PrADo: Preparación Automatizada de Documentos

TIC 2000-1681

Informe técnico

Marzo de 2004

Lourdes Aguilar, Àlex Alsina, Anna-Belén Avilés, Toni Badia, Sergio Balari, Gemma Boleda, Stefan Bott, Jenny Brumme, Carme Colominas, Anna Espunya, Josep Fontana, Jordi Fontseca, Àngel Gil, Carmen Hernández, Laia Mayol, Louise McNally, Carme de la Mota, Martí Quixal, Yolanda Rodríguez, Oriol Valentín, Enric Vallduví, Teresa Vallverdú

Contenidos

Índice de tablas

Índice de figuras

1 Introducción

El proyecto Preparación Automatizada de Documentos, PrADo, se inició a partir del interés de los dos grupos investigadores, de la Universitat Pompeu Fabra y de la Universitat Autònoma de Barcelona, en el procesamiento de textos reales no restringidos y en el establecimiento de métodos de procesamiento estables y eficientes para las lenguas habladas en Cataluña, el catalán y el castellano. Ver en el apéndice C la lista de investigadores de cada grupo.

1.1 Objetivos del proyecto

Los objetivos iniciales del proyecto eran:

- el desarrollo de dos prototipos de correctores gramaticales (uno para el catalán y otro para el español), a partir de:
    - las gramáticas de marcaje de textos catalanes y castellanos en el formato de la *Constraint Grammar*
    - la gramática del corrector propiamente dicho
    - la cadena de procesamiento general en la que se insertan las gramáticas de marcaje y del corrector
- la elaboración de un modelo de usuario para el corrector, que tuviera en cuenta:
    - las interferencias entre catalán y castellano en los usuarios de ambas lenguas en Cataluña,
    - las interferencias del inglés en los usuarios cultos de las dos lenguas habladas en Cataluña
- la preparación de las bases teóricas (lingüísticas e informáticas) para que en una etapa posterior se pudiera proceder principalmente a la construcción de dos correctores de estilo (para las dos lenguas estudiadas en el proyecto)

1.2 Desarrollo del proyecto

Cuando se puso en marcha el proyecto, el grupo de la UPF tenía bastante adelantado el desarrollo de una gramática de marcaje morfosintáctico para textos catalanes, en el formalismo

de la *Constraint Grammar*. Por otra parte, la empresa Conexor oy, que ofrece el formalismo a los grupos de investigación interesados, disponía de una gramática de marcaje morfosintáctico para el español y la ponía a disposición de los grupos de investigación a un precio razonable. Por ello, el presente proyecto tomaba como punto de partida las dos gramáticas de marcaje morfosintáctico, y se proponía en los primeros meses terminar la del catalán y adaptar la del español a las necesidades y objetivos del proyecto.

No obstante, como se indicó en la memoria del primer año de proyecto, la evaluación que llevamos a cabo sobre los resultados de la gramática de marcaje morfosintáctico para el español de Conexor no proporcionaba los resultados mínimos que eran de esperar y la empresa no estaba dispuesta a facilitarnos los medios para mejorarla. Consecuentemente, tomamos la decisión de implementar una nueva gramática de marcaje del español, partiendo de nuestra experiencia y conocimientos sobre la *Constraint Grammar* y sus características. Por lo tanto, el proyecto se vio reorientado en el sentido de tener que construir una gramática de marcaje nueva, con el tiempo de desarrollo que ello comporta.

Simultáneamente, iniciamos la investigación sobre las interferencias entre el catalán y el castellano en los hablantes de ambas lenguas residentes en Cataluña, y también sobre las interferencias del inglés sobre el catalán y el castellano de estos mismos hablantes. A finales del primer año contábamos pues con los resultados iniciales de esta investigación y pudimos iniciar el estudio de los modos de formular las reglas del corrector.

En el segundo año del proyecto se terminó la gramática de marcaje morfosintáctico para el catalán y se inició la evaluación de la misma. A partir de estos resultados se pudo empezar a trabajar en la versión definitiva de la gramática de marcaje sintáctico del catalán (cuyo desarrollo dependía, como es natural, de los resultados del marcaje morfosintáctico). Por otra parte, una vez establecidos los modos y el entorno de trabajo para el desarrollo de la gramática del español (proyección morfológica, corpus de desarrollo y programa de preprocesamiento), se empezó el desarrollo de la gramática de marcaje morfosintáctico del castellano. Para ello, se estableció una estrategia para aprovechar, siempre que fuera posible, la gramática catalana para poder llevar a cabo el desarrollo de la castellana en relativamente poco tiempo.

En este mismo año, en relación con el prototipo de corrector, se determinaron las distintas formulaciones posibles de las reglas del mismo, atendiendo tanto a sus características formales como a las descriptivas lingüísticas.

Finalmente, el tercer año del proyecto se dedicó a:

- terminar la gramática de marcaje sintáctico del catalán y a efectuar su evaluación
- terminar la gramática de marcaje morfosintáctico del español y a efectuar su evaluación
- elaborar los módulos del corrector del catalán que iban a formar parte del prototipo final del proyecto
- preparar la memoria final del proyecto y poner en orden los resultados del mismo

1.3 Metodología general

En este apartado reseñamos los aspectos más relevantes de la metodología seguida. En este contexto, deben tenerse en cuenta los factores condicionantes del proyecto:

- se trata un proyecto coordinado en el que participan dos grupos de investigación distintos, con muchos puntos de contacto, pero con características distintas en cuanto a tamaño, historia reciente, capacidad técnica, etc.,
- se trata de un proyecto de implementación de gramáticas lingüísticas, que no puede ser realizado por una sola persona en un breve plazo de tiempo, pero que comporta muchas dificultades de coordinación cuando se desarrolla en equipo,
- el proyecto supone integrar dos tipos de gramáticas claramente distintos, la gramática de marcaje (morfosintáctico y sintáctico) y la gramática del corrector, de manera que las estrategias de formulación del conocimiento lingüístico son diferentes, y no pueden mezclarse.

Una de las primeras preocupaciones del equipo de investigación ha sido garantizar la cobertura de las gramáticas de marcaje. Estas gramáticas deben ser suficientemente amplias y robustas como para marcar adecuadamente los textos a corregir, tanto si contienen errores como si no. Es básico que los desarrolladores del corrector gramatical conozcan plenamente el comportamiento de la gramática ante los textos que se aplicarán al corrector.

El equipo de trabajo del proyecto ha sido en algunos momentos bastante numeroso, entre los investigadores miembros de los grupos de investigación y los investigadores en formación que han colaborado en el mismo. Como estas gramáticas no están escritas en un formalismo de alto nivel (sino en máquinas de estados finitos), el proceso de escribir las mismas requiere un altísimo control por parte de los gramáticos. Por ello la metodología seguida intentaba garantizar un desarrollo seguro que impidiera la presencia de desviaciones en el proceso. Podemos resumir los aspectos esenciales de la siguiente manera:

- en todo el proceso de creación de las gramáticas de marcaje, el equipo de trabajo ha mantenido con regularidad sesiones de discusión sobre el tratamiento lingüístico incorporado en las gramáticas de marcaje,
- la implementación se ha llevado a cabo siempre en equipo,
- en todo momento, se ha llevado a cabo un control teórico sobre la implementación, de manera en cualquier momento se podía replantear cualquier decisión tomada (a la vista de las consecuencias que tiene en otras áreas de la gramática),
- el resultado ha sido evaluado por parte de expertos lingüistas, con unos criterios y una metodología preestablecidos,
- finalmente, las gramáticas han sido modificadas según los resultados de las evaluaciones

Otro aspecto metodológico importante ha sido el establecimiento inicial, y posterior mejora, del flujo general del procesamiento, que toma un texto en formato ASCII y lo retorna convenientemente marcado y analizado. Conviene notar que este flujo de procesamiento es utilizado en otros proyectos (notablemente, ALLES, BancTrad y eTitle; véase ap. 7).

Mención aparte merece el estudio por parte de expertos en didáctica de la lengua y en interferencias lingüísticas de las interferencias que incluimos entre las que debían ser tratadas por el corrector. Se empezó el trabajo recopilando la información, para proceder luego al análisis formal de las mismas y a su clasificación según sus características formales. A continuación, se llevó a cabo una implementación de prueba de las primeras reglas, cuyos resultados fueron analizados teniendo en cuenta sobretodo la interacción entre las reglas básicas de marcaje y las reglas del corrector. Finalmente se ha procedido a la creación de un prototipo de corrector para el catalán. Los estudios conducentes a este corrector han sido contrastados con los de otros proyectos (especialmente, el proyecto europeo ALLES).

Finalmente, una vez realizado el prototipo de corrector se ha procedido a la integración, en una única cadena general de procesamiento, de la gramática de marcaje del catalán con el prototipo de corrector. En este último estadio, por ejemplo, se ha modificado el acceso al léxico de la gramática de marcaje, de manera que en este nivel ya se procede al primer eslabón de la corrección, el de la corrección ortográfica.

Antes de terminar estas notas introductorias sobre la metodología seguida, conviene mencionar que en el seno de los grupos de investigación se han llevado a cabo estudios y discusiones sobre las posibilidades de ampliar el corrector para convertirlo en un corrector de estilo. Para

ello, conviene conseguir un tratamiento relevante de la semántica y la pragmática. Aunque estos estudios y discusiones están todavía en un estado incipiente, se puede reseñar que existen dos estrategias distintas:

- por un lado, se puede intentar implementar gramáticas de bajo nivel (en máquinas de estados finitos, probablemente en el formalismo de la *Constraint Grammar* mismo) que traten fenómenos y aspectos semánticos y pragmáticos para dar cuenta de las características de los errores de estilo comprendidos en textos producidos por los hablantes tipo del corrector
- por otra parte, se puede conectar una gramática de alto nivel (posiblemente implementada en un formalismo de unificación) al resultado de la gramática de marcaje obtenida en el presente proyecto, de manera que sea esta gramática de alto nivel la que dé cuenta de las características semánticas y pragmáticas que puedan reconocer y explicar los errores de estilo.

Será objeto de otro proyecto futuro seguir con estas investigaciones.

La presente memoria se estructura como sigue: en primer lugar, presentamos el sistema de procesamiento desarrollado durante el proyecto: los módulos lingüísticos para catalán y español (aps. 2 y 3), los resultados de su evaluación (ap. 4), y el módulo de corrección (ap. 5). A continuación, describimos los trabajos complementarios realizados sobre el léxico catalán (ap. 6), los proyectos en que se reutilizan las herramientas desarrolladas en PrADo (ap. 7) y los resultados de investigación y las publicaciones derivadas del proyecto (ap. 8). Finalmente, la memoria termina con unas conclusiones y reflexiones de cara a proyectos futuros (ap. 9), así como una serie de anejos.

2 Descripción del sistema: módulo para el catalán (CatCG)

El sistema de procesamiento lingüístico para el catalán, denominado CatCG, tiene una estructura altamente modular (Badia et al. 2001). Su arquitectura está representada en la Figura 1. El sistema consta de un módulo de preproceso (ap. 2.1), un módulo de proyección morfológica, una gramática de desambiguación morfosintáctica (ap. 2.2), y dos gramáticas sintácticas (proyección y desambiguación; ap. 2.3). Los primeros dos módulos están implementados en un lenguaje de programación, las gramáticas en el formalismo *Constraint Grammar* (Karlsson 1995, Tapanainen 1996).

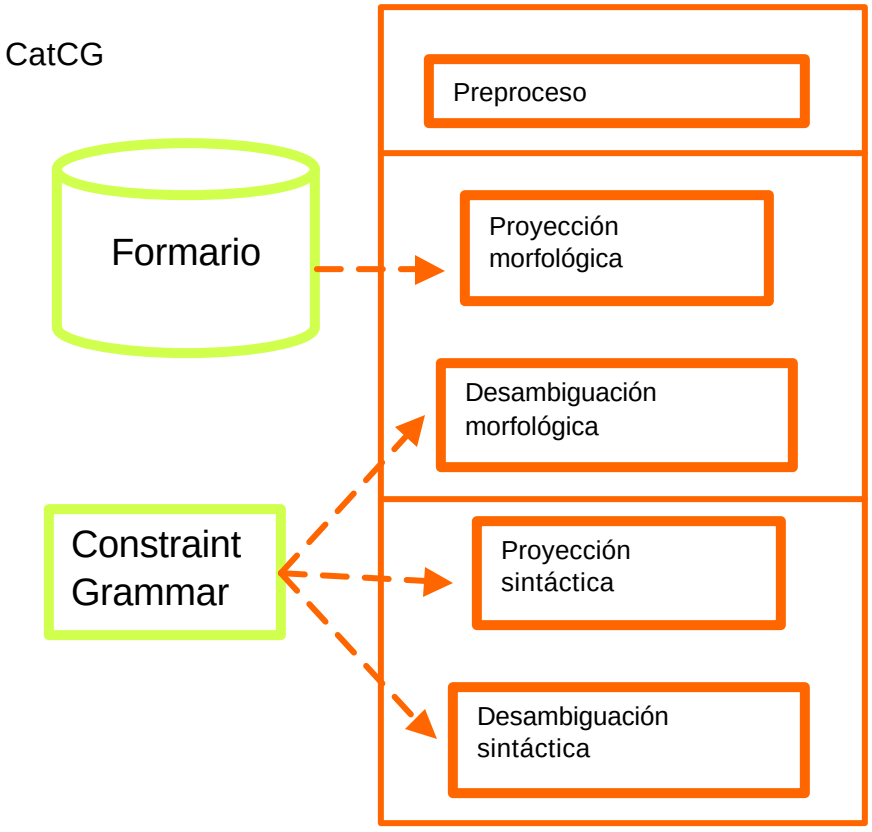

Figura 1 Esquema de la arquitectura de análisis superficial morfosintáctico para el catalán

2.1 Preproceso

El preproceso para textos catalanes ha sido desarrollado enteramente dentro del marco del proyecto PrADo. La función del preproceso es preparar el texto de entrada para un tratamiento gramatical posterior. Esto incluye todos los pasos que no se pueden realizar en el formalismo principal en que se realiza la gramática, en nuestro caso el formalismo de la *Constraint Grammar* La tarea principal del preproceso es la segmentación de palabras y unidades mayores como oraciones y párrafos. Todas estas unidades se pueden detectar, en la gran mayoría de los casos, mediante un tratamiento formal de los textos sin tener que recurrir a conocimiento lingüístico.

