APLICACION DE LA RELAJACION GRADUAL DE RESTRICCIONES PARA LA DETECCION Y CORRECCION DE ERRORES SINTACTICOS

Koldo Gojenola, Kepa Sarasola
Informatika Fakultatea 649 Postakutxa
20080 Donostia
Tfno: 943-218000
jipgogak@si.ehu.es

En este artículo se presenta un sistema realizado para la detección y corrección de errores gramaticales. El sistema se ha desarrollado para tratar un conjunto de errores representativos de las anomalías sintácticas encontradas en textos escritos en euskara. En los numerosos trabajos realizados hasta el momento sobre corrección de textos basada en la sintaxis se ha utilizado la técnica de relajación de restricciones sintácticas, de forma que se permite el análisis de oraciones en las que no se cumplen todas las restricciones gramaticales exigidas por el lenguaje. La técnica se ha utilizado para lenguas con una concordancia relativamente pobre entre constituyentes como el inglés, francés o español, pero tiene problemas de eficiencia en el caso de los errores tratados para el euskara. En este trabajo se generaliza la técnica de relajación para su aplicación a un mayor número de errores y a lenguas con una concordancia rica entre constituyentes. Se proponen criterios para el diagnóstico y corrección de esos errores mediante la relajación gradual de restricciones gramaticales. La implementación realizada según estos criterios mejora la eficiencia del corrector sintáctico.

Palabras clave: Análisis sintáctico, corrección gramatical, sistemas robustos.

APLICACION DE LA RELAJACION GRADUAL DE RESTRICCIONES PARA LA DETECCION Y CORRECCION DE ERRORES SINTACTICOS

Koldo Gojenola, Kepa Sarasola
Informatika Fakultatea 649 Postakutxa
20080 Donostia
Tfno: 943-218000
jipgogak@si.ehu.es

1. Introducción

Una de las primeras aplicaciones del procesamiento del lenguaje natural ha sido el tratamiento de lenguaje "no gramatical", utilizándose el término de sistema robusto para denominar a sistemas que no se colapsan al enfrentarse a lenguaje no permitido por una gramática dada.

En este artículo se presenta un sistema realizado para la detección y corrección de errores gramaticales. El sistema se ha desarrollado para tratar un conjunto de errores representativos de las anomalías sintácticas encontradas en textos escritos en euskara. En los numerosos trabajos realizados hasta el momento sobre corrección de textos basada en la sintaxis se ha utilizado la técnica de relajación de restricciones sintácticas, de forma que se permite el análisis de oraciones en las que no se cumplen todas las restricciones gramaticales exigidas por el lenguaje. La técnica se ha utilizado para lenguas con una concordancia relativamente pobre entre constituyentes como el inglés, francés o español, pero tiene problemas de eficiencia en el caso de los errores tratados para el euskara. En este trabajo se generaliza la técnica de relajación para su aplicación a un mayor número de errores y a lenguas con una concordancia rica entre constituyentes. Se proponen criterios para el diagnóstico y corrección de esos errores mediante la relajación gradual de restricciones gramaticales. La implementación realizada según estos criterios mejora la eficiencia del corrector sintáctico.

El artículo se estructura de la siguiente manera: en el segundo punto se describe el método de relajación de restricciones sintácticas generalmente empleado en los corretores gramaticales. Se destacan los problemas de eficiencia que presenta este método y se plantea utilizar un método de relajación a varios niveles. En el tercer punto se presentan los tipos de errores sintácticos que se han recopilado y clasificado para el euskara. El apartado cuarto describe la arquitectura del sistema diseñado. Finalmente se evalúan los resultados obtenidos.

2. Correctores gramaticales basados en relajación de restricciones

Desde el primer momento en que se empezó a realizar el análisis sintáctico de forma automática se vio la necesidad de tratar lenguaje fuera de la cobertura estricta de una gramática dada. Esta necesidad se hizo evidente en las primeras aplicaciones diseñadas en análisis sintáctico, los compiladores de lenguajes de programación, en los que el tratamiento de errores es una componente sustancial y de la que depende en gran medida el éxito de un compilador independientemente de su eficiencia. Generalmente, un compilador necesita de varias etapas para detectar un error y, opcionalmente, generar correcciones posibles.

