistemas existentes para que funcionen a nivel de morfema, opción muy costosa e incluso imposible en muchos casos, o de adaptar el análisis lingüístico al nivel de la palabra, como se ha hecho en el caso de la desambiguación morfológica (Aduriz et al., 1997).

- Las reglas de PATR son más complejas, lo que lleva a que una regla en el formalismo PATR puede llevar a un gran número de reglas en RASP. La tabla 2 muestra un ejemplo de una misma regla en ambos formalismos. El formalismo PATR estructura las reglas como una colección de ecuaciones, mientras que RASP codifica la regla mediante estructuras de rasgos que pueden compartir valores mediante variables (valores que comienzan con el símbolo @).

```
X0 → X1 X2
   X0/Cat  = sn
   X1/Cat  = det
   X2/Cat  = n
   X0/Caso = X2/Caso
   X0/Num  = X2/Num
   X1/Num  = X2/Num
```

```
[Cat sn, Caso @1, Num @2] →
       [Cat det, Num @2]
       [Cat n, Caso @1, Num @2]
```

Tabla 2. Ejemplos de la regla '*sn → det n*' en PATR (arriba) y RASP (abajo).

En concreto, la implementación particular de PATR usada (Aldezabal, Gojenola y Sarasola, 2000) ha sido dotada de mecanismos que le dotan de mayor flexibilidad, como el uso de la conjunción y disyunción de ecuaciones. Esto hace la gramática más potente a la vez que aumenta la complejidad de las reglas individuales. Por el contrario, RASP también tiene sus propios mecanismos de manejo de la complejidad, como mecanismos de propagación de valores y definición de alias. Como resumen, podemos decir que la gramática PATR tendrá un número de reglas sensiblemente inferior a RASP, aunque las reglas de RASP son más legibles y concisas que las de PATR.

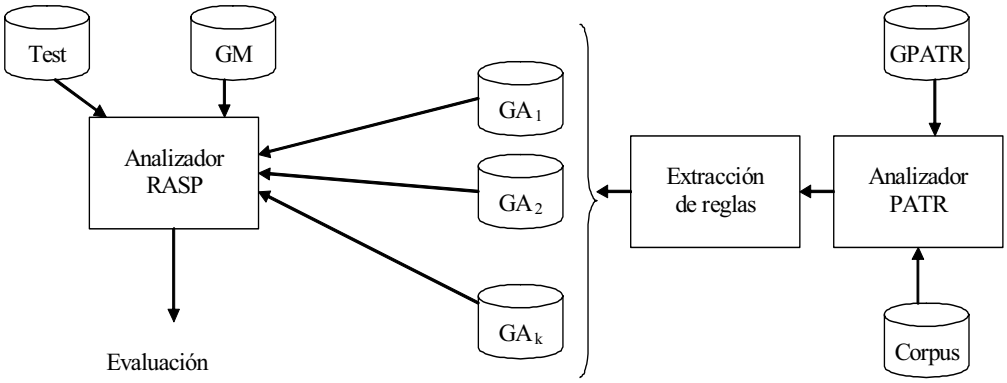

Figura 2. Esquema general del trabajo.

Estos dos factores llevan a que la traducción de reglas de un formalismo a otro sea una tarea no trivial ni directa. Aunque la gramática PATR existente pueda ser usada como consulta, el desarrollo de una gramática RASP por un lingüista supone realizar el diseño partiendo prácticamente de cero.

2 Descripción del sistema

La figura 2 muestra una descripción general del trabajo realizado. Se dispone de una gramática implementada en el formalismo PATR (GPATR) y se quiere desarrollar una gramática en el formalismo RASP (equivalente o de mayor cobertura que la original). Por un lado se ha desarrollado una nueva gramática de manera manual (GM) en el formalismo RASP. En paralelo, se ha aplicado la gramática antigua (GPATR) a un corpus, obteniéndose un conjunto de reglas de manera automática, a partir de los patrones de sintagmas nominales encontrados en el corpus. Aplicando diferentes criterios y diferentes tamaños de corpus se han obtenido varias gramáticas RASP de manera automática (ver sección 3). Finalmente, se han comparado la gramática manual (GM) y las automáticas (GA_1, …GA_k) frente a un conjunto de ejemplos analizados manualmente, para determinar la cobertura de la gramática correspondiente.