Aunque la separación de palabras parece una tarea fácil, en la práctica se presentan una serie de problemas. En la mayoría de los casos hay espacios o signos de puntuación señalando los límites entre las palabras. Pero ni es verdad que los signos de puntuación siempre separen palabras ni es verdad que las palabras siempre tengan un separador explícito. A continuación veremos los casos más problemáticos.

- En catalán existen palabras como *paral·lel* que se escriben con punto volado. Sin embargo, muchos usuarios usan el punto normal en lugar del punto volado para separar las dos eles. En este caso, el punto no marca ni una frontera de palabra ni una frontera de oración.
- En las gramáticas que constituyen la CatCG se consideran las formas contraídas como *del* una unión entre dos palabras. Consecuentemente se separan estas dos palabras y se reemplazan por las formas no contraídas, *de el*. La forma original contraída se puede recuperar mediante una etiqueta SGML que mantiene la forma en un atributo. En el ejemplo presente la etiqueta SGML sería la siguiente: <contrac forma="del">.
- Los clíticos catalanes se adjuntan al verbo y se separan de este mediante un guión o una comilla sencilla. Aunque los guiones y comillas son separadores, consideramos en estos casos que forman también parte de la palabra. Separamos por ejemplo la palabra compleja *m'agrada* ('me gusta') en *m'* y *agrada*.

Otras entidades que se marcan en el preproceso son nombres propios, abreviaturas y fechas. Las abreviaturas se detectan utilizando una lista de abreviaturas conocidas, la búsqueda de fechas se realiza mediante el reconocimiento de patrones. Para la detección de nombres propios se utiliza una estrategia mixta: por un lado se marcan nombres propios conocidos y por el otro lado se usa una heurística (detección de palabras no conocidas que empiezan por mayúscula).

La detección de oraciones y párrafos sigue una estrategia simple que interpreta los signos de puntuación como fronteras de oración. Para evitar interferencias con los puntos que se usan en las abreviaturas, se aplica el módulo de detección de abreviaturas en una fase anterior. Se marcan las oraciones y párrafos con etiquetas SGML y se asigna a cada unidad un número de identificación. El último paso del preproceso consiste en una verticalización del texto, donde cada palabra o signo de puntuación ocupa una línea separada en el formato de salida.

Después del etiquetado morfológico del texto, que se describe más abajo, se convierte el formato etiquetado en el formato específico que exige la *Constraint Grammar*. Esto se consigue

con una simple conversión de etiquetas morfológicas en una versión desplegada de la información que contiene la etiqueta.

La Figura 2 representa el módulo de preproceso para el catalán de una manera detallada.

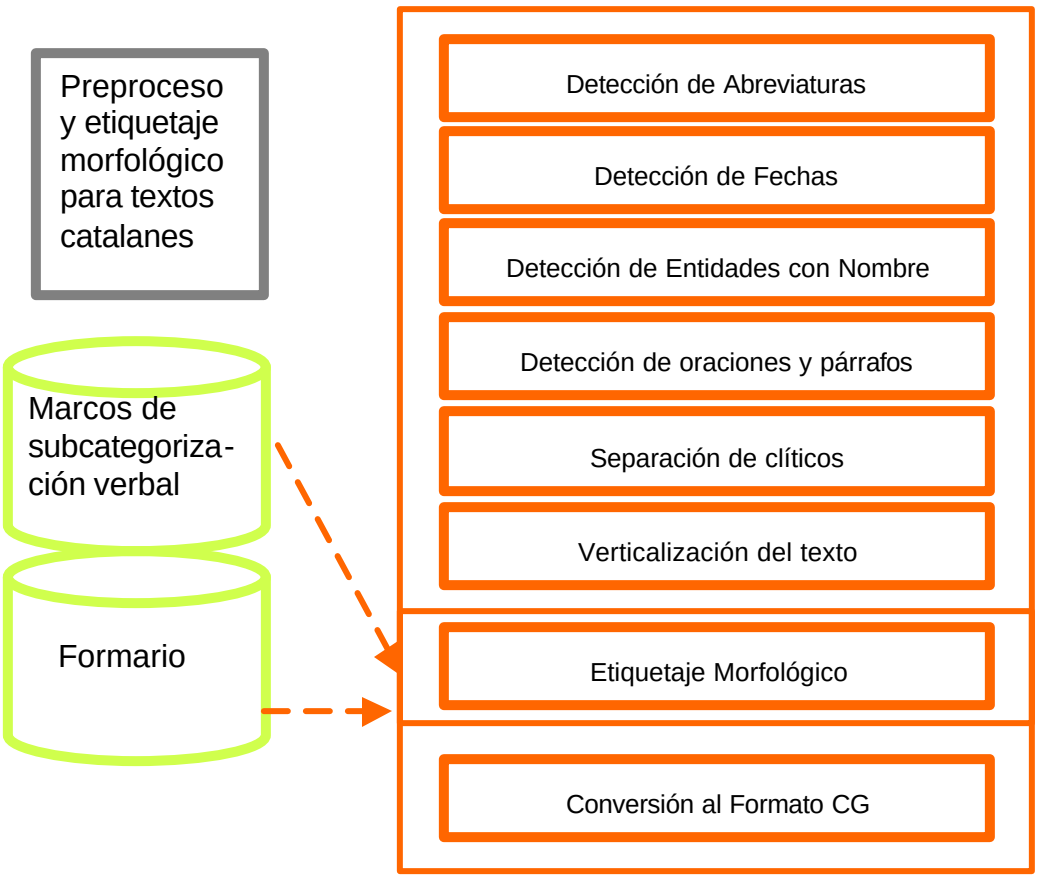

Figura 2 Esquema del preproceso en el módulo para el catalán

2.2 Morfología

Dentro de CatCG, el módulo que realiza el análisis morfológico se compone esencialmente de dos submódulos: el primero de ellos, el de proyección morfológica, asigna a cada una de las palabras del texto todas sus lecturas posibles. El segundo de ellos, el módulo de desambiguación morfológica, tiene por objetivo desambiguar aquellas palabras que presenten más de una lectura en su estado anterior (resultado de la proyección morfológica), de tal modo que sólo permanezcan las lecturas que son realmente posibles en función del contexto inmediato (y que idealmente debería ser sólo una).

2.2.1 Módulo de proyección morfológica

Tal como se refleja en la Figura 1 al principio del apartado 2, la asignación de lecturas posibles a cada palabra (proyección morfológica) se realiza justo después del módulo de preproceso. En este estadio, tal y como se ha explicado en el apartado 2.1, los elementos del texto han sido segmentados en palabras, oraciones y párrafos. Asimismo, se han identificado algunas entidades textuales algo más peculiares como fechas, nombres propios, etc.

La proyección morfológica se realiza mediante un programa codificado en el lenguaje de programación PERL (*Practical Extraction and Retrieval Language*). Este programa recorre todas las palabras aparecidas en el texto y las coteja con todas las lecturas disponibles en el *formario* (diccionario de formas), el cual consta de unas 600.000 formas con la información siguiente: categorías morfosintáctica principal y secundaria; en caso de que corresponda, género, número, tiempo, modo, persona, etc.

2.2.1.1 Construcción y mantenimiento del formario

La construcción del formario se realizó a partir de un módulo de análisis morfológico en línea preexistente llamado CATMORF y descrito en Tuells (1998). Esta herramienta, construida a partir del DIEC (1996) y del DLC (1995), contiene alrededor de 90.000 lemas. Dado que CATMORF fue concebido como un analizador y no como un generador morfológico, el proceso de vaciamiento del diccionario se ha realizado de forma *ad hoc* Esencialmente, se han generado de forma automática todas las formas posibles de todos los lemas (en función de sus categorías posibles y de su caracterización paradigmática) y éstas han sido analizadas mediante CATMORF. Como resultado se ha obtenido una lista de formas que conforman el subconjunto de palabras del catalán contenido en forma de lemas en CATMORF. Esta versión, revisada y ampliada de forma semi-automática, es la que ahora conforma la base de datos léxica de CatCG.

Asimismo, existe una base de datos de patrones de subcategorización para unos 8.000 verbos. Esta base de datos ha sido tomada, revisada y ampliada de un trabajo anterior realizado en colaboración con el Institut d'Estudis Catalans. Actualmente, se está trabajando en métodos de adquisición léxica para poder extraer a partir de corpus información sobre los patrones de subcategorización de los verbos que todavía no aparecen en la base de datos anteriormente mencionada (véase ap. 6.2).

Como se explicará más adelante, este módulo de proyección morfológica deberá sufrir algún que otro cambio para que sea útil para la corrección de textos.

2.2.2 Módulo de desambiguación morfológica

El módulo de desambiguación morfológica está implementado en el formalismo *Constraint Grammar*, en adelante CG (Karlsson 1995, Tapanainen 1996). La estrategia esencial de esta aproximación consiste en obtener un análisis morfosintáctico parcial a partir de la información contextual proporcionada en cada oración.

El formalismo CG, desarrollado en la Universidad de Helsinki, se caracteriza por facilitar la escritura de reglas de restricciones que permiten definir los contextos en los que se deben seleccionar o eliminar lecturas. Este formalismo está implementado en el lenguaje de programación *C* y, desde un punto de vista computacional, se fundamenta en máquinas de estados finitos (compiladas en cascadas) que determinan las secuencias (de palabras) que deben darse para que se apliquen las operaciones de eliminación o selección. Las reglas reflejan la estructura de las máquinas de estados finitos y podrían describirse como expresiones regulares complejas.

2.2.2.1 Las reglas en el formalismo CG

Las reglas en el formalismo CG responden todas a la siguiente sintaxis:[1]

OPERADOR (ETIQUETA) **TARGET** (ETIQUETA) **IF** (**-1** SET) (**0** SET) (**1** SET) ;

El primer elemento de la regla es el tipo de operación a realizar: REMOVE (eliminar), SELECT (seleccionar), MAP o ADD (proyectar), REPLACE (reemplazar), o NEW (añadir). Como veremos a lo largo de la descripción de esta arquitectura, los dos primeros se usan esencialmente en los módulos de desambiguación, mientras que el resto se usa en los módulos de proyección.

El segundo elemento indica la etiqueta afectada por el operador, es decir, aquella que será eliminada, seleccionada, proyectada, etc. La palabra TARGET es opcional (pero si aparece, es un literal) y se utiliza para indicar la categoría morfológica de la palabra sobre la que se realiza la operación –esta opción se usa esencialmente en los módulos de proyección.

[1] Lo que se explica a continuación en relación con el formalismo CG es válido tanto para el catalán como para el español, y tanto para la morfología como para la sintaxis, e incluso la detección de errores.

El literal IF indica la separación de las reglas en dos partes: la parte izquierda que determina la acción de la regla; y la parte derecha que indica la descripción de la secuencia afectada por la regla. Si la descripción que se halla en una regla casa con la que se halla en el texto, entonces se aplica la acción (parte izquierda de la regla).

La parte descriptiva de la regla indica la secuencia de palabras que deben hallarse para que se aplique la regla. La posición cero (0) indica la palabra sobre la que nos hallamos (y sobre la que se realizará la operación). El resto de posiciones expresan lugares a la izquierda (enteros negativos) o a la derecha (enteros positivos) respecto a la palabra cero.

Uno de los aspectos esenciales de la CG es la flexibilidad y el potencial expresivo de las reglas: se pueden expresar posiciones relativas o absolutas (y posiciones relativas o absolutas respecto a éstas, y así recursivamente); se pueden expresar negaciones; se pueden expresar posiciones "más allá de", mediante el símbolo * (estrella de Kleene); se pueden crear barreras (BARRIER) para que se deje de buscar contexto si se da una determinada condición, etc. Algunas de estas opciones se pueden ver ejemplificadas a lo largo de este informe; para otras aconsejamos la lectura de Tapanainen (1996).

2.2.2.2 Descripción del fichero de reglas CG para la desambiguación morfológica del catalán

La implementación del módulo de desambiguación morfológica del catalán se ha fundamentado en una serie de decisiones teóricas. En general, se ha optado por una teoría subyacente a la implementación que:

- respetara los datos lingüísticos hallados en los corpus de desarrollo (esencialmente dos: un corpus compilado a partir de la web, y otro tomado del CTILC (Rafel 1994)),
- fuera coherente con la investigación actual en lingüística, sin desviarse innecesariamente de la tradición,
- permitiera una implementación coherente y robusta de las reglas

A efectos prácticos esto implicó la determinación de una lista concreta de categorías principales y secundarias, fundamentadas en un conjunto de criterios de orden morfológico, sintáctico y, menos frecuentemente, semántico. Los detalles de la propuesta, basada también en varias gramáticas para el catalán (Badia 1995, Solà, et al., 2002), se describen extensamente en Quixal (2003). La lista de categorías y subcategorías principales se halla en el Anejo A.