En las aplicaciones de procesamiento del lenguaje natural (PLN) también se ha afrontado el problema del tratamiento de lenguaje extragramatical:

a) interfaces en lenguaje natural ([Young et al., 91])
b) comprensión de textos, recuperación de información ([Carbonell y Hayes, 83], [Karlsson et al., 92])
c) tratamiento automático de lenguaje hablado ([Tomabechi, 93])
d) correctores ortográficos y gramaticales ([Heidorn et al., 82], [Rodríguez, 91], [Rabinovitz, 93])
e) estudio del aprendizaje de segundas lenguas ([Maritxalar, 93])

La no gramaticalidad de una cadena con respecto a un sistema de PLN se puede definir en dos ámbitos diferentes:

- Gramaticalidad absoluta con respecto a una lengua como por ejemplo el inglés. Este es un término difícil de manejar computacionalmente al no existir actualmente un analizador de un lenguaje que asigne una estructura a todas las cadenas de una lengua y solamente a esas cadenas.
- Gramaticalidad relativa con respecto a una gramática particular de una lengua. Una oración no es gramatical si no es cubierta por la gramática utilizada en un sistema, que generalmente cubre sólo un subconjunto de la lengua.

Esta distinción no suele ser importante en los sistemas de tipos a), b) y c), donde no es tan importante determinar la gramaticalidad de una oración como asignar una estructura (gramatical o no) a esa oración, de forma que el resultado pueda ser procesado posteriormente. En los sistemas de tipos d) y e), sin embargo, es importante determinar la gramaticalidad de forma absoluta.

Los errores gramaticales se han clasificado de acuerdo a los tipos de conocimiento lingüístico ([Lapalme, 85]):

- Errores de tipo ortográfico. En este tipo se incluyen los errores cometidos al escribir una palabra. Estos errores pueden ser debidos al desconocimiento de la ortografía de las palabras, de las reglas morfológicas del lenguaje, o a errores tipográficos. Se producen normalmente en palabras aisladas, y son tratados por correctores ortográficos, de uso común en los procesadores de texto actuales ([Peterson, 80], [Agirre et al., 92]).
- Errores sintácticos. Entre los muchos tipos de errores debidos al uso incorrecto de reglas sintácticas, se pueden destacar los errores de concordancia sintáctica en género, número persona y caso entre los componentes de una oración.
- Errores de tipo semántico y pragmático. Se producen cuando el significado de la oración no corresponde con el pensado por el autor, o se tiene una oración correcta sintácticamente pero sin sentido.
- Errores de puntuación. En varias clasificaciones de errores gramaticales, estos se agrupan con los errores sintácticos, aunque en muchos casos el uso de la puntuación también tiene una estrecha relación con la semántica de la oración ([Nunberg, 91]).

La importancia del uso de conocimiento sintáctico en el tratamiento de errores es clara. La proporción de errores sintácticos puede variar desde un 36 a un 60% del total de errores para textos escritos en inglés ([Kukich, 92]), o llegar hasta un 45% según un estudio realizado para determinar la ocurrencia de diferentes tipos de errores en diarios franceses ([Lapalme, 85]). Por otro lado, un gran porcentaje de los errores de tipo ortográfico (en palabras aisladas) produce palabras correctas, no pudiendo ser detectados sin la ayuda de conocimiento sintáctico o semántico ([Mitton, 87]). Por último, el conocimiento sintáctico también es útil para discriminar las propuestas de corrección generadas por un corrector ortográfico ([Agirre et al., 94]).

En este trabajo se analiza la utilización de conocimiento sintáctico para tratar errores en textos escritos en euskara. No existe una estimación de la proporción de errores sintácticos para el euskara, aunque el hecho de ser ésta una lengua recientemente normalizada permite suponer que tanto el número como la variedad de errores sean importantes.

2.1 Método de relajación de restricciones

Un método comúnmente utilizado para detectar errores sintácticos consiste en utilizar una gramática independiente del contexto como esqueleto del análisis, aumentada con restricciones de tipo sintáctico que definen los requisitos que deben

cumplir los componentes de las reglas. Este método es especialmente adecuado para ser utilizado con formalismos basados en unificación (PATR-II, LFG, HPSG, ...) en los que la mayor parte del conocimiento lingüístico se encuentra expresado en estructuras del tipo rasgo-valor, y en los que las restricciones que cumplen las estructuras se definen por medio de ecuaciones.