3 Desarrollo del nuevo analizador

En esta sección se describirá el proceso de obtención de la gramática RASP desarrollada manualmente (apartado 3.1) y las gramáticas obtenidas de manera automática (apartado 3.2).

3.1 Gramática manual

El desarrollo de la gramática manual se ha llevado a cabo mediante el examen de corpus y gramáticas del euskera (Goenaga, 1980; Zubiri y Zubiri, 2000). La gramática desarrollada hasta el momento tiene un total de 59 reglas, más un conjunto de reglas de propagación automática de rasgos. Este trabajo ha sido desarrollado por una lingüista durante un periodo de tres semanas a dedicación completa.

3.2 Gramáticas (semi)automáticas

Para las gramáticas obtenidas de manera automática, se tuvieron en cuenta los siguientes aspectos:

- Elección de los terminales a considerar. Cada elemento que forma un sintagma nominal (nombre, adjetivo, adverbio, determinante, pronombre) tiene una serie de rasgos asociados. Se han escogido cinco rasgos para representar estos elementos: categoría, subcategoría, definido/indefinido, número[2] y caso. El uso de todos o parte de los rasgos dará lugar a un número diferente de reglas. En nuestro trabajo

[2] La definición del número en euskera se hace normalmente basándose en dos rasgos: definido/indefinido y singular/plural. Esto da lugar a tres posibles combinaciones de valores: definido-singular, definido-plural e indefinido.

Sintagma analizado	***Atzerriko*** extranjero (N-gen-s)	***hiri*** ciudad (N)	***handi*** grande (ADJ)	***bateko*** una (DET-gen-s)	***etxe*** casa (N)	***horiak*** amarillas (ADJ-abs-p)
	Las casas amarillas de una ciudad grande del extranjero					
a) Patrones obtenidos (todos los rasgos: +num, -gen)	SN → N-gen-s N ADJ DET-gen-s N ADJ-caso-p					
b) Patrones obtenidos (solo categoría, subcategoría: -num, -gen)	SN → N-gen N ADJ DET-gen N ADJ-caso					
c) Patrones obtenidos (todos los rasgos + reglas genitivo: +num, +gen)	SN → SN-gen N ADJ-caso-p SN-gen → N-gen-s SN-gen → SN-gen N ADJ DET-gen-s					
d) Patrones obtenidos (solo categoría, subcategoría + reglas genitivo: -num, +gen)	SN → SN-Gen N ADJ-caso SN-gen → N-gen SN-gen → SN-gen N ADJ DET-gen					

Tabla 3. Ejemplo de sintagma nominal y posibles patrones de reglas RASP.

hemos distinguido dos alternativas: todos los rasgos o únicamente categoría+subcategoría+caso. La primera opción pretende capturar fenómenos de concordancia entre diferentes palabras, al precio de aumentar el número de posibles patrones a abarcar. La segunda es más compacta, produciendo menos reglas y patrones más generales, aunque a priori podría presentar el problema de la sobregeneración, al no tener en cuenta posibles problemas de concordancia en número. Esta distinción es similar a la descrita por Cowan y Collins (2005) para el análisis del español.

- Tipo de reglas a extraer. La alternativa más sencilla consiste en extraer todos los patrones encontrados y asociar una regla diferente a cada uno. De esta manera se contará con un conjunto de reglas extraídas directamente de los patrones de sintagmas nominales encontrados en un corpus. La cobertura de la gramática mejorará aumentando el tamaño del corpus de extracción. El esquema de las reglas extraídas es:

  (1) SN → elemento elemento(caso)*

  Donde *elemento* hace referencia a nombre, adjetivo, adverbio, determinante o pronombre, junto con sus rasgos asociados. El último elemento de un sintagma nominal tiene siempre una marca de caso (los casos del euskera vienen a coincidir básicamente con las preposiciones del español). Hay 15 casos definidos para el euskera.