Para dar cuenta del tipo de reglas que se incluyen en el módulo morfológico presentamos algún ejemplo:

```
REMOVE (VERB) IF (-1C PREP) (0 NOMBRE) (NOT 0 VERBO-NO-FINITO) ;
```

Esta regla sostiene que si se halla una palabra (*target*) que tiene entre sus lecturas posibles la de nombre y va precedida de una palabra que sólo tiene la lectura de preposición (marcado por la **C** que sigue al **-1**), se debe eliminar de entre las lecturas posibles de la palabra *target* la de verbo (si es que la tuviera, por supuesto). Además, existe una condición adicional marcada por **NOT 0** mediante la que se evita que la lectura eliminada sea una lectura de infinitivo, gerundio o participio (formas del verbo que, en ocasiones, presentan un comportamiento propio del nombre).

Una regla como esta se aplicaría a una secuencia como *de tanques* (que tiene dos traducciones literales posibles: '*de cierras', o 'de cercas'), cuyo análisis morfológico inicial sería:

```
"<de>"
        "de" Prep P
"<tanques>"
        "tanca" Nom com fem pl N5-FP
        "tancar" Verb MInd Pres 2pers sg VDR2S-
```

Figura 3 Ambigüedad resultante de la proyección morfológica para *de tanques*

Dando como resultado el siguiente análisis:

```
"<de>"
        "de" Prep P
"<tanques>"
        "tanca" Nom com fem pl N5-FP
```

Figura 4 Desambiguación de la secuencia *de tanques*

Otro tipo de regla más compleja sería el que sigue:

```
REMOVE (NOM) IF (-1C INTENSIFICADOR_ADJ) (0 ADJ + FS) (**-2 NOM + FS BARRIER
    DELIMITADOR_CLAUSULA OR VERBO-FINITO LINK -1 DET + FS) ;
```

Esta regla se aplicaría en una oración como *costejats en una proporció més gran pels més rics*, que viene a querer decir “costeados en una proporción mayor por los más ricos”. Lo que en palabras define la regla anterior es que:

- si hallamos una palabra que tiene entre sus lecturas la de adjetivo
- si esta palabra va precedida de un intensificador adjetival (*més, menys, tan*, etc.)
- si además se halla un nombre en la posición o posiciones anteriores al intensificador, siempre que no se traspase un delimitador de cláusula o un verbo finito (como indica la barrera, **BARRIER**)
- si además este nombre va precedido a su vez de un determinante
- entonces podemos eliminar la lectura nombre de la palabra en la posición cero (que tiene seguro una lectura adjetival)

Para ejemplificar algunos detalles de la versatilidad del formalismo CG, hemos ilustrado en esta regla el funcionamiento de los operadores **BARRIER** y **LINK**. Asimismo, observamos que en la posición -2 hay dos asteriscos en lugar de uno: esto implica que la operación de búsqueda no se detiene en el primer nombre que encuentra sino el segundo, lo cual permite que exista entre la palabra afectada y el nombre precedido de determinante alguna estructura algo más compleja.

Una regla como esta permite analizar correctamente la frase antes mencionada dando como resultado el siguiente:

```
"<Costejats>"
        "costejar" <S> <O> <NA> Verb Part masc pl maj VC--MP
"<en>"
        "en" Prep P
"<una>"
        "un" Det card-indef fem sg E6--FS
"<proporció>"
        "proporció" Nom com fem sg N5-FS
"<més>"
        "més" Adv D4
"<gran>"
        "gran" Adj qual masc-fem sg JQ--6S
<contrac forma="pels">
"<per>"
        "per" Prep P
"<els>"
        "el" Det art masc pl EA--MP
</contrac>
"<més>"
        "més" Adv D4
"<rics>"
        "ric" Adj qual masc pl JQ--MP
"<$.>"
```

Figura 5 Desambiguación de nombre y adjetivo en una estructura compleja

2.2.2.3 El fichero de reglas de desambiguación morfológica en cifras

Este módulo de desambiguación morfológica tiene actualmente 1.117 reglas estructuradas en dos bloques. El primer bloque está pensado para la eliminación de lecturas muy frecuentes o para resolver casos muy *ad hoc*, y contiene 42 reglas. El segundo bloque, conformado por las 1.075 reglas restantes, se encarga de resolver las ambigüedades más frecuentes según los corpus de entrenamiento.

Para una visión más detallada de los porcentajes de desambiguación previos y posteriores a la aplicación de este módulo, véase el apartado relativo a los resultados de la evaluación (4.2.1).

2.3 Sintaxis

El módulo de sintaxis de la CatCG proporciona información sobre la función sintáctica de cada palabra. Es un analizador superficial (*shallow parser*), es decir, no anota de manera que se pueda reconstruir un árbol sintáctico completo, sino que da información subespecificada, que se puede refinar en módulos posteriores.

A diferencia de otros modelos de análisis superficial, denominados *chunkers*, la CatCG no proporciona información sobre constituencia, sino que focaliza en información parcial sobre *función* y *dependencia*.

En la CatCG, la función sintáctica que se asignaría a un sintagma o constituyente (p.ej. complemento directo o sujeto) se asigna a una sola palabra, a la que se considera el núcleo de este sintagma. Por ello es vital decidir en el ámbito teórico cuál es el núcleo de cada constituyente; por ello también son problemáticas estructuras como las comparativas, en que no es fácil decidirlo. Para el resto de palabras se anota fundamentalmente información de dependencia, que servirá para identificar el núcleo del cual depende. Evidentemente, esta información se puede reutilizar en un módulo posterior para identificar constituencia, pero la CatCG no da esta información directamente.

Veamos un ejemplo. En un sintagma como *la casa de la cantonada* ('la casa de la esquina'), la preposición *de* recibiría la función de complemento del nombre (<CN), pues es el núcleo del constituyente *de la cantonada* y este constituyente modifica el nombre *casa*. Al nombre *cantonada*, en cambio, se le asigna simplemente función de complemento o término de la preposición (<P); de manera similar, al artículo *la* se le asignará determinante de nombre (DN>).

Los símbolos '<' y '>' que forman parte de las etiquetas funcionales sirven para indicar la dirección en que se encuentra el núcleo: en el caso de *cantonada*, sabemos que la preposición se encuentra a la izquierda (<P), pero nada nos indica si el núcleo es *de* o bien otra preposición situada más a la izquierda. La Figura 6 muestra el análisis completo de la frase *la casa de la cantonada és molt maca* ('la casa de la esquina es muy bonita')[2]:

[2] La herramienta se puede probar en la dirección web http://mutis.upf.es/catcg/

```
"<la>"
        "el" Det art fem sg EA--FS @DN>
"<casa>"
        "casa" Nom com fem sg N5-FS @Subj
"<de>"
        "de" Prep P @Advl
"<la>"
        "el" Det art fem sg EA--FS @DN>
"<cantonada>"
        "cantonada" Nom com fem sg N5-FS @<P
"<és>"
        "ser" <SS> <A> Verb MInd Pres 3pers sg VDR3S- @VPrin
"<molt>"
        "molt" Adv D4 @AA/A>
"<maca>"
        "maco" Adj qual fem sg JQ--FS @Atr
```

Figura 6 Análisis para el sintagma *la casa de la cantonada*

Los criterios que se han seguido para el desarrollo de la gramática son enteramente paralelos a los criterios para el módulo de morfología, repetidos aquí:

- respetar los datos lingüísticos hallados en los corpus de desarrollo,
- ser coherente con la investigación actual en lingüística, sin desviarse innecesariamente de la tradición,
- permitir una implementación coherente y robusta de las reglas

Con estos criterios, se ha llegado a definir las 25 etiquetas sintácticas actuales, cuyo nombre y glosa se reseña en la Tabla 1:

Función	Glosa	**Función**	Glosa
<AA/A, AA/A>	Adjunto a un adverbio o adjetivo	**<NN**	Nombre complemento de nombre
Advl	Adverbial (adjuntos verbales y oracionales)	**<P, P>**	Complemento de una preposición
AP	Aposición	**P-CD**	Preposición introductora de CD
Atr	Atributo en una oración copulativa	**Pred**	Complemento predicativo
<C	Complemento de conjunción completiva	**Pr-reflex**	Pronombre reflexivo

CD	Complemento directo	**P-Subj**	Preposición introductora de sujeto
CD-clt	Complemento directo (pronombre clítico)	**<QP**	Preposición complemento de cuantificador
CI-clt	Complemento indirecto (pronombre clítico)	**Subj-clt**	Sujeto (pronombre clítico)
<CN, CN>	Complemento de un núcleo nominal	**VAux>**	Verbo auxiliar
Conj	Elemento conjuntivo	**VPrin**	Verbo principal
DN>	Determinante de nombre		

Tabla 1 Funciones sintácticas en la CatCG

En cuanto a la implementación, el módulo de sintaxis de la CatCG consta de dos componentes: el de **proyección** de funciones sintácticas y el de **desambiguación** de las funciones.

2.3.1 Módulo de proyección

A diferencia de la proyección morfológica (v. ap. 2.2.1), la proyección sintáctica se realiza de manera controlada, es decir, evitando proyectar lecturas ambiguas en contextos suficientemente seguros. A continuación tenemos un ejemplo de regla de proyección:

```
MAP (ATR) TARGET (ADJ) IF (-1 VCOP) (NOT *1 NOM BARRIER BAR-DF OR COMA) ;
```

En esta regla se proyecta la función atributo (ATR) en adjetivos (ADJ) precedidos por un verbo copulativo (VCOP) y no seguidos de nombre (NOT NOM). Así, esta función, menos habitual que las de modificador de nombre, no se proyecta indiscriminadamente y el módulo de desambiguación es más ligero.

El módulo de proyección tiene actualmente 171 reglas y está finalizado.

2.3.2 Módulo de desambiguación

Las reglas de desambiguación sintáctica son totalmente paralelas a las de desambiguación morfológica (v. ap. 2.2.2). La siguiente regla (simplificada) indica que se debe eliminar la lectura sujeto (SUBJ) de un nombre (NOM) si delante tiene una preposición (PREP):

```
REMOVE TARGET (SUBJ) IF (0 NOM) (-1C PREP);
```

El módulo de desambiguación tiene actualmente 1343 reglas y está en una fase de desarrollo muy avanzada; aproximadamente al 90%. En el apartado 4.2.2 se ofrece un análisis cuantitativo y cualitativo de los resultados del módulo.

3 Descripción del sistema: módulo para el español (CastCG)

En el marco del proyecto PrADo y ya desde sus inicios se definió un modelo de usuario al que irían dirigidas en su momento las herramientas. Para el desarrollo de las herramientas ha sido necesario recopilar un corpus con textos cuyas características se correspondieran con las del perfil de un hipotético usuario predeterminado:

- Usuario trilingüe catalán, español e inglés (los textos son todos peninsulares)
- Usuario actual (en ningún caso textos anteriores a 01-01-2000)
- Usuario de nivel cultural y de redacción medio-alto

Los textos que conforman dicho corpus fueron recopilados en la Universitat Autònoma de Barcelona entre Marzo y Mayo de 2002. Actualmente consta de 238.766 palabras y está organizado jerárquicamente siguiendo una estructura temática. Así, dispone de textos tanto de lenguaje no especializado (correo electrónico, web, literatura y prensa actual) como de lenguaje especializado (derecho y lingüística, básicamente).

La necesidad de disponer de dicho corpus de desarrollo se justifica en tanto que hay un usuario predeterminado que generará un tipo de textos más o menos predecibles y de los cuales los distintos módulos de la herramienta deben poder dar cuenta con una alta eficacia. Además, contar con un amplio corpus de textos del tipo de los que la herramienta deberá tratar permite disponer de un potencial subcorpus de errores tipo para desarrollar las herramientas de corrección.

Para la explotación del corpus, se diseñaron una serie de herramientas informáticas (disponibles actualmente en el sitio web del proyecto: http://prado.uab.es). Las herramientas se dividen en dos bloques: en primer lugar, aquellas que trabajan con el corpus en formato texto, y, en segundo lugar, aquellas que lo manejan ya etiquetado. Esto permite hacer búsquedas de distinto grado de complejidad.

La primera de estas herramientas permite recuperar todas aquellas frases (entendidas como cualquier secuencia de texto separada por un punto) que contengan aquella cadena de caracteres que el usuario elija. Esto permite crear un corpus específico y adecuado a las necesidades puntuales del usuario. Las otras dos herramientas, en cambio, trabajan sobre un corpus ya etiquetado. La primera de ellas permite obtener una lista de todas aquellas palabras que compartan un cierto tipo de ambigüedad categorial: el usuario puede especificar hasta tres categorías distintas, entre las cuales las palabras de la lista serán ambiguas. La segunda

herramienta recupera concordancias morfológicas, es decir, recupera todas aquellas secuencias de palabras que concuerden con los patrones morfológicos deseados.

Estas herramientas fueron fundamentales para la búsqueda y estudio de los distintos fenómenos lingüísticos que se debían contemplar a la hora de construir el módulo de desambiguación morfológica. Cabe decir, por último, que la última de las herramientas disponibles en línea es una *demo* del módulo de etiquetado y desambiguación morfológica para el español, en su fase actual.