La mayor parte de los errores sintácticos se puede tratar como la violación de una restricción. Por ejemplo, dada la regla sintáctica

```
Oración ->     SN(Num)   SV(Num)
```

con la restricción de concordancia en número entre el sintagma nominal sujeto y el sintagma verbal, y la oración no gramatical "*Juan y Pedro come", se puede detectar el error si se elimina la restricción. Para generar una propuesta de corrección se puede elegir según diferentes criterios cuál de los dos elementos tiene el rasgo correcto, y corregirlo posteriormente con un generador morfológico. La gran ventaja del método de relajación de restricciones es que permite utilizar la misma gramática para el análisis de oraciones correctas y oraciones no gramaticales, relacionando los análisis de estas últimas con oraciones correctas.

[Vosse, 92, 93] propone asignar un peso a cada una de las restricciones expresadas en reglas gramaticales, de forma que ese peso exprese la penalización que el escritor de la gramática asigna a su violación. Durante el análisis, al comprobar las restricciones sintácticas, se guardan los análisis resultantes de aplicar o no esas restricciones, con lo que al final se tendrá, generalmente, un gran número de análisis, de los cuales se elegirán los de peso mínimo (si la oración era gramatical, habrá al menos un análisis de peso cero). También [Tomabechi, 93] utiliza esta aproximación, guardando análisis de componentes que violan ciertas restricciones para evitar repetición de tratamientos en caso de oraciones incorrectas. El gran problema de esta solución es la adición de multitud de análisis erróneos, calculados incluso cuando la oración es gramatical, que se añaden a la intrínseca ambigüedad del lenguaje natural.

Sin embargo, el tratamiento uniforme para todas y cada una las restricciones resulta muy ineficiente. El número de combinaciones de posibles relajaciones aumenta prohibitivamente con el número de restricciones. Piénsese por ejemplo cuál sería el número de combinaciones para una regla con diez restricciones. Este efecto se multiplica en el caso de la aplicación conjunta de varias reglas (por ejemplo, al verificar concordancias entre elementos de la oración principal y subordinadas). Esto no ha sido un grave problema en el caso del inglés, lengua a la que se han dedicado gran parte de los trabajos realizados en corrección gramatical, debido al relativo orden fijo de los constituyentes principales de una oración y a la pobre concordancia entre

ellos. En el caso de otras lenguas como el español y francés, con un orden más libre y mayor concordancia, no se han citado problemas importantes ([Genthial, 92], [Rodríguez, 91]). Sin embargo, en lenguas con orden libre de componentes al nivel de la oración principal y concordancia rica entre ellos, como el euskara, ruso o latín ([Miller, 87]), el tratamiento de errores gramaticales difiere en gran medida del tratamiento para oraciones correctas, ya que la relajación de las múltiples restricciones sintácticas genera un número alto de posibilidades a examinar. Estos fenómenos exigen la adecuación de la técnica de relajación de restricciones para permitir su uso eficiente.

2.2 Relajación gradual de restricciones

En este trabajo se ha estudiado otra alternativa consistente en ir relajando las restricciones de forma gradual, de manera que en cada nivel se relajen diferentes subconjuntos de restricciones, lográndose así reducir el número de combinaciones de relajación. En una primera opción se realizará el análisis con la gramática completa. En caso de no obtener ningún análisis se relajará un primer conjunto de restricciones, si aún así no se obtienen análisis a continuación se relaja un segundo conjunto de restricciones, y así sucesivamente. Los diferentes niveles tratarán diferentes tipos de errores, típicamente de menor a mayor gravedad o coste computacional. Esta técnica ha sido propuesta por [Douglas y Dale, 92] para el tratamiento de errores sintácticos, y por [Chanod et al., 93] como modelo de un analizador sintáctico robusto, aunque en ninguno de los trabajos citados se han descrito fenómenos que precisen de más de un nivel de relajación (aparte del nivel básico de la gramática completa), debido a la simplicidad de los errores tratados en ambos casos. Este tratamiento permite ampliar la cobertura de la gramática, desde el nivel gramatical mínimo G_0, que define la gramática de las oraciones totalmente correctas, hasta el nivel máximo posible de relajación G_n (Fig. 2), que comprendería el esqueleto de la gramática independiente del contexto subyacente (esta relación de inclusión entre cobertura de gramáticas correspondientes a distintos niveles de relajación sólo se dará en caso de que cada nivel comprenda las restricciones del nivel anterior).