  Otro aspecto que se ha tenido en cuenta es que los complementos del nombre, formados por el uso de las dos marcas de caso del genitivo pueden aplicarse de manera recursiva. Este hecho puede servir para disminuir el número de reglas extraídas y también para que la gramática resultante pueda generalizar mejor la estructura de sintagmas nominales complejos. Teniendo esto en cuenta, el esquema de las dos clases de reglas extraídas será:

  (2) SN → elemento elemento(caso)*
  SN-Gen → elemento elemento(genitivo)*

  Donde *elemento* puede ser ahora cualquiera de las categorías básicas más la categoría sintáctica sintagma nominal genitivo (SN-gen) correspondiente a un complemento del nombre. El hecho de que las reglas obtenidas sean ahora recursivas puede ayudar a mejorar la cobertura de la gramática, debido a su mayor generalización, y también a reducir considerablemente el número de reglas.

Teniendo en cuenta estas dos opciones, se puede producir una gramática de cuatro maneras diferentes:

- usando todos los rasgos para definir los elementos léxicos de la gramática (+num),
- usando únicamente la categoría y subcategoría (-num),
- usando o no reglas específicas para los genitivos (+gen y -gen).

La tabla 3 muestra las 4 diferentes posibilidades de extracción de reglas sobre un patrón de un sintagma nominal analizado con la gramática PATR. Aunque las opciones c) y d) generan más de una regla, su efecto global es reducir en gran medida el número total de reglas, ya que se produce una mayor generalización en la gramática resultante (solo se guarda una instancia de las reglas repetidas).

Para la extracción de las reglas de manera automática, se ha experimentado con tres tamaños diferentes del corpus de extracción. Esto permitirá evaluar el impacto del tamaño del corpus de extracción sobre la gramática resultante. Como hipótesis inicial se puede pensar que un mayor tamaño del corpus proporcionará una mayor cobertura, al tenerse en cuenta un mayor número de patrones de reglas. Por otro lado, esto puede tener el efecto de producir sobregeneración, esto es, que la gramática proporcione ambigüedad en forma de análisis redundantes.

		Número de reglas
Gramática manual (GM)		59
Gramática automática. Corpus 1 (C1) 1.300 palabras	a) -gen +num	199
	b) -gen -num	160
	c) +gen +num	129
	d) +gen -num	81
Gramática automática. Corpus 2 (C2) 10.000 palabras	a) -gen +num	968
	b) -gen -num	781
	c) +gen +num	406
	d) +gen -num	223
Gramática automática. Corpus 3 (C3) 70.000 palabras	a) -gen +num	3.807
	b) -gen -num	3.073
	c) +gen +num	850
	d) +gen -num	455

Tabla 4. Relación entre los criterios de obtención de patrones y el tamaño de la gramática resultante.

Se tomaron como corpus de extracción de patrones textos periodísticos de 1.300, 10.000 y 70.000 palabras, donde cada corpus comprende al anterior. Cada uno de los corpus dio lugar a cuatro gramáticas diferentes, de acuerdo a los criterios apuntados en la tabla 3. Esto dio lugar a 12 gramáticas diferentes a comparar con la gramática desarrollada manualmente.

La tabla 4 muestra el número de reglas de cada gramática. Se observa que el número de reglas aumenta con el tamaño del corpus de extracción. Por otro lado, la eliminación de restricciones de número (-num) y la inclusión de reglas para el tratamiento del genitivo (+gen) disminuyen drásticamente el tamaño de la gramática resultante. En un principio, una gramática con más reglas es candidata a tener una mayor precisión, al reconocerse la presencia de unos patrones muy concretos, mientras que puede tener el inconveniente de no obtener ningún análisis para patrones que no han aparecido en el corpus. Por otro lado, las gramáticas más compactas pueden tener el inconveniente de la sobregeneración, es decir, generar análisis innecesarios, al ser menos restrictivas y generalizar más de lo necesario.

Este trabajo puede encuadrarse en la línea de obtención de gramáticas a partir de *treebanks* (Charniak, 1996) con la diferencia de que aquí no se toma un *treebank* como punto de partida, sino un corpus analizado con una gramática. Por otro lado, (Krotov et al., 1998) muestra cómo se puede reducir de manera considerable el número de reglas extraídas (pasa de 16.000 a 6.000 reglas) sin efectos destacables en la cobertura y precisión, basándose en la eliminación de reglas redundantes. En nuestro caso, sin embargo, el método de compactación de reglas usado está basado en un mínimo conocimiento lingüístico del euskera y, por otro lado, es relativamente sencillo debido al ámbito restringido de aplicación del experimento.