Este apartado está dedicado al módulo lingüístico para el español, describiendo cada uno de sus módulos: preproceso (ap. 3.1), etiquetado morfológico (ap. 3.2) y etiquetado sintáctico (ap. 3.3).

3.1 Preproceso

A diferencia del preproceso para textos catalanes, se usa una herramienta externa para el preproceso castellano. Se trata del etiquetador MACO+, desarrollado en la Universitat Politécnica de Barcelona. Aunque MACO+ es en principio un etiquetador, incorpora un preproceso propio. Esto ha facilitado la creación de un preproceso por un lado, pero por otro lado se ha tenido que adaptar el formato de salida a las necesidades de una gramática CG.

En primer lugar, se ha tenido que implementar una detección de abreviaturas. MACO+ asigna una etiqueta especial para las abreviaturas, pero esta etiqueta no tiene valor para la creación de reglas gramaticales. Más importante que el hecho de que una unidad es una abreviatura es la categoría morfológica a la que pertenece esta unidad. Por ejemplo “Prof.” abrevia la palabra “Profesor” y pertenece a la categoría *nombre* y aunque es una abreviatura no pertenece a una categoría morfológica relevante que se llame *abreviatura*.

El preproceso interno de MACO+ tampoco detecta las fronteras de oraciones, por lo que tuvimos que crear un módulo que marcara estas unidades con una etiqueta SGML.

Un punto más importante aún es que MACO+ no distingue entre verbos sencillos y verbos que llevan un pronombre cliticizado o una secuencia de clíticos, como por ejemplo el imperativo *dámelo*. A diferencia del catalán, en la ortografía del español no se usa un carácter de puntuación para separar el pronombre clítico del verbo. Este hecho dificulta la detección de los clíticos. Hemos creado un módulo que los detecte basándose en la forma verbal (detectada por MACO+) y la terminación del verbo. El etiquetario que usa MACO+ no prevé esta información, así que la anotamos directamente sobre el formato que utiliza la *Constraint Grammar*. Para un

verbo imperativo como *dámelo*, por ejemplo, se añade la información *cl1: me cl2:lo*. Aunque los clíticos en español no se representan como unidades autónomas, la información se conserva en la anotación sobre el verbo y las reglas de la CG pueden usar esta información.

La Figura 7 representa el módulo de preprocesamiento para el español.

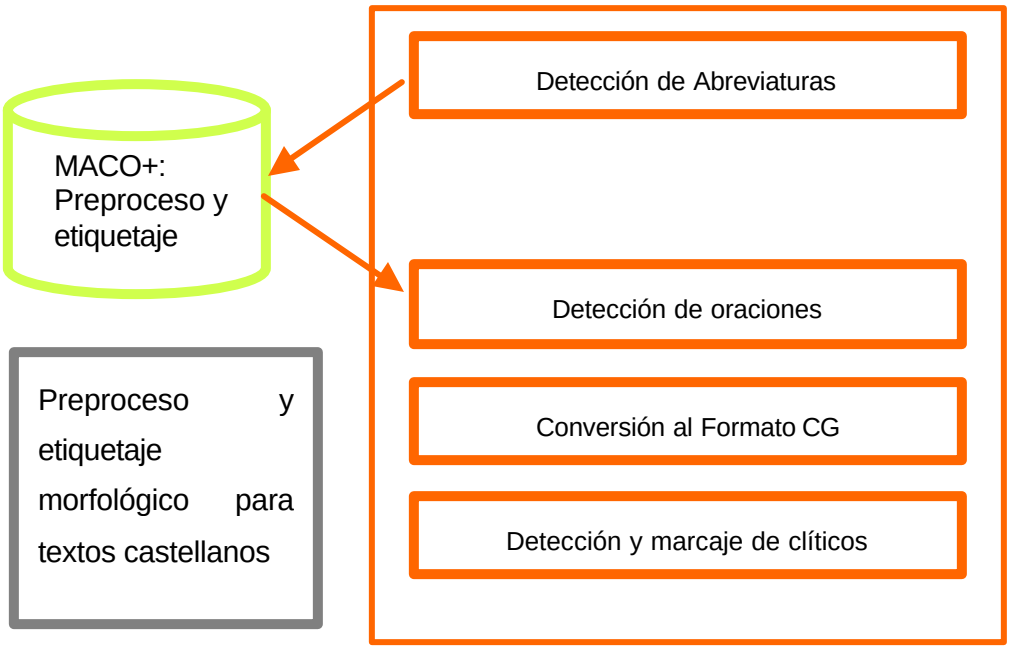

Figura 7 Esquema del módulo de preproceso para el español

Por último, cabe señalar que a raíz de los problemas de compatibilidad señalados, y en el marco del Programa de Cooperación Interuniversitaria AL.E.2003, Paola Caymes Scutari ha desarrollado un módulo de preproceso alternativo que identifica diferentes categorías que en las etapas posteriores no son reconocidas, como fechas, nombres, monedas, siglas, abreviaturas, números, direcciones de Internet y correos electrónicos. A cada una de ellas las etiqueta como corresponde y expande siglas y abreviaturas. Este módulo está pensado para constituir la base de un analizador morfológico, de modo que en un futuro sea posible prescindir de MACO y del costoso proceso de conversión de etiquetas.

3.2 Morfología

Para la proyección morfológica, como ya se ha señalado, el punto de partida es el analizador morfológico desarrollado por la Universitat Politécnica de Catalunya, MACO+. Tras un proceso

automático, las etiquetas de MACO (versión española del estándar europeo de PAROLE) son reconvertidas a un formato reconocible por el motor de la *Constraint Grammar*.

Además, es en este momento cuando se introducen determinados cambios en el etiquetado basados en evidencias lingüísticas que permiten el tratamiento unitario de distintas palabras con un funcionamiento morfosintáctico idéntico, que las gramáticas y diccionarios han tratado separadamente, como categorías morfológicas distintas (entiéndase categoría morfológica en el sentido más tradicional del término). Es el caso, por ejemplo, de la tradicional distinción entre determinantes, adjetivos y pronombres, muchas veces de frontera confusa o incluso difusa; todas aquellas palabras cuyos lemas permiten tanto la función pronominal como la determinativa han sido analizadas bajo la etiqueta unitaria de *Especificador*. Por ejemplo, en el caso de **ambas**, en el que podría haberse encontrado una triple ambigüedad, se dan únicamente dos posibles lecturas, ya que *determinante* y *pronombre* son una sola:

```
"<ambas>"
        "amba" Nom com fem pl – NCFP000
        "ambos" Esp card fem pl DC3FP0
```

Figura 8 Lecturas de la palabra *ambas*

En el caso de **la,** en cambio, se mantiene la distinción entre *pronombre* y *especificador*, ya que se trata de lemas distintos:

```
"<la>"
        "el" Esp art fem sg DA3FS0
        "la"  Nom com masc sg NCMS000
        "él" Pron person febl 3pers fem pl acus PP3FSA00
```

Figura 9 Lecturas de la palabra *la*

La información morfológica aportada por el *input*, el texto analizado morfológicamente con todas las lecturas posibles, se modifica a partir de reglas restrictivas basadas en información lingüística contextual cuya ventana máxima es la oración. El objetivo del proceso de desambiguación morfológica es, precisamente, eliminar, o reducir al máximo, estas ambigüedades a partir de la implementación de reglas en el formalismo CG explicado en el apartado 2.2.2.1.

La gramática para el español se estructura de manera similar a la del catalán, en dos bloques. El primer bloque de reglas consiste en una serie de reglas particulares que eliminan aquellas

ambigüedades que son producto de un etiquetado demasiado amplio y que no está en consonancia con nuestro perfil de usuario, por tratarse de anacronismos, americanismos, etc.

El resto de reglas se organizan a partir de las distintas categorías morfológicas: en primer lugar se escribieron las reglas referentes a categorías cerradas (especificador, conjunción, preposición, pronombre y adverbio), ya que trabajar con paradigmas cerrados permitió posteriormente delimitar mejor el resto de categorías (a saber, nombre, adjetivo, verbo) y sus rasgos menores (género, número, modo, etc.).

Para la delimitación de la frontera categorial, además de los ejemplos reales aportados por el corpus (que como ya se ha dicho se corresponden con el tipo de lengua utilizada por el usuario preestablecido) se ha recurrido a dos tipos de fuentes bibliográficas: diccionarios (especialmente el DRAE) y gramáticas (GDLE). Asimismo, se ha intentado mantener una coherencia teórica con el módulo del catalán, para asegurar así una buena interacción entre ambos componentes en la fase de corrección.

3.2.1 Estrategias de desambiguación entre nombre y verbo

La ambigüedad entre nombre y verbo era una de las más importantes en el momento de iniciar el proceso de desambiguación morfológica (14138 formas ambiguas, es decir, un 5,92% del corpus). Las principales estrategias que se siguieron para desambiguarlas son las siguientes:

Falta de concordancia: se pueden desambiguar algunas ocurrencias detectando casos en los que no hay concordancia. Por ejemplo, podremos descartar la combinación ESPECIFICADOR + NOMBRE en aquellos casos en los que ambas palabras no concuerden en género y nombre (por ejemplo, en la secuencia *La recuerdo*). Este método, aunque sólo permite desambiguar un número limitado de casos, es muy seguro y no presenta problemas de sobredesambiguación.

Restricciones distribucionales: la mayor parte de reglas que hemos creado se sirven de criterios distribucionales para eliminar lecturas; es decir, se debe detectar un contexto en el que la lectura nominal o la lectura verbal no sea posible. Por ejemplo, podremos descartar la lectura verbal cuando la palabra ambigua sea un verbo potencial en forma personal y esté precedido de una preposición. La lectura nominal puede ser eliminada si la forma ambigua esta precedida por un pronombre fuerte, demostrativo o relativo (con la excepción de *cuyo*).

Presencia de otro verbo principal: una estrategia muy útil para resolver este tipo de ambigüedad consiste en intentar detectar otro verbo principal en la frase. Si es posible detectarlo, la forma ambigua no será un verbo (puesto que hay otro verbo principal) y, por

consiguiente, se tratará de un nombre. Para que no sobredesambigüe en el caso de oraciones relativas, subordinadas o coordinadas, se controla el ámbito de la regla.

Problemas del formario: En algunos casos, el formario tiene algunos problemas que crean más ambigüedad de la que realmente existe. Por ejemplo, MACO asigna la lectura verbal a formas imposibles de VERBO FINITO + CLÍTICO, como, por ejemplo, *asistente* Para eliminar la lectura verbal de esta forma, se ha creado un grupo de formas que nunca podrán recibir la interpretación verbal.

Después de aplicar todas las reglas de desambiguación, quedan 2595 formas ambiguas. Es decir, la ambigüedad se reduce más del 81%.

3.2.2 Estrategias de desambiguación entre pronombre y especificador

La ambigüedad entre nombre y verbo afectaba también a un número elevado de formas (6965, es decir, el 2,91% del conjunto del corpus). Además, se trata de una ambigüedad que afecta a palabras de uso muy frecuente, como, por ejemplo, los artículos (*la, las, los*). La estrategia más común para desambiguar estos casos, como en el caso de los verbos, es la de restringir el contexto de manera que sólo una de las dos lecturas sea posible.

Se puede eliminar la lectura de pronombre en los siguientes contextos:

- si la forma no va seguida de verbo: *las flores* vs. *las cantas*
- si la forma va seguida de un nombre no ambiguo: *las niñas* vs. *las cuentas*

Se puede eliminar la lectura de especificador en los siguientes contextos:

- si la forma va seguida de verbo no ambiguo: *los contratamos.*

Después del proceso de desambiguación, quedan 1.064 formas ambiguas entre especificador y pronombre, lo que supone una reducción del 84,7%.

3.2.3 Estrategias de desambiguación entre preposición y cualquier otra categoría

Antes de iniciar el proceso de desambiguación, se podían encontrar en el corpus de desarrollo 21420 palabras (un 8,97%) con una lectura prepositiva entre sus posibles categorías léxicas. Las principales estrategias para proceder a la desambiguación se dividen según las categorías afectadas:

Preposición y adverbio: la mayoría de los casos de ambigüedad entre ambas categorías se pueden resolver en uno u otro sentido partiendo del criterio de transitividad, es decir, la preposición es siempre transitiva y, por tanto, siempre está inmediatamente seguida de algún tipo de estructura nominal, mientras en el caso del adverbio no es así.

Preposición y nombre: para proceder a la resolución de esta ambigüedad se usaron criterios distribucionales de combinación posible y válida en una oración tipo, por ejemplo, implementando reglas que eliminen la posibilidad de que dos preposiciones se encuentren seguidas o de que una preposición vaya precedida de un especificador.

Preposición y verbo: referidas a los casos de "bajo" y "entre", es la sección de reglas más *ad hoc* de las referidas a preposiciones, puesto que resulta casi imposible generalizar. Su alto índice de frecuencia de aparición justifica la necesidad de contar con este bloque de reglas.

Tras el proceso de desambiguación de esta categoría, el número de casos ambiguos queda en 66, lo que en términos estadísticos supone una reducción de la ambigüedad del 99'7%.

3.2.4 Estrategias de desambiguación entre conjunción y cualquier otra categoría

Para la desambiguación de aquellas palabras cuyas posibles lecturas contenían una conjuntiva (17635 casos, es decir, un 7,38%) se procedió de la misma manera que en el caso de las preposiciones, distinguiendo entre subtipos de ambigüedades:

Conjunción y nombre: para este caso se recurrió a estrategias de tipo distributivo, por ejemplo, la imposibilidad de estructuras del tipo *Artículo* + *Conjunción*.