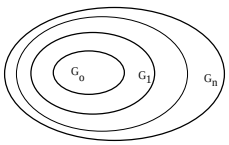


Figura 2.

3. Caracterización de errores sintácticos para el euskara

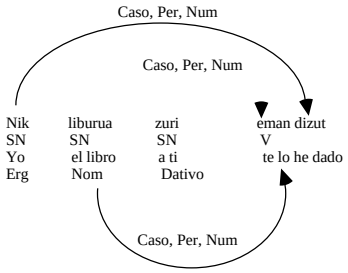
Los tipos de errores más comúnmente tratados en los sistemas de corrección gramatical ([Miller, 86], [Rodríguez, 91]) son los siguientes:

- Errores de concordancia. Un ejemplo típico son las restricciones de concordancia en género, número, caso y persona entre los componentes de una oración.
- Confusión léxica. Se confunde una palabra con otra diferente. Aquí se incluyen los casos de sustitución de una palabra por otra similar u homófona, y también los errores de utilización de un paradigma de declinación incorrecto para una palabra.
- Omisión o duplicación de elementos. El elemento erróneo puede ser una palabra, morfema o componente complejo.

En el caso del euskara, se han escrito varios estudios ([HABE, 85]) sobre diferentes tipos de errores, la mayoría realizados en el contexto del aprendizaje de segundas lenguas. Previamente a la realización del presente trabajo también se examinó un conjunto de 100 redacciones escritas por estudiantes de euskara de un nivel medio-alto.

Dentro de la gran variedad de errores encontrados, destaca el papel de los errores de concordancia, que comprenden un porcentaje muy alto del total. Se pueden establecer distintas subclases de errores de concordancia:

- Concordancia entre elementos de la oración y el verbo.



"Yo te he dado el libro"

Figura 1.

Los sintagmas nominales nucleares (sujeto, objeto y objeto indirecto) concuerdan en número, persona y caso con el verbo (Fig. 1). El orden de los constituyentes es libre y se permite la elipsis de cualquiera de los sintagmas nominales. Esto lleva a que el examen de sentencias correctas (donde una diferente marca de caso distingue a cada sintagma nominal) sea mucho más sencillo que el de sentencias incorrectas, donde la posible incorrección o ausencia de marcas de caso, persona y número lleva a tener que comprobar numerosas hipótesis de error.

- Errores en el ámbito del sintagma nominal, como la adición, supresión o confusión de determinantes y numerales.

- Complementos oracionales, en los que se suprime o se confunde el sufijo o marca correspondiente. Estos errores se pueden ver también como un caso de concordancia (los complementos oracionales carecen de o tienen una marca diferente de la esperada). Por ejemplo:

```
Nik eman dizudan liburua polita da
(el libro que te he dado es bonito)

*Nik eman dizut     liburua     polita da
te he dado          el libro    es bonito
(omisión de la marca de oración de relativo)
```

- Omisión o confusión del régimen de subcategorización del verbo:


```
Jon mailuaz baliatu zen   (Jon se valió del martillo)
*Jon mailuarekin baliatu zen  (Jon se valió con el martillo)
```

4. Arquitectura del sistema

A continuación se presenta la arquitectura del sistema que se ha utilizado para estudiar el uso de la relajación gradual de restricciones. En la figura 3 se presentan los módulos en que se divide el sistema.

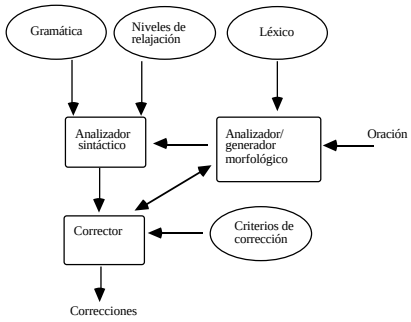


Figura 3.