4 Evaluación

Para comprobar el funcionamiento de cada una de las gramáticas se ha tomado un conjunto de 50 sintagmas nominales extraídos de un corpus periodístico. Este conjunto de prueba comprende 50 diferentes tipos de sintagmas nominales, es decir, no hay tipos de sintagmas repetidos. Dos sintagmas se han considerado como de tipos diferentes en caso de que tengan diferentes elementos o diferentes rasgos sintácticos (por ejemplo, nombre+caso, det+adj+nombre+caso, …). Para cada uno de ellos se han determinado la(s) interpretación(es) correcta(s). Estas interpretaciones se han comparado con el resultado de cada uno de los analizadores. Los principales problemas de una gramática computacional son la

sobregeneración (obtención de más análisis de los correctos) y la infrageneración (no obtención de análisis que son correctos). Una gramática poco restrictiva tendrá una cobertura cercana al 100%, pero esto no será útil si produce multitud de análisis redundantes. Por el contrario, si la gramática es demasiado restrictiva, habrá elementos válidos que no recibirán ningún análisis.

La tabla 5 muestra los resultados de aplicar las diferentes gramáticas al conjunto de prueba. Como primera medida se ha evaluado el número de ejemplos para los que la gramática produce un análisis correcto (aunque en la mayoría de los 50 ejemplos el sintagma tiene un solo análisis correcto, también hay casos en los que hay más de un análisis correcto, debido a la ambigüedad sintáctica). Este valor aparece en la columna de en medio de la tabla, y da una medida de la cobertura de la gramática.

Gramática		Obtención de al menos un análisis correcto		Diferencia respecto al número de análisis correctos
GM		39	78%	1,18
C1	a	25	50%	0,72
	b	27	54%	0,81
	c	29	58%	1,24
	d	32	64%	0,66
C2	a	32	64%	1,44
	b	33	66%	1,03
	c	33	66%	2,91
	d	34	68%	0,74
C3	a	35	70%	1,71
	b	36	72%	1,03
	c	36	72%	3,00
	d	37	74%	0,68

Tabla 5. Evaluación de los analizadores sobre 50 ejemplos de sintagmas nominales.

Se puede comprobar cómo la gramática desarrollada manualmente (GM) obtiene los mejores resultados en cuanto a cobertura (78%). Como se puede suponer, en las gramáticas automáticas el aumento del tamaño del corpus de extracción permite obtener mayor cobertura, acercándose a GM en el caso C3 (corpus de mayor tamaño), llegando al 74%. Por otro lado, para medir el grado de sobregeneración e infrageneración, se ha contado también la diferencia en número de análisis correctos entre el resultado correcto y las gramáticas desarrolladas. Para dar una medida simplificada, se ha calculado la media de la diferencia en valor absoluto entre los resultados esperados y los obtenidos (es decir, hemos dado el mismo "peso" a tener un análisis extra que uno menos). Se comprueba cómo las gramáticas del grupo c (con información sobre número y complementos de genitivo) tienden a sobregenerar. Por ejemplo, la gramática C3c produce de media una diferencia de 3,00 análisis respecto al número real de interpretaciones. En la gran mayoría de los casos examinados esto es debido a la sobregeneración, que en este caso no se produce por la generación de análisis extra incorrectos, sino por la aparición de múltiples análisis correctos que son redundantes. Es por este motivo que no hemos calculado la precisión como tal, ya que al ser los análisis obtenidos correctos en su gran mayoría, la precisión supera el 95%, siendo el mayor problema la obtención de resultados repetidos.

Un resultado que puede parecer sorprendente es que las gramáticas de clase d (sin información sobre número y generalización de complementos de caso genitivo), que en principio eran candidatas a generar en exceso, debido a que no ponen restricciones de concordancia en número, son las que menos sobregeneración producen, superando incluso a la gramática manual, con 0,68 análisis extra en C3d, frente a 1,18 en la gramática manual. La principal razón podría ser que este conjunto de gramáticas reduce en gran medida el número de reglas (al generalizar, se mantiene un solo ejemplar de las reglas repetidas).