Conjunción y pronombre relativo: este tipo de ambigüedad se encuentra en 7696 casos. En el caso del castellano no se ha podido contar con información de subcategorización léxica, por lo que se complicó enormemente la tarea de desambiguación entre conjunción y pronombre dentro del Sintagma Verbal. En el caso del Sintagma Nominal que inicia la oración, la tarea fue mucho más sencilla, ya que con estrategias de concordancia entre el núcleo nominal y el verbal el problema se resuelve de manera muy efectiva. Finalmente, también se recurrió a estrategias que tuvieran en cuenta la coordinación entre dos oraciones (sustantivas o relativas, según el caso). Tras el proceso de desambiguación, restan 2687 palabras ambiguas entre ambas categorías, es decir, la ambigüedad se ha reducido un 65,09%.

Tras todo el proceso, el número de casos ambiguos se reduce a 3150 (incluidos aquí los 2687 especificados en el apartado anterior), es decir, la ambigüedad se reduce un 82,14%.

3.2.5 Estrategias de desambiguación entre verbo y (otro) verbo

Fue necesario crear un apartado propio para la ambigüedad entre verbos, que afectaba a 24.795 palabras (un 10,38% del corpus) y que puede deberse al hecho de tratarse de lemas distintos o, dentro de un mismo lema, personas, tiempos o modos distintos. Las estrategias básicas que se han seguido son:

Concordancia: número y persona

Modo: presencia de nexos que introduzcan modalidad, presencia de verbos en modo indicativo...

Eliminación de lecturas no procedentes: hay un gran número de verbos cuya ambigüedad se explica por la posibilidad de ser confundidos con verbos desusados o americanismos cuya lectura se ha eliminado ya que no procede según el modelo de usuario.

La efectividad de este bloque de reglas es del 87,41%. En todos los tipos de ambigüedad se ha preferido siempre la corrección por encima de la efectividad y, de hecho, es muy difícil hacer este tipo de desambiguación de manera coherente, sin generalizaciones que puedan causar errores, ya que hay algunos rasgos morfológicos, especialmente en el caso de los verbos, que no es posible desambiguar sin contar con información de tipo sintáctica, de subcategorización léxica o incluso semántica y pragmática.

3.2.6 Estrategias de desambiguación entre adverbio y verbo

La ambigüedad entre adverbios y verbos, identificables tanto por sus distintas propiedades morfológicas como por su distribución, afectaba a 1014 casos (un 0,42% del corpus). Algunos de los casos tomados como ambiguos presentaban ambigüedad sólo si se tiene en cuenta la diacronía. Dado que las acepciones en desuso no han de ser desambiguadas por el sistema, en formas como *antiguo* o *claro*, no se tiene en cuenta la lectura verbal.

Se detectaron también errores de formario, ya que MACO, tras haber aislado erróneamente un pronombre enclítico en formas como *bastante,* daba como posible la interpretación verbal. En otros casos, el error consistía en etiquetar conjuntamente construcciones que admiten valores categoriales distintos a los tratados en las expresiones cohesionadas. Por ejemplo, *aparte de* es de uso frecuente también con la lectura verbal (“apartar a alguien de algo”). Para alguna forma no se generaba una categoría posible, como en *conforme*, cuya lectura nominal, con escasa frecuencia de uso pero registrada en el DRAE, no estaba prevista.

La identificación de la categoría es relativamente sencilla en algunos casos. Tal es el caso de *adentro o demasiado*, que de ser verbos son necesariamente pronominales. Para la mayoría, no obstante, se han creado diversas reglas que tienen en cuenta la capacidad flexiva de los verbos, la selección de complementos (a veces precedidos de preposiciones concretas), el orden de palabras en la oración y la inclusión de los adverbios en contextos fijos. Algunas de estas reglas tienen un alcance amplio y permiten identificar buena parte de los verbos y adverbios. En otros casos, las reglas se han creado para resolver la ambigüedad de unos pocos casos. El sistema posee capacidad predictiva más allá del corpus de entrenamiento, puesto que existen reglas capaces de resolver ambigüedades que realmente no causan problemas en el corpus, debido a la escasísima frecuencia de uso de una de las categorías asociadas a las formas. Así sucede con el uso verbal de *abajo, adelante*, *adentro*, *apenas*, *así, cerca, demasiado, encima*, *justo, medio, nada* o *tarde* En la actualidad, los casos de ambigüedad en el corpus se han reducido a 93, por lo que el porcentaje de ambigüedad ha disminuido un 90,83%.

3.2.7 Otras estrategias de desambiguación para adverbios

La estrategia era basarnos en desambiguar cualquier elemento que tuviera asociada la categoría de adverbio. Para ello, se llevó a cabo una búsqueda con las herramientas disponibles en http://prado.uabes sobre el corpus etiquetado: en concreto, se obtuvo una lista de los tipos de ambigüedades que se dan entre las diferentes categorías y la categoría de *adverbio*, con un par de ejemplos de cada tipo.

La clasificación que se obtiene es la que se muestra en la tabla siguiente:

Ambigüedad	**Ejemplos**
ADVERBIO /NOMBRE	"<acaso>"
ADVERBIO/CONJUNCIÓN	"<aun>"
ADVERBIO/INTERJ/CONJUNCIÓN	"<entonces>"
ADVERBIO/INTERJECCIÓN	"<atrás>"
ADVERBIO/INTERJECCIÓN/NOMBRE	"<despacio>"
ADVERBIO/INTERJECCIÓN/VERBO	"<abajo>"
ADVERBIO/INTERJECCIÓN/VERBO	"<fuera>"
ADVERBIO/NOMBRE/ADJETIVO/ESPECIFICADOR/VERBO	“<medio>"
ADVERBIO/NOMBRE/PRONOMBRE/ VERBO	"<nada>"
ADVERBIO/NOMBRE/PRONOMBRE/ESPECIFICADOR	"<cuanto>"
ADVERBIO/NOMBRE/VERBO	"<cerca>"

ADVERBIO/PREPOSICIÓN	"<hasta>"
ADVERBIO/PRONOMBRE	"<algo>"
ADVERBIO/PRONOMBRE	"<sí>"
ADVERBIO/PRONOMBRE/ESPECIFICADOR	"<demás>"
ADVERBIO/VERBO	"<apenas>"
ADVERBIO/VERBO/PRONOMBRE/ESPECIFICADOR	"<bastante>"

Tabla 2 Ambigüedades en que interviene la categoría adverbio

La información del corpus se complementó con un análisis de los criterios usados en la tradición gramatical para la correcta asignación de una etiqueta categorial a las unidades del español. El manual teórico básico consultado ha sido la *Gramática descriptiva de la lengua española*, compilada por I. Bosque y V. Demonte (Espasa Calpe, Madrid, 1999), para delimitar las clases gramaticales, aunque también se ha recurrido a la información gramatical contenida en el *Diccionario de la lengua española* de la Real Academia Española (Espasa Calpe, 2001, 22ª edición) para consultar la etiqueta gramatical de determinadas piezas léxicas.

Enunciamos a continuación las principales estrategias adoptadas en el proceso de desambiguación morfológica.

- Dado el carácter particular de algunas de las piezas léxicas que pueden funcionar como adverbios, se ha utilizado como estrategia básica de desambiguación la separación en conjuntos formados por uno o más lemas: así distinguimos entre adverbios transitivos y adverbios intransitivos, o entre cuantificadores de función adjetiva y función adverbial. En ocasiones, ha sido necesario formular reglas para un lema en concreto, como es el caso de “sí” o “qué”.

- Las lecturas ambiguas con la etiqueta ‘Interjección’ se restringen, mediante una regla referida a los lemas que la contengan, a los ámbitos señalados por una marca de exclamación. Asimismo, algunas piezas léxicas que solo funcionan como interjección en ámbitos de habla relegados de nuestro estudio (americanismos o arcaísmos, básicamente) se han considerado problema del diccionario de formas, adoptándose la decisión de eliminar dichas lecturas de los lemas implicados.

- Dado su particularmente complicado funcionamiento, se ha diseñado un bloque de reglas específicas para los adverbios que aparecen en estructuras comparativas.

- Para desambiguar adverbios y conjunciones, se ha propuesto un análisis basado en el ámbito oracional: a modo de ejemplo, “mientras” será conjunción (en este caso, nexo de simultaneidad) a no ser que tras él aparezca algún signo de puntuación (equivale entonces a “mientras tanto”).

- Para desambiguar adverbios y pronombres, se ha adoptado el criterio propuesto en Brucart (1999) (dentro de la *Gramática descriptiva de la lengua española*) según el cual *donde, cuando, como* son siempre adverbios relativos (a diferencia de la tradición gramatical, que consideraba una doble categorización, dependiendo de la presencia o ausencia de un antecedente: en el primer caso, sería un adverbio relativo, y en el segundo una conjunción). Ahora bien, la salida que ofrece **maco+** no contempla la lectura adverbial para estas piezas; ante esta situación, se ha optado por seleccionar la lectura pronominal en todos aquellos casos en que debería ser adverbio, con el fin de poder reemplazar esta lectura por la adverbial en un estado futuro del módulo.
- Para desambiguar adverbios, especificadores y pronombres, se ha recurrido al criterio de la concordancia. Por ejemplo, se pueden descartar las lecturas de especificador en aquellos casos en que no concuerden en género y número con la palabra siguiente.
- Para desambiguar adverbios y nombres, hemos usado básicamente criterios distribucionales para eliminar lecturas imposibles en contextos determinados: en otras palabras, se identifican dominios sintácticos en que la pieza que estamos analizando concurra con otras categorías específicas. A título de ejemplo, los adverbios no admiten las expansiones propias de los nombres, por lo que no pueden aparecer en combinación con los artículos, demostrativos, adjetivos y cuantificadores nominales.
- Para desambiguar adverbios y verbos, en los casos en que no concurre otra categoría y que no hayan quedado tratados por reglas anteriores, se han formulado reglas puntuales directamente sobre los lemas implicados: es el caso de “antiguo”, “tarde” o “medio”.

3.3 Sintaxis

El procedimiento que se ha seguido para la creación del módulo sintáctico de la CastCG, que se aplica sobre el resultado del módulo de desambiguación morfológica, ha sido el de procesar textos en español con los analizadores creados para el catalán y trabajar esos módulos a partir de los errores y ambigüedades detectados.

Los errores y ambigüedades detectados y que se han debido trabajar de forma específica para estos módulos se detallan a continuación.

3.3.1 Errores

3.3.1.1 Por categorías morfológicas

- Nombres temporales. En la CatCG se detectaban en un módulo del preproceso, en la CastCG se han convertido en una lista de palabras en el módulo de proyección.
- Pronombres personales. Los oblicuos recibían función de sujeto. Se han modificado las reglas que asignaban erróneamente esta función.
- Verbos de oraciones subordinadas. La falta de patrones de subcategorización para los verbos en el lexicón español provoca numerosos errores, algunos de los cuales se han podido enmendar a partir de la creación y modificación de reglas de proyección y desambiguación sintácticas.
- Tiempos verbales perifrásticos. El auxiliar recibía la función de verbo principal (VPrin).
- Nombres después de infinitivo. Recibían sujeto, sistemáticamente.
- Verbos copulativos seguidos de forma no personal. Eran considerados auxiliares.
- Enumeraciones. Además del problema arrastrado de la CatCG de la desambiguación entre enumeraciones y aposiciones, en muchos casos no se proyectaba la función pertinente.

3.3.1.2 Por función sintáctica

- No se proyectaba la función *atributo* en copulativas con el verbo ‘ser’.
- Se proyectaba *atributo* en oraciones predicativas.
- Errores en la proyección de complemento de preposición cuando el nombre no iba precedido inmediatamente de preposición.
- Errores de proyección en los pronombres átonos.
- Error de selección de complemento directo para algunos sujetos.
- Desambiguación de funciones que la CastCG debería dejar ambiguas (entre algunos casos de atributo y sujeto y de complemento directo y sujeto).

3.3.2 Ambigüedades

- Verbos que recibían todas o casi todas las funciones, y verbos que no recibían ninguna debido a errores en el fichero de proyección sintáctica.
- Ambigüedades entre sujeto, complemento directo y atributo, tanto en oraciones copulativas como en oraciones predicativas.
- Ambigüedades entre sujeto y complemento directo en oraciones con verbo en forma no personal.
- Ambigüedades entre aposiciones y complementos directos en contextos donde no debería permanecer la lectura de aposición.

Una vez localizados los errores y las ambigüedades que se daban al aplicar los módulos sintácticos de la CatCG, se han revisado los ficheros de reglas (tanto el de proyección como el desambiguación sintáctica) y se hacen las modificaciones pertinentes. Ha sido útil la creación de un corpus *ad hoc* de unas mismas oraciones en catalán y en castellano para determinar por qué algunos casos quedaban solucionados en la CatCG y daban un resultado erróneo en la CastCG. Una vez modificados los ficheros de reglas, quedan algunos problemas pendientes de ser solucionados (para más detalles, ver 4.3.2).

4 Evaluación de los módulos lingüísticos

4.1 Metodología

Hasta el momento se han realizado las evaluaciones de los siguientes módulos: morfológico de la CatCG, sintáctico de la CatCG y morfológico de la CastCG. Los resultados se mostrarán en apartados específicos destinados a los resultados de las tres evaluaciones que se han llevado a cabo hasta el momento. El módulo sintáctico de la CastCG, que ha sido el último en desarrollarse, todavía está pendiente de ser evaluado.