- Gramatica y léxico. El sistema será tanto más robusto cuanta más información se pueda especificar por medio de restricciones, en detrimento de la componente independiente del contexto. Se ha elegido un formalismo sintáctico no condicionado a ninguna teoría sintáctica, afín a PATR-II ([Shieber, 86]). La elección de un formalismo declarativo permitirá la generalidad y aplicabilidad del método a otros sistemas. Por ejemplo, la regla que establece la correspondencia entre el sujeto y el sintagma verbal de una oración se expresa de la siguiente forma[1] (las restricciones se han numerado para poder referenciarlas en el caso de que se relajen):

[1]En la regla, *as* e *is* designan el sintagma nominal sujeto y el sintagma verbal, respectivamente.


```
erregela 1 X0 ---> X1, X2
  (1, X0/cat <=> as),
  (2, X1/cat <=> is),
  (3, X2/cat <=> as),
  (4, X1/agr/case <=> X2/subj/agr/case),
  (5, X1/agr/num <=> X2/subj/agr/num),
  (6, X1/agr/per <=> X2/subj/agr/per),
  (7, X0/subj/agr/per <=> X2/subj/agr/per),
  (8, X0/subj/agr/num <=> X2/subj/agr/num),
  (9, X0/subj/agr/case <=> X2/subj/agr/case),
  (10, X0/obj <=> X2/obj),
  (11, X0/obj2 <=> X2/obj2),
  (12, X0/daughters/first <=> X1),
  (13, X0/daughters/rest/first <=> X2),
  (14, X0/order <=> X2/order),
  (15, X0/subj/def <=> X1/def),
  (16, X2/mode <=> normal),
  (17, X2/order <=> left),
  (18, X0/subj/lex <=> X1/lex)
```

Un inconveniente del método es que si se quiere reutilizar una gramática previamente definida (como la LFG definida por [Abaitua, 88] para el euskara) es preciso poner las restricciones de forma explícita, perdiéndose en parte las generalizaciones lingüísticas que hacen compactas y elegantes las modernas gramáticas de unificación .

La gramática diseñada para el estudio trata oraciones enunciativas permitiéndose el uso de oraciones subordinadas de relativo. Las reglas definidas son suficientes para tratar los tipos de errores descritos en el apartado anterior.

- Información sobre las restricciones a relajar en cada una de las reglas por cada uno de los niveles de relajación. Se debe determinar qué restriciones se van a relajar (vendrán determinadas por los errores a tratar) y a qué nivel (esto influirá en la eficiencia: si sólo hay un nivel, se tendrán numerosos análisis, muchos de ellos poco plausibles). Esta información tiene relación con la distinción realizada por [Uszkoreit, 91] entre competencia gramatical y ejecución ("performance"), donde se propone que las restricciones declarativas de las gramáticas de unificación se vayan examinando teniendo en cuenta su capacidad de discriminación, de cara a lograr mayor eficiencia. Por ejemplo, para la regla anterior se tiene:

```
Nivel 0:
   rest. obligatorias  [2,3,1,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18],
   rest. a relajar     []
Nivel 1:
   rest. obligatorias  [2,3,1,10,11,12,13, 14, 15, 16, 17, 18],
   rest. a relajar:
            [4,9],   "no concuerdan el caso del sujeto y el del verbo"
            [5,8],   "no concuerdan el número del sujeto y el del verbo"
            [6,7],   "no concuerdan la persona del sujeto y la del verbo"
Nivel 2:    ...
Nivel 3:    ...
```

- Analizador sintáctico (parser). Se trata de un analizador ascendente simple que, de acuerdo al nivel de relajación actual, testea o no las restricciones

gramaticales. Aunque una restricción se pueda relajar a un nivel dado, el analizador sólo lo hace en caso de que no se cumpla, de forma que no se incremente el número de combinaciones de relajación en caso de que esa restricción sea aplicable.