5 Conclusiones y trabajo futuro

Este trabajo ha mostrado el proceso de construcción de una gramática sintáctica de los sintagmas nominales del euskera en el sistema RASP. Por un lado, se ha desarrollado la gramática de manera manual por un experto lingüista, mientras que por otra parte se ha generado un conjunto de gramáticas de manera casi automática, usando una gramática en el formalismo PATR y un corpus. En este segundo caso, se ha obtenido un conjunto de gramáticas de manera automática con un esfuerzo pequeño, a partir de:

- una gramática PATR previa,
- un mínimo conocimiento lingüístico respecto del tipo de reglas a extraer (especificando el conjunto de rasgos

sintácticos relevantes y el uso de un único tipo ó dos tipos de reglas),

- un corpus del que se han extraído patrones de los sintagmas analizados con la gramática antigua.

Se ha comprobado la cobertura de las diferentes gramáticas, midiendo tanto la sobregeneración como la infrageneración respecto a un conjunto de sintagmas nominales de prueba. Uno de los resultados principales es que en varios casos las gramáticas generadas de manera automática se aproximan a los resultados de la gramática manual. Las principales conclusiones del trabajo son:

- Aún cuando en nuestra opinión es preferible una gramática manual, hemos comprobado que una gramática automática puede ser una alternativa válida en casos en que no se disponga de recursos humanos suficientes. Examinando los errores cometidos por la gramática automática, se ha comprobado que debido a un error en el proceso de obtención de patrones, las gramáticas automáticas han fallado en 4 casos de prueba, lo que daría mayor cobertura a la gramática automática (82% frente 78% de la manual). En cualquier caso, nuestro principal objetivo es una gramática de amplia cobertura, y la gramática manual tiene ventajas en cuanto a menor número de reglas, modularidad y modificabilidad, características importantes de cara a un futuro mantenimiento.
- Las gramáticas obtenidas de manera automática sirven como punto de comparación con la gramática manual, proporcionando información sobre errores o casos no contemplados por el lingüista. El hecho de basarla en corpus hace que el proceso de obtención de la gramática final sea más sistemático.

Aunque los patrones de sintagmas nominales obtenidos automáticamente se han aplicado directamente al RASP, esto tiene el efecto de que los analizadores sintácticos basados en gramáticas independientes del contexto pierden eficiencia al aumentar el tamaño de la gramática. Existen otras posibilidades no exploradas, como la inducción de gramáticas a partir de los patrones obtenidos, o la conversión a un autómata o transductor de estado finito, mecanismos más compactos y eficientes que la solución adoptada en este trabajo.

Bibliografía

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. *Proceedings of RANLP'1997*.

Aldezabal I., M. Aranzabe, A. Atutxa, K. Gojenola, M. Oronoz, y K. Sarasola. 2003. Application of finite-state transducers to the acquisition of verb subcategorization information. *Natural Language Engineering*, Vol. 9 - 01. Cambridge University Press.

Aldezabal I., M. Aranzabe, A. Atutxa, K. Gojenola, K. Sarasola, y P. Goenaga. 2001. Extracción masiva de información sobre subcategorización verbal vasca a partir de corpus. *Actas del XVII Congreso de la SEPLN* Universidad de Jaen, 2001.

Briscoe, E., y J. Carroll. 1993. Generalised probabilistic LR parsing of natural language (corpora) with unification-based grammars. *Computational Linguistics,* 19(1). 25-59.

Briscoe, E., y J. Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation,* Las Palmas, Gran Canaria.

Charniak E. 1996. Tree-bank Grammars. *National Conference on Artificial Intelligence AAAI/IAAI*, Vol. 2.

Cowan B., y M. Collins. 2005. Morphology and Reranking for the Statistical Parsing of Spanish. *Proceedings of EMNLP'2005*.

Gazdar G., E. Klein, G. Pullum, y I. Sag. 1985. *Generalized Phrase Structure Grammar*. Harvard University Press.

Goenaga P. 1980. *Gramatika bideetan*. Erein.

Krotov A., M. Hepple, R. Gaizauskas, y Y. Wilks. 1998. Compacting the Penn Treebank. *Proceedings of the ACL'COLING Joint Conference, Montreal.*

Shieber S.M. 1986. *An Introduction to Unification-Based Approaches to Grammar*. CSLI Lecture Notes, 4, Stanford.

Zubiri E., y I. Zubiri. 2000. *Euskal Gramatika osoa.* (2º ed.). DIDAKTIKER.