Las evaluaciones de los módulos morfológicos y del módulo sintáctico han sido planteadas de formas distintas debido a que para las evaluaciones de los módulos morfológicos contábamos con corpus etiquetados que nos servían como corpus de comparación.

El objetivo final de la evaluación de los módulos morfológicos era comparar (parcialmente) la salida desambiguada proporcionada por nuestros sistemas con el etiquetado de los corpus de

comparación. Mediante un proceso automático se comparan ambos análisis y se considera que, en caso de coincidencia, el análisis es correcto. En el caso de hallar una divergencia o de que nuestras gramáticas dejen algún tipo de ambigüedad, se considera un error. Aquí entra en juego el papel del evaluador humano, que decide qué herramienta ha hecho el análisis correcto y qué lectura debería haber sido la elegida en el caso de la ambigüedad renuente.

Durante este proceso de evaluación, nos limitamos a cotejar la categoría principal de las palabras correspondientes. Una evaluación más detallada se ha dejado para etapas posteriores.

Para la evaluación del módulo morfológico de la CatCG disponíamos de un fragmento del corpus CTILC, de l'Institut d'Estudis Catalans (Rafel 1994). Este corpus ha sido etiquetado semiautomáticamente y corregido manualmente. Para la evaluación del módulo morfológico de la CastCG contábamos con el corpus LexEsp (Sebastián, et al., 2000), etiquetado automáticamente.

Antes de la comparación de etiquetas se obviaron las divergencias de tipo teórico como, por ejemplo, el hecho de que muchas palabras que las CGs consideran especificadores, en el IEC y en el LexEsp son considerados pronombres. Una vez tuvimos los corpus comparados y alineados, la evaluación se planteó de la siguiente manera: se repartieron textos periódicamente entre lingüistas voluntarios (miembros del proyecto), a los que se facilitaron instrucciones de evaluación e información sobre los principales presupuestos teóricos y los análisis previstos. La evaluación se facilitó también mediante el uso de una barra de macros (de MS Word) que contenía las etiquetas a evaluar. Los evaluadores marcaron, usando la macro correspondiente, la categoría de cada una de las palabras divergentes.

La evaluación de los módulos sintácticos se tuvo que plantear de un modo distinto, ya que no disponíamos de corpus etiquetados sintácticamente que nos pudieran servir como corpus de comparación. El hecho, además, de que tuviéramos más de veinte funciones sintácticas distintas, hizo conveniente dividir la evaluación por categorías morfológicas, de manera que se evaluara por separado cada una de ellas. El procedimiento para la evaluación del módulo sintáctico de la CatCG fue el siguiente: se ha procesado el corpus con la CatCG y cada semana los evaluadores (lingüistas voluntarios, miembros del proyecto) recibieron un texto y la información necesaria sobre las posibles funciones que puede realizar la categoría morfológica que está evaluando, así como de nuevo la barra de macros con la que se marcaba el texto. La Tabla 3 resume los datos principales de las evaluaciones.

módulo	submódulo	corpus de comparación	núm. palabras
CatCG	morfología	CTILC (Rafel 1994)	220.000
	Sintaxis	-	12.397
CastCG	morfología	LexEsp (Sebastián, et al., 2000)	105.000

Tabla 3 Datos de las evaluaciones para los módulos lingüísticos

4.2 Resultados y estado de la cuestión para el catalán

4.2.1 Morfología

El corpus de evaluación del módulo de desambiguación morfológica consistía en un subcorpus del CTILC (Rafel 1994) de 252.841 instancias textuales, con unas 220.000 palabras, puesto que el resto eran signos de puntuación. Recuérdese que el CTILC es un corpus etiquetado automáticamente y corregido a mano.

De las 115.219 formas léxicas originalmente ambiguas, un total de 7077 (6,1%) permanecieron ambiguas (dos o más lecturas) después de haber procesado el corpus. Además, la cantidad de formas léxicas con todavía 3 posibles lecturas se había reducido a 168 (de originalmente 20446), y la cantidad de formas léxicas con 4 lecturas había disminuido a 1 (de originalmente 7971).

De las 1117 reglas de la gramática de desambiguación, solamente 773 se llegaron a aplicar (en este fragmento preciso del corpus de CTILC). De estas 773, sólo 19 de ellas se aplicaron 1000 veces o más, y solamente 142 se aplicaron 100 veces o más. Estas cifras reflejan cómo disminuye la cantidad de veces que se aplica una regla a medida que aumenta el número de reglas. Por tanto, queda también demostrada para esta aplicación lingüística la ley de Zipf: la mayor parte del trabajo se puede hacer con un pequeño esfuerzo, mientras el resto se logrará solamente dedicando comparablemente mucho más tiempo.

4.2.1.1 Errores más frecuentes

A continuación presentamos una tabla en la que las columnas (por orden) indican:

- etiqueta recibida según nuestro etiquetador
- etiqueta deseada según el criterio de los evaluadores
- total de casos hallados en el corpus de evaluación
- porcentaje que este tipo de error representa sobre el resto de errores

- porcentaje que este tipo de error representa en el total del corpus

RECIBIDA	DESEADA	TOTAL	% ERRORES	% CORPUS
NOM	+ADJ	399	12,87%	0,18%
PRON	+CONJ	357	11,52%	0,16%
VERB	+NOM	237	7,65%	0,11%
PRON	+DET	236	7,61%	0,11%
NOM	+VERB	213	6,87%	0,10%
ADJ	+NOM	197	6,35%	0,09%
Subtotal			**52,87%**	**0,75%**
NOM	+DET	114	3,68%	0,05%
ADV	+DET	95	3,06%	0,04%
CONJ	+PRON	87	2,81%	0,04%
NOM	+?	87	2,81%	0,04%
DET	+PRON	74	2,39%	0,03%
DET	+ADV	65	2,10%	0,03%
NOM	+INTERJ	63	2,03%	0,03%
DET	+NOM	55	1,77%	0,03%
DET/NOM	+?	54	1,74%	0,02%
NOM	+ADV	47	1,52%	0,02%
ADV	+NOM	46	1,48%	0,02%
ADV	+CONJ	41	1,32%	0,02%
NOM	+PREP	39	1,26%	0,02%
ADV	+PRON	37	1,19%	0,02%
ADJ	+ADV	34	1,10%	0,02%
ADJ/NOM	+?	32	1,03%	0,01%
Otros	-	491	15,84	0,22%
Total		**3100**	**100%**	**1,42%**

Tabla 4 Errores detectados en el corpus de evaluación: morfología CatCG

La tabla anterior refleja los errores más frecuentemente cometidos por nuestro *tagger*. De hecho, solamente refleja los ocurridos un mínimo de 30 veces o más (hecho que nos asegura que supongan al menos un 1% del total). El resto de errores se agrupa en una última línea titulada “otros”. Es alentador que el porcentaje total de error sea solamente de 1,42%. Este porcentaje confirma que hemos logrado controlar la aplicación de las reglas: dicho de otra manera, hemos logrado evitar que se apliquen las reglas en casos donde no estábamos seguros del acierto. Sin embargo, como veremos más adelante, el porcentaje de ambigüedad restante está alrededor del 7%-8% (incluyendo los elementos que no se han tenido en cuenta en la evaluación –principalmente participios o gerundios ambiguos con lecturas adjetivales).

4.2.1.2 Ambigüedades más frecuentes

A continuación presentamos una tabla en la que las columnas (por orden) indican:

- etiquetas recibidas según nuestro etiquetador

- etiqueta deseada según el criterio de los evaluadores
- total de casos hallados en el corpus de evaluación
- porcentaje que este tipo de ambigüedad representa sobre el resto de errores
- porcentaje que este tipo de ambigüedad representa en el total del corpus

RECIBIDA	DESEADA	TOTAL	% AMBIG.	% CORPUS
ADJ/NOM	+ADJ	3268	46,18%	1,49%
ADJ/NOM	+NOM	1701	24,04%	0,78%
NOM/VERB	+NOM	1120	15,83%	0,51%
Subtotal			**86,04%**	
DET/NOM	+NOM	309	4,37%	0,14%
NOM/VERB	+VERB	138	1,95%	0,06%
ADJ/VERB	+ADJ	112	1,58%	0,05%
DET/NOM	+DET	98	1,38%	0,04%
ADJ/VERB	+VERB	88	1,24%	0,04%
ADV/ADJ/NOM	+ADJ	55	0,78%	0,03%
Subtotal			**97,34%**	
ADJ/NOM/VERB	+VERB	45	0,64%	0,02%
ADJ/NOM/VERB	+ADJ	43	0,61%	0,02%
ADV/DET	+DET	27	0,38%	0,01%
ADJ/NOM/VERB	+NOM	20	0,28%	0,01%
ADV/DET	+ADV	12	0,17%	0,01%
ADV/NOM	+ADV	10	0,14%	0,00%
ADJ/PREP	+PREP	8	0,11%	0,00%
ADJ/PRON	+PRON	8	0,11%	0,00%
ADV/ADJ/NOM	+NOM	3	0,04%	0,00%
NOM/PREP	+PREP	2	0,03%	0,00%
NOM/PRON	+NOM	2	0,03%	0,00%
ADJ/NOM/PREP	+PREP	1	0,01%	0,00%
ADJ/PREP	+ADJ	1	0,01%	0,00%
ADJ/PRON	+ADJ	1	0,01%	0,00%
ADV/ADJ/NOM	+ADV	1	0,01%	0,00%
ADV/ADJ/NOM/VERB	+ADJ	1	0,01%	0,00%
ADV/VERB	+VERB	1	0,01%	0,00%
INTERJ/NOM	+INTERJ	1	0,01%	0,00%
NOM/PRON	+PRON	1	0,01%	0,00%
AMBIGÜEDAD TOTAL		**7077**	**100,00%**	**--**
Total de palabras		**219060**	**--**	**3,23%**

Tabla 5 Ambigüedades detectadas en el corpus de evaluación: morfología CatCG

Debemos destacar que el total de ambigüedad real es algo mayor que el que en esta tabla se indica: probablemente alrededor del 7% o 8%. Esto es debido al hecho de que había dos tipos de ambigüedad que conscientemente no queríamos evaluar: la ambigüedad entre participio/adjetivo y la ambigüedad entre gerundio/adjetivo. La razón de esta decisión es que en muchos casos no existía un criterio claro para distinguir entre ellas. Por tanto, decidimos

posponer el tratamiento de estos dos grupos de ambigüedad a un momento posterior de desarrollo.

Por otro lado, quizá lo más destacable de esta tabla sea que la ambigüedad nombre/adjetivo suma el 70% del total de ambigüedad restante. Esto demuestra algo conocido y estudiado por la teoría lingüística, a saber, que la distinción entre estas dos categorías es difícil de sentenciar sin información semántica, pero también que la transcategorización entre ellas es un hecho habitual. Asimismo, cabe destacar que la ambigüedad verbo/nombre también es bastante elevada, cosa que se puede explicar por la cantidad de participios que también presentan lecturas nominales (*pintada, rentada, batut, imprès*, etc.).

4.2.2 Sintaxis

Los resultados del proceso de evaluación para la sintaxis se reflejan en la tabla siguiente, con datos por categoría (en orden alfabético):

Categoría	**Precisión**[3]	**Cobertura**[4]	**F-score**[5] **(a = 0.5)**
Adjetivo	0,97	0,96	0,97
Adverbio	0,98	1	0,99
Conjunción	0,95	0,95	0,95
Determinante	0,96	0,95	0,95
Nombre	0,74	0,76	0,75
Preposición	0,93	0,55	0,74
Pronombre	0,90	0,69	0,79
Verbo	0,87	0,66	0,77
Media	*0,91*	*0,82*	*0,86*

Tabla 6 Resultado de la evaluación del módulo sintáctico

Vemos en la Tabla 6 que la precisión es en casi todos los casos mayor a la cobertura,[6] siguiendo una filosofía que implica asignar una función sólo cuando la decisión es fiable, tal y como se explicaba en el apartado anterior. Ello resulta en general en una pérdida de cobertura, que se puede subsanar de diferentes maneras, como veremos.

[3] Precisión = Aciertos / Errores + Aciertos.

[4] Cobertura = Aciertos / Ambigüedades restantes +Aciertos.

[5] F-score = αPrecisión + αCobertura. En este caso, como $\alpha = 0.5$, precisión y cobertura tienen el mismo peso, como es estándar en tareas de etiquetado morfosintáctico.

[6] La única excepción es la categoría nombre; véase apartado 4.2.2.1.

Los resultados indican que la herramienta está en un estado muy avanzado: para la mitad de las categorías (adjetivo, adverbio, conjunción, determinante), el coeficiente *F* es igual o superior a 0.95, es decir, la calidad del módulo sintáctico es comparable a un etiquetador morfológico estándar. Para el resto (nombre, preposición, pronombre y verbo), los coeficientes están alrededor del 0,76, indicando que son categorías en las que todavía hay que trabajar.

Cabe señalar que es de esperar que los peores resultados se obtengan en categorías mayores como nombre y verbo, pues son éstas las que reciben mayor número de funciones, y dichas funciones pueden presentar dependencias a larga distancia (sujeto, complemento directo, verbo principal o no, etc.). Sin embargo, se pueden mejorar los resultados en muchos casos si se aumenta la información léxica.