- Analizador/generador morfológico. Aunque para esta prueba se ha utilizado un diccionario reducido, con un análisis ad hoc, se piensa integrar el analizador morfológico XUXEN, de gran cobertura, desarrollado para el euskara ([Agirre et al., 92]).
- Corrector y criterios de correción. Dados varios árboles de análisis para una oración, en los que se han relajado una o más restricciones, el corrector se valdrá de estos criterios para elegir la(s) corrección(es) más adecuadas. Primeramente se deben seleccionar las alternativas a proponer como corrección. El primer criterio de discriminación utilizado en este trabajo ha sido tratar los análisis con el número mínimo de errores (restricciones relajadas). En los errores estudiados para el euskara la oración correcta se encuentra, en la gran mayoría de los casos, entre las correcciones de menor número de errores (generalmente uno o dos). En los demás casos es prácticamente imposible, incluso para un corrector humano, decidir cuál es la corrección a la oración dada. En caso de haber varios análisis con el número mínimo de errores se deberán determinar criterios de desambiguación, o bien se pueden presentar todos como correcciones plausibles. Por ejemplo, en caso de que haya una corrección que exija un cambio de la persona de un constituyente y otra para el número, es más plausible corregir el número. Otra opción a considerar en el futuro será la asignación de pesos a cada una de las restricciones.

5. Evaluación de resultados. Ejemplo de aplicación

A partir de la clasificación de errores confeccionada a partir de bibliografía y de un conjunto de 100 redacciones escritas por estudiantes de euskara de un nivel medio-alto se han definido los siguientes niveles abstractos de relajación de restricciones:

- Primer nivel. Se examinan todas las restricciones.
- Segundo nivel. Se relajan las restricciones de concordancia en caso, persona y número de los sintagmas nominales con relación al verbo.
- Tercer nivel. Relajación de las restricciones sobre los tipos de complementos oracionales. En este caso se incluye el error relativamente común de confundir u omitir el sufijo de los complementos oracionales.

- Cuarto nivel. Combinación de las relajaciones de segundo y tercer nivel. La eficiencia del analizador disminuye considerablemente en los casos en que ha de contemplar este nivel, debido al gran número de análisis obtenidos. Su definición se justifica porque los errores de concordancia del segundo nivel se dan con mayor frecuencia en oraciones complejas.

Para probar el sistema se cogió un grupo de 53 oraciones incorrectas extraídas del corpus señalado. Estas oraciones son representativas de los diferentes tipos de errores a tratar. En el conjunto de prueba se tienen treinta oraciones simples con errores de concordancia, diez oraciones compuestas con errores de concordancia, ocho oraciones compuestas con errores en oraciones de relativo, y cinco oraciones con combinaciones de ambos tipos de errores. Los resultados se indican en la tabla 1, donde se muestra la diferencia entre mantener uno o varios niveles de relajación:

	Un solo nivel	Tres niveles
Oraciones correctas	0.3 / 4.5 / 2.0	0.4 / 6.0 / 2.63
Errores de concordancia	0.56 / 15.0 / 4.1	0.6 / 16.0 / 4.9
Errores en or. de relativo	53.0	10.0
Combinación de errores	94.0	101.0

Tabla 1.-Tiempo de análisis medio en segundos según tipo de errores y número de niveles de relajación[2].

En el caso de oraciones correctas los tiempos de análisis son ligeramente superiores en el caso de utilizar varios niveles de relajación. Para los errores de concordancia el tiempo es parecido en los dos casos, en contra de lo esperado, ya que se pensaba que el utilizar un nivel específico para estos errores mejoraría los resultados respecto a la utilización de un solo nivel, en el que la relajación de restricciones correspondientes a otros errores podía aumentar el tiempo de análisis. La explicación de que no suceda puede ser que las restricciones de oraciones de relativo no interaccionan con las restricciones de concordancia. Por otro lado, en el caso de errores en oraciones de relativo el tiempo medio de análisis es cinco veces menor al utilizar varios niveles de

[2]En las casillas con tres valores:
tiempo medio en or. simples / t. medio en or. complejas / t. medio en total

relajación, debido a la interferencia de las restricciones de concordancia, que aumentan el número de posibilidades a considerar. En el caso de tener combinaciones de errores el análisis por niveles es menos eficiente que el de un solo nivel, al irse probando los distintos niveles hasta llegar al último. En cualquier caso, esta ligera ineficiencia se ve compensada por el hecho de que los errores múltiples son menos frecuentes, no siendo necesario descender hasta ese nivel normalmente.