El Anejo B contiene datos detallados del estado y los tipos de problemas de cada categoría, desglosados en ambigüedad restante y errores. Aquí nos centraremos sin embargo en el análisis de las cuatro categorías más problemáticas, que serán las que habrá que revisar.

4.2.2.1 Nombre

En el nombre, tanto la precisión como la cobertura tienen valores bastante bajos: 0,74 y 0,76. De hecho, el valor de precisión es el más bajo de todas las categorías, es decir, es la categoría en que la función asignada es errónea en más casos.

Sin embargo, la mayoría de los errores (un 52%) son de función desconocida, es decir, el corrector no sabía cuál de las funciones previstas asignar a las ocurrencias. Este error es debido fundamentalmente a que no hay una función prevista para nombres en fragmentos: títulos, pies de fotografía, listas, etc. Habría que añadir una función “nombre principal” o similar, equivalente a la función de “verbo principal”, para estos casos. Cabe señalar que esta función sería relativamente sencilla de desambiguar. Sin estos errores, la precisión sería 0,86. Aunque esta cifra es sensiblemente mejor que la actual, todavía está lejos del umbral mínimo deseable, 0,95, es decir, habría que revisar también los otros errores.

En cuanto a la cobertura, más de un 60% de la ambigüedad restante corresponde a casos en que intervienen las funciones *sujeto* y *complemento directo* En una lengua de orden relativamente libre como es el catalán, creemos que la ambigüedad respecto a esta categoría sólo puede disminuirse con información de semántica léxica. Los casos que se desambiguan actualmente (aprox. 35%) se pueden detectar gracias a factores como la concordancia y la información de subcategorización, disponible en el léxico catalán.

4.2.2.2 Preposición

En el caso de la preposición, la precisión es aceptable (0,93), lo que indica que los casos que se desambiguan se desambiguan fiablemente. Es la cobertura (0,55) lo que hace bajar el coeficiente. En este caso, la ambigüedad más importante es la ambigüedad entre *complemento del nombre* y *adverbial* (equivalente a complemento del verbo, el SV o la oración): es el famoso problema del *PP-attachment* (adjunción del SP), sin duda el problema más grave para cualquier analizador o etiquetador sintáctico. Estos casos suman un 87% de los casos de ambigüedad restante. Para resolverlos, de nuevo, se necesita más información lexicosemántica, para poder determinar las relaciones entre los núcleos verbales o nominales y los SSPP.

4.2.2.3 Pronombre

En el caso de los pronombres la precisión (0,90) todavía es aceptable, aunque inferior a la deseable. El problema, pues, es la cobertura (0,69), que es muy baja por causa de fundamentalmente dos problemas: en primer lugar, la ambigüedad *sujeto*/*complemento directo*, que ya hemos comentado en el caso de los nombres, y que suma aprox. un 55% de los casos de ambigüedad restantes. En segundo lugar, la ambigüedad en los pronombres clíticos, a la que corresponde otro 37% de la ambigüedad restante. Estos casos se podrían disminuir en la propia CatCG sin necesidad de más información.

4.2.2.4 Verbo

En el verbo, la precisión (0,87) es de nuevo mayor que la cobertura (0,66), pero relativamente baja. Un 30% de los errores vienen de la incorrecta identificación del tipo de oración en que se encuentra el verbo: relativa, subordinada o principal.

Este es también el mayor problema para la cobertura, sobre todo la distinción entre subordinada y principal, ya que un 45% de la ambigüedad restante se debe a que el sistema no puede decidir qué tipo de oración es. Probablemente, una parte de estos errores se podría depurar todavía en la misma CatCG, sin información añadida, pero otros casos serán irresolubles con esta técnica: se necesitaría un formalismo que pudiera tener en cuenta constituyentes, no sólo la cadena de palabras.

4.2.2.5 Conclusión

Del análisis realizado se desprende que para las categorías adjetivo, adverbio, determinante y conjunción se está bastante cerca del umbral que se puede alcanzar, pues en todos los casos

el coeficiente *F* es igual o superior a 0,95. Para el resto de categorías, vemos que habría que refinar la gramática para pronombre y verbo y, secundariamente, para nombre y preposición. Sin embargo, para estas dos últimas categorías se está cerca del umbral, y para conseguir una mayor cobertura es necesario disponer en el léxico de una información lexicosemántica más amplia, en la línea de los trabajos que se están realizando ya en la Universitat Pompeu Fabra (véase ap. 6).

4.3 Resultados y estado de la cuestión para el español

4.3.1 Morfología

Actualmente, el desarrollo del módulo de desambiguación se considera terminado. El archivo de reglas contiene 743 reglas sobre ambigüedad morfológica, que en el corpus de desarrollo se aplican un total de 130.751 veces. El corpus de desarrollo recién etiquetado presenta un grado de ambigüedad del 64,78%, incluyendo aquí cualquier tipo de ambigüedad, tanto de categorías mayores, como en categorías menores. Después del proceso, el grado de ambigüedad se reduce a un 13,86%.

El corpus de evaluación del módulo de desambiguación morfológica procede del LexEsp y consta de unas 120.000 instancias textuales, de las que 104.591 son palabras. De este total, 4.923 permanecen ambiguas después de procesar el corpus.

Errores más frecuentes

A continuación presentamos una tabla en la que las columnas (por orden) indican:

- etiqueta recibida según nuestro etiquetador
- etiqueta deseada según el criterio de los evaluadores
- total de casos hallados en el corpus de evaluación
- porcentaje que este tipo de error representa sobre el resto de errores
- porcentaje que este tipo de error representa en el total del corpus

RECIBIDA	DESEADA	TOTAL	% DEL TOTAL DE ERRORES	% RESPECTO AL TOTAL DEL CORPUS
NOM	+ADJ	147	17,76	0,14
ESP_PRON	+DET	84	10,16	0,08
NOM	+VERB	54	6,53	0,05
NOM_VERB	+ADJ	54	6,53	0,05
CONJ	+PRON	46	5,56	0,04
ESP_NOM_ADV	+DET	33	3,99	0,03

Subtotal			50,5	0,40
ESP_NOM	+DET	28	3,39	0,03
ADJ	+NOM	24	2,90	0,02
NOM	+ADV	23	2,78	0,02
PRON	+ESP	22	2,66	0,02
NOM_ADV	+ADJ	21	2,54	0,02
PRON	+CONJ	19	2,30	0,02
ESP_ADV	+DET	14	1,69	0,01
VERB	+NOM	14	1,69	0,01
NOM	+ESP	12	1,45	0,01
NOM_NUM	+ESP	11	1,33	0,01
ADV	+ADJ	10	1,21	0,01
ESP	+ADJ	10	1,21	0,01
NOM_ADV	+ESP	10	1,21	0,01
ADV	+PRON	9	1,09	0,01
Otros		182	22,0	0,17
Total		827	100	0,79

Tabla 7 Errores detectados en el corpus de evaluación: morfología CastCG

La tabla anterior refleja los errores más frecuentemente cometidos por nuestro *tagger* De hecho, solamente refleja los ocurridos un mínimo de 9 veces o más (hecho que nos asegura que supongan al menos un 1% del total). El resto de errores se agrupa en una última línea titulada "otros". Es alentador que el porcentaje de error sea solamente de 0,79%. Este porcentaje es la prueba de que se ha preferido la corrección sobre la eficacia. Así hemos logrado controlar la aplicación de las reglas: dicho de otra manera, hemos logrado evitar que se apliquen las reglas en casos donde no estábamos seguros del acierto. De todos modos, como veremos más adelante, el porcentaje de ambigüedad restante está alrededor del 4,70%.

Ambigüedades más frecuentes

A continuación presentamos una tabla en la que las columnas (por orden) indican:

- etiquetas recibidas según nuestro etiquetador
- etiqueta deseada según el criterio de los evaluadores
- total de casos hallados en el corpus de evaluación
- porcentaje que este tipo de ambigüedad representa sobre el resto de errores
- porcentaje que este tipo de ambigüedad representa en el total del corpus

RECIBIDA	DESEADA	TOTAL	% DEL TOTAL DE AMBIG.	% RESPECTO AL TOTAL DEL CORPUS
CONJ/PRON	+PRON	816	16,58	0,78
ADJ/NOM	+ADJ	686	13,93	0,66
ADJ/NOM	+NOM	584	11,86	0,56
Subtotal			42,37	1,99
NOM/VERB	+NOM	495	10,05	0,47
CONJ/PRON	+CONJ	407	8,27	0,39
NOM/VERB	+VERB	335	6,80	0,32
ESP/PRON	+ESP	304	6,18	0,29
NOM/ADV	+ADV	261	5,30	0,25
ADJ/VERB	+ADJ	142	2,88	0,14
Subtotal			81,86	3,85
ESP/NOM	+ESP	96	1,95	0,09
CONJ/VERB	+CONJ	89	1,81	0,09
CONJ/ADV	+ADV	49	1,00	0,05
ESP/ADV	+ADV	48	0,98	0,05
ESP/NOM/ADV	+ESP	44	0,89	0,04
ADJ/NOM/ADV	+ADJ	38	0,77	0,04
ADJ/VERB	+VERB	38	0,77	0,04
ADJ/NOM/VERB	+ADJ	32	0,65	0,03
ADV/VERB	+ADV	32	0,65	0,03
PRON/ADV	+PRON	31	0,63	0,03
ADJ/ADV	+ADJ	30	0,61	0,03
ESP/NOM/PRON/ADV	+ADV	30	0,61	0,03
NOM/ADV	+NOM	27	0,55	0,03
ESP/NOM	+NOM	25	0,51	0,02
PRON/ADV	+ADV	23	0,47	0,02
ADJ/ESP	+ADJ	19	0,39	0,02
ESP/VERB	+ESP	18	0,37	0,02
ADJ/NOM/ADV	+ADV	17	0,35	0,02
ESP/PRON/ADV	+ADV	16	0,33	0,02
AMBIGÜEDAD TOTAL		**4923**	100	4,71
Total de palabras		**104591**		

Tabla 8 Ambigüedades detectadas en el corpus de evaluación: morfología CastCG

Debemos destacar que el total de ambigüedad real es algo mayor que el que en esta tabla se indica. Esto es debido al hecho de que había un tipo de ambigüedad que conscientemente no queríamos evaluar: la ambigüedad entre participio y adjetivo. La razón de esta decisión es que en muchos casos no existía un criterio claro para distinguir entre ambas lecturas. Por tanto, decidimos posponer el tratamiento de estos dos grupos de ambigüedad para un momento posterior de desarrollo.

Por otro lado, quizá lo más destacable de esta tabla sea que la ambigüedad nombre/adjetivo suma el 26% del total de ambigüedad restante. Esto demuestra algo conocido y estudiado por la teoría lingüística, a saber que la distinción entre estas dos categorías es difícil de sentenciar

sin información semántica, pero también que la transcategorización entre ellas es un hecho habitual. Asimismo, cabe destacar que la ambigüedad conjunción/pronombre también es igualmente elevada, cosa que se puede explicar por la falta de información de subcategorización léxica: la tarea de desambiguación entre conjunción y pronombre dentro del Sintagma Verbal es muy difícil de abarcar si no se sabe de antemano, por ejemplo, qué verbos aceptan completivas y cuáles no. Es por ello que está previsto incorporar esta información sobre subcategorización en una futura fase de revisión del proceso de etiquetado morfológico.

4.3.2 Sintaxis

Aunque para la gramática sintáctica del español no se ha realizado una evaluación con correctores humanos, cabe señalar algunos problemas pendientes de ser solucionados:

- La falta de una subcategorización verbal para el castellano. Con una subcategorización verbal mínima se podrían reducir los casos de ambigüedad entre complemento directo y sujeto. También serviría para dar un trato diferenciado a los complementos verbales regidos, que de momento reciben la misma función que los complementos adjuntos.
- Algunas categorías como verbos o conjunciones deberán tratarse mejor a partir de la creación de nuevas reglas para depurar algunas ambigüedades solucionables a partir de los módulos existentes en la CastCG.

Muchos de los problemas restantes no son específicos de la CastCG, sino compartidos con la CatCG. Los listamos a continuación:

- Problemas puntuales como el análisis de los títulos, números, fechas, porcentajes… (este problema debe solucionarse en el preproceso).
- La ambigüedad entre sujeto y complemento directo. Se necesitaría algún tipo de información de carácter semántico para poder desambiguar algunos casos de ambigüedad entre ambas funciones.
- La ambigüedad entre algunos casos de sujeto y atributo en oraciones con ‘ser’ con dos sintagmas nominales. La teoría que hemos propuesto para el tratamiento de estas oraciones en la CatCG y en la CastCG deja algunos casos en los que no es posible decidir entre una u otra función.
- Casos de adjunción de preposición (*PP-Attachment*), en que actualmente no es posible decidir a nivel superficial si es complemento del nombre o del verbo.

5 Descripción del sistema: módulo de corrección

5.1 Modelo de usuario y tipología de errores

En el presente proyecto el usuario tipo se ha planteado según un perfil sociolingüístico, como un sujeto conocedor de las dos lenguas oficiales habladas en Cataluña, en sus variedades estándar escritas, con una competencia de redacción media o alta, aunque vulnerable a la producción de interferencias en ambas lenguas. El hecho de habernos planteado el perfil del usuario tipo en los términos expuestos ha respondido a la necesidad de construir un prototipo de corrector gramatical capaz de superar la limitación impuesta por la figura de un usuario general.