Por otra parte, sería útil establecer, para una gramática dada y unos tipos de errores a tratar, un afinado óptimo para la relación entre el número de niveles y la eficiencia. Un mayor número de niveles produce mejores resultados para oraciones con errores de niveles bajos, mientras que repite tratamientos para errores graves. El número de niveles y la distribución óptima de restricciones habrán de obtenerse tras un estudio minucioso de los errores producidos. Por ejemplo, la supresión de las restricciones de persona, número y caso de sintagmas nominales y verbos en un solo nivel de relajación comporta un gran número de análisis posibles, de los que muchos no serán tenidos en cuenta posteriormente por su elevado número de errores. En este caso será más adecuada una nueva subdivisión, de manera que se relajen primero las restricciones que son causa de un mayor número de errores (por ejemplo, es un error frecuente omitir la marca de caso del ergativo).

Otra posible mejora consistiría en la adaptación dinámica del número de niveles necesarios según el tipo de errores dominante en la aplicación concreta (por ejemplo, en un entorno de enseñanza de segundas lenguas).

En la tabla 2 se muestra el número de propuestas (árboles de análisis) presentado por el analizador utilizando cuatro niveles de relajación:

An. correctos	Una corrección	2 correcciones	Más de 2 correcciones
11	14	17	11

Tabla 2.-Distribución de las oraciones analizadas en función del número de correcciones (análisis) propuestas.

Como primer resultado a destacar se tiene que once sentencias producen análisis correctos sintácticamente. Estos análisis serán incorrectos semántica o pragmáticamente. Este hecho fija un límite inferior en cerca de un 20% de errores de concordancia que no se pueden detectar únicamente con conocimiento sintáctico. La introducción de conocimiento semántico para discriminar los análisis obtenidos, como por ejemplo se propone en [Agirre et al., 94], será la única forma de detectar estos errores.

En cuanto al número de errores (relajaciones) a corregir en cada propuesta de corrección el sistema presenta uno o dos errores en la gran mayoría de los casos. En su implementación actual el sistema muestra todas las propuestas. Se planea discriminar las propuestas de acuerdo a los tipos de errores producidos.

6. Conclusiones

En este trabajo se ha aplicado el método de detección y correción de errores sintácticos propuesto por [Douglas y Dale, 92] y [Chanod et al., 93], basado en la ampliación gradual de la cobertura de una gramática por medio de la supresión de restricciones sintácticas. La relajación de restricciones se ha aplicado en varios niveles, generalizándose el tratamiento descrito hasta ahora para lenguajes con una concordancia más pobre que el euskara.

Se ha escogido un formalismo gramatical basado en la unificación, del tipo de PATR-II, como base del sistema para permitir, gracias a sus características (declaratividad, uniformidad, ...), la reutilización de cualquier gramática tras una adecuación más o menos laboriosa.

Se ha realizado una recopilación y clasificación de errores sintácticos en textos escritos en euskara, haciendo una propuesta de criterios para su diagnóstico y corrección mediante la relajación gradual de restricciones. El método de relajación es válido para un amplio grupo de errores sintácticos que se pueden describir como violaciones de restricciones sintácticas, y se ha probado en el tratamiento de un conjunto de oraciones encontradas en textos reales y descritas en clasificaciones de errores típicos del euskara, mostrando una mayor eficiencia.

La generación de correcciones para los errores examinados consistirá en la modificación de los rasgos de una palabra y la utilización del generador morfológico ya implementado.

Referencias

Abaitua, J. 1988 "Complex predicates in Basque: from lexical forms to functional structures" *Tesis doctoral*, Univ. de Manchester

Aduriz, I., Agirre, E., Alegría, I., Arregi, X., Arriola, J.M., Artola, X., Díaz de Ilarraza, A., Ezeiza, N., Maritxalar, M., Sarasola, K. and Urkia, M. 1991 "A Morphological Analysis Based Method for Spelling Correction" in *Proceedings of the E.A.C.L.*, Utrecht, The Netherlands

Agirre, E., Alegria, I., Arregi, X., Artola, X., Díaz de Ilarraza, A., Maritxalar, M., Sarasola, K. and Urkia, M. 1992 "XUXEN: A Spelling Checker/Corrector for Basque Based on Two-Level Morphology" in *Proceedings of the third conference on Applied Natural Language Processing*, Trento, Italy

Agirre, E., Arregi, X., Artola, X., Díaz, A., Sarasola, K. 1994 "Lexical-semantic information and the automatic correction of spelling errors" *in Proceedings of the workshop on Semantics and Pragmatics of NaturallLanguage: Logical and computational aspects*, Sara, France

Carbonell, J., Hayes P. J. 1983 "Recovery strategies for parsing extragrammatical language" *American Journal of Computational Linguistics*

Chanod, J. P., Montemagni S., Segond F. 1993 "Multiple-pass parsing and dynamic relaxation: a text-driven approach to parsing" *Proceedings of the International Conference on Expert Systems and Natural Language Processing*, Avignon, France

Douglas, S., Dale R. 1992 "Towards Robust PATR" *Proceedings of the International Conference on Computational Linguistics (COLING)*, Nantes

Genthial, D. 1991 "Contribution à la construction d'un système robuste d'analyse du français" *Thèse de l'université Joseph Fourier* , Grenoble I

HABE 1985 "Akatsen bilduma analitikoa" *Zutabe Aldizkaria*

Heidorn G. E., Jensen K., Miller L. A., Byrd R. J., Chodorow M. S. 1982 "The EPISTLE text-critiquing system" *IBM Systems Journal*, Vol. 21, No. 3

Karlsson, F., Voutilainen A., Heikkilä J., Anttila A. 1992 "Constraint Grammar: a language-independent system for parsing unrestricted text"

Kukich, K. 1992 "Techniques for Automatically Correcting words in Text" *ACM Computing Surveys*, Vol. 24, No. 4, Diciembre

Lapalme, G., Brunelle, E. 1985 "La détection automatique des fautes de syntaxe en français"

Maritxalar, M. 1993 "Integration of NLP techniques in the ICALL systems field: the treatment of incorrect knowledge" *Proceedings of the AI-ED 93, World Conference on Artificial Intelligence in Education*, Edinburgh

Miller, L. A. 1986 "Computers for composition: a stage model approach to helping" *Visible Language*, XX(2)

Mitton, R. 1987 "Spelling checkers, spelling correctors, and the misspellings of poor spellers" *Information Processing and Management*, Vol. 23, No. 5

Nunberg, G. 1991 "The linguistics of punctuation" *CSLI Lecture Notes*, Stanford

Peterson J. L. 1980 "Computer programs for detecting and correcting spelling errors" *CACM* Volume 23

Rabinovitz R. 1993 "Better writing through electricity" *PC Magazine*, Mayo

Rodríguez, C. 1991"CORRECTOR: un sistema de verificación sintáctica y estilística de textos". *VII congreso de la SEPLN*

Shieber, S. 1986 "An introduction to unification-based approaches to grammar" *CSLI Lecture Notes nº 4*, Stanford, CA

Tomabechi, H. 1993 "A soft graph unification method for robust parsing" *Proceedings of the Third International Workshop on Parsing Technologies (IWPT-93)*, Tilburg (The Netherlands) and Durbuy (Belgium)

Uszkoreit, H. 1991 "Strategies for adding control information to declarative grammars" *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*

Veronis A. 1988 "Morphosyntactic correction in natural language interfaces" *Proceedings of the 12th International Conferenc on Computational Linguistics*, Budapest

Vosse, T. 1992 "Detecting and Correcting Morpho-syntactic Errors in Real Texts" *Proceedings of the 3rd International Conference on Applied Natural Language Processing*, Trento, Italy

Vosse, T. 1993 "Detecting and Correcting Morpho-syntactic Errors in Real Texts" *Proceedings of the TWLT* , Enschede, The Netherlands

Young, C. W. Eastman C. M., Oakman.R. L. 1991 "An analysis of ill-formed input in natural language queries to document retrieval systems" *Information Processing and Management*, Vol. 27, No. 6