Hemos incluido también entre el usuario potencial aquel que se ocupa, de forma profesional o eventual, de traducir al español o al catalán textos especializados o pseudo-especializados redactados originalmente en inglés, susceptible, pues, de interferencias de esta lengua. Exceptuando tal vez la traducción de textos de un cierto valor literario, de edición más cuidada y con menores restricciones de tiempo, existe un importante corpus de textos, fundamentalmente técnicos o pseudo-técnicos, cuya traducción debe efectuarse en un breve espacio de tiempo y para cuya revisión puede ser de mucha utilidad contar con una herramienta como la que hemos propuesto.

La tipología de errores que se propone es la de Hernández (1998), según la cual, de manera muy resumida y teniendo en cuenta que hay muchas zonas de intersección entre los distintos grupos, los errores de interferencia se podrían clasificar a partir de seis niveles:

1. Nivel fónico
2. Nivel gramatical (interferencia morfológica e interferencia sintáctica)
3. Nivel léxico-semántico (calco formal y calco semántico)
4. Nivel discursivo
5. Nivel pragmático
6. Nivel gráfico

Para comprobar y completar la tipología de errores que se acaba de presentar, se han vaciado los siguientes documentos de las parejas de lenguas que se indican:

1. **Castellano-inglés / inglés-castellano**. Vaciado de los casos de interferencia aparecidos en las siguientes obras:

    a. Prado, M. (2001) *Diccionario de falsos amigos inglés-español*, Madrid, Gredos, contrastando las decisiones de M. Prado con las definiciones de las palabras en el DRAE (2001).

    b. Thompson, J. (1989) *Los timadores*, Júcar, Etiqueta Negra. Traducción de M. A. F. Álvarez-Nava.

    c. Gómez Torrego, L. (1995) *El léxico en el español actual: uso y norma*, Madrid, ArcoLibros.

    d. García Yebra, V. (1982) *Teoría y práctica de la traducción*. Madrid: Gredos.

    Asimismo, se han analizado diversas secuencias de los servicios informativos de los medios de comunicación orales (radio, TV).

2. **Castellano-catalán / Catalán-castellano** Vaciado de los casos de interferencia que aparecen documentados en:

    a. Solà, Joan (1995) *Llibre d'estil de l'Ajuntament de Barcelona*. Amb la col·laboració de Xavier Fargas, Anna Gudiol i Alba Fraser, Barcelona, Consorci per a la Normalització Lingüística.

    b. Solà, Joan (1989) *Qüestions controvertides de sintaxi catalana,* 2a ed. Barcelona, Edicions 62.

    c. Badia i Margarit, Antoni M. (1995) *Gramàtica de la llengua catalana descriptiva, normativa, diatòpica, diastràtica*, Barcelona, Proa.

    d. Ruaix i Vinyet, Josep (1996) *Diccionari auxiliar,* Barcelona/Moià.

Presentamos a continuación una breve muestra de análisis de algunas interferencias catalán-castellano seleccionadas al azar:

1. Cambías, diferencías
2. Ir **a la nuestra**
3. Hace **una calor...**
4. *Haber* más artículo definido (*vas a casa de unos amigos y **hay la** tele puesta*)

5. Deber + infinitivo con valor de probabilidad (***debe hacer*** *quince días o tres semanas que estoy por aquí*)
6. *Tú* en vez de *ti* (*con un resumen que ellos te hacen* ***a tú*** *ver...*)
7. *Por* en vez de *para (será jodido* ***por*** *la persona que entre...)*
8. Queísmo (*es consciente* ***que****)*
9. *Doble negación preverbal (esta última semana la gente* ***tampoco no*** *ha estado tanto hablando del Gran Hermano)*
10. Construcciones con *hacer*: *cuando hacen futbol, hacer años, hacer fuera...*
11. *Todo y así* (aun así) (*todo el día con el Canal Digital que... que te enteras de todo, pero* ***todo y así***)
12. *Claredad, arideza, enseñamiento…*
13. *Mandra, plegar...*

5.1.1 Análisis

1. **"Cambías", "diferencías":** ortográfica. La detectaría el corrector.
2. **"Ir a la nuestra":** podría ser morfológica -cambio de género- o léxica. Habría que indicar que se trata de una estructura formada por el verbo *ir + prep. + art. det. + posesivo*. Habría que decir también que desde la preposición hasta el posesivo siempre se mantiene en singular, mientras que el verbo va cambiando (no hace falta conjugar el verbo: con el infinitivo es suficiente).
3. "Hace **una calor":** morfológica. Aplicación del género equivocado al sustantivo "calor". Simplemente habría que indicar el cambio de género. Otros ejemplos: la análisis, dientes perfectas, la olor...
4. **"Hay la tele** puesta": se trata de una interferencia sintáctica (haber + artículo determinado). El verbo *haber*, salvo raras excepciones, no puede ir seguido del artículo determinado. Se trataría de sustituirlo por la 3ª p.sing. del verbo "estar". La sustitución puede ser de diversas maneras: *esta noche hay los fuegos artificiales* (son los/ hay fuegos...); *mañana hay la fiesta mayor de Calella* (es la...). Como es complicado porque no todos los ejemplos se sustituirían por la misma forma, deberíamos reunir el máximo número de casos posibles e indicar la solución de cada caso. En una fase posterior se

vería si se puede unificar de algún modo o cómo habría que indicarlo, pero, por el momento, sería suficiente con esto.

5. **“Debe hacer** 15 días que estoy por aquí”: se trata de una interferencia gramatical; el verbo “deber” con valor de probabilidad siempre va seguido de la preposición “de”. Se trata de añadirla. En algunos casos la dificultad estribaría en detectar el valor de probabilidad (con frecuencia resulta ambiguo), pero solucionar este problema también quedaría para una fase posterior.
6. “Un resumen que ellos te hacen **a tú”:** gramatical; pronombres. El pronombre personal de 2ª persona precedido de preposición no es “tú” sino “ti”. Se trata de reemplazar uno por otro.
7. “Será jodido **por la persona** que entre”: gramatical; preposiciones. En este caso, aunque no se vea en el fragmento seleccionado, el valor del “por” no es causal (será jodido a causa de la persona que entre), sino final -con reservas lo de “final”- (será jodido para la persona que entre). Como en los demás casos dudosos, sería suficiente con indicarlo: la solución se vería más adelante.
8. **“Es consciente que”:** gramatical; de régimen. Determinados verbos o expresiones rigen la preposición “de” ante la conjunción “que” (ser consciente de que, estar seguro de que...). Se trataría, pues, de añadir dicha preposición. Habría que listar todos los verbos o expresiones que la rigen. Otros ejemplos de error: quejarse que, darse cuenta que, el hecho que...
9. “La gente **tampoco no ha hablado** tanto del Gran Hermano”: gramatical; negación; doble negación preverbal. En español no puede haber dos partículas negativas en situación preverbal. Se trataría de eliminar el “no”. Habría que buscar todas las posibles dobles negaciones más frecuentes e introducirlas con la indicación de eliminar el “no”. Otro ejemplo de error: nadie no.
10. **“Hacen fútbol, hacer años, hacer fuera...”:** léxico-semánticas. Es complicado definir el fenómeno. Son expresiones que en una lengua funcionan y en la otra, no. La solución, por tanto, sería acumular el máximo de expresiones “incorrectas” con “hacer” e indicar directamente la corrección. Soluciones: hay fútbol, dan fútbol, cumplen años, echan, etc.
11. “Te enteras de todo, pero **todo y así...”:** conectores; discursiva. Se trata simplemente de sustituir el conector incorrecto por “aun así”.

12. “Claredad”, “arideza”, “enseñamiento”...: derivación; morfológica. Simplemente habría que indicar la forma correcta.
13. “Mandra”, “plegar”: léxicas. Indicar forma correcta.

5.2 Implementación

El prototipo de corrector gramatical, cuya implementación presentamos a continuación, responde a las necesidades previamente mencionadas, es decir, se considera que el usuario es nativo y tiene un conocimiento medio o alto tanto del español como del catalán. La corrección ortográfica y gramatical tiene en cuenta, pues, los fenómenos de interferencia, así como otros que son más independientes de la lengua que se esté tratando, tales como errores de omisión o duplicación de palabras o de relajación a la hora de escribir un texto en el que deliberadamente se omite sistemáticamente por ejemplo, todo tipo de acento o similar.

La implementación del corrector se ha realizado hasta ahora para el catalán, tanto a nivel ortográfico como gramatical. En cuanto al español, se espera que la implementación del corrector sea bastante automática dada la gran semejanza estructural entre catalán y español. Las estrategias que se han decidido para el catalán, son, pues perfectamente generalizables para el español. Los submódulos encargados de la corrección son **CATSPEL** y **CGGRAM**.

5.2.1 CATSPEL (Catalan Spelling)

El módulo CATSPEL, diseñado en C++, es el encargado de la detección y posterior corrección de aquellas palabras que no pertenecen al formario (*palabras desconocidas* en adelante). También es el responsable de la posible corrección de errores gramaticales, como se explicará más adelante.

Para toda palabra desconocida, CATSPEL busca una lista de candidatos correctos, es decir, pertenecientes al formario, que ordena según unos criterios de verosimilitud que más tarde explicaremos.

La estructura de datos asociativa de tipo clave/valor que se ha usado en la implementación ha sido un *multi-way balanced tree* o simplemente *b-tree* Las claves son las formas, mientras que los valores son cadenas de caracteres que concatenan las diferentes informaciones sintácticas, es decir, lema más información categorial y eventualmente de subcategorización.

Las búsquedas en el formario se realizan en memoria secundaria, razón por la cual es del todo necesario minimizar los accesos a disco en tiempo de ejecución. Probablemente, los *b-tree*

constituyen la mejor manera de lograrlo. Éstos son básicamente árboles m-arios balanceados, que aseguran un coste logarítmico (en función del tamaño del formario). La condición de balanceado asegura que el árbol no degenere en una lista, por lo que entonces la búsqueda tendría un coste lineal. Haciendo pruebas empíricas, se ha encontrado que la clave del *b-tree* (número de ítems por nodo) que minimiza el tiempo de búsqueda es K=12, para un formario de 400000 formas.

Otra manera, muy eficiente pero costosa en espacio, son los *tries*, básicamente árboles alfabéticos que cumplen la condición del prefijo (muy útil en la detección de fusión de palabras, como p.ej., 'otramujer'), es decir, dada una tira de caracteres w_1 prefijo de w, en la búsqueda de w, se comprueba directamente si w_1 pertenece al formario. En caso de que no pertenezca se aborta la búsqueda. Se espera que, como trabajo futuro, se implemente un trie y se compare la eficiencia con la estructura de B-tree.

Veamos a continuación las dos operaciones básicas del módulo: generación de candidatos y filtrado.

Generación de candidatos

Las operaciones básicas de edición de palabras son las de (d), (i), (s) y (t), que corresponden respectivamente a las de borrado (*deletion*), inserción (*insertion*), sustitución (*substitution*) y transposición (*transposition*). Se puede comprobar que (s) y (t) pueden ser definidas en términos de (d) e (i).

La generación de candidatos se efectúa con un número de cómo máximo tres operaciones de edición. Veamos un ejemplo:

> Rergesabaq $\rightarrow^{(t)}$ Regresabaq $\rightarrow^{(d)}$ Regresaba, palabra correcta que ha sido generada gracias a una transposición (t) y un borrado (d).

Nótese que no todas las combinaciones de edición se llevan a cabo, pues muchas de ellas conducen a formas ortográficas del todo imposibles, por lo que una búsqueda en el diccionario electrónico sería del todo inútil. Veamos un ejemplo:

> *arrosego $\rightarrow^{(i)}$ *arroszego (Cast: Arrastro, Forma correcta en Cat: arrossego)

Es decir se ha llevado a cabo una inserción de la letra z por lo se que llega a la secuencia sz que es del todo imposible en catalán. Una búsqueda en el diccionario de la palabra sería pues inútil.

Téngase en cuenta, que dado que la longitud media en caracteres de las formas del formario es de aproximadamente de diez caracteres. Si se tiene que generar candidatos para un número de cien palabras desconocidas, sólo con la operación de inserción generalizada obtenemos el siguiente número de búsquedas: 35*(10+1)*100, es decir aproximadamente 35000 búsquedas.

Filtrado y ordenación de los candidatos

El conjunto de candidatos generado debe ser filtrado. Deben ser ordenados de acuerdo a un criterio de verosimilitud. En efecto, si $\{w_1, \ldots, w_n\}$ es el conjunto de candidatos, y w es una palabra no perteneciente al formario, se calcula el w_i que minimiza $\{d(w, w_i)\}_{i=1, \ldots,n}$ donde d es un versión modificada de la distancia de edición de Levenshtein, que asigna probabilidades a cada operación de edición.

Las probabilidades se calculan a partir de un corpus de entrenamiento. Este procedimiento basado en inferencia Bayesiana, se inspira en Kernighan et al. (1990). En el proceso de ordenación o *ranking*, también se han incorporado algunas heurísticas que son sensibles al tipo de usuario, p.ej. nativo o no nativo. También se ha incorporado la heurística que tiene el contexto, en este caso, conocer las categorías mayores previas y posteriores a la palabra objetivo.

5.2.2 CGGRAM (Constraint Grammar GRAMmar checking)

El módulo de revisión gramatical CGGRAM se encarga de la detección de errores gramaticales. Se ha implementado con la ayuda del sistema de análisis superficial morfosintáctico basado en restricciones *Constraint Grammar* (Karlsson et al. 1995, Tapainen 1996).

El proceso de detección consta de las siguientes fases, todas implementadas en *Constraint Grammar*:

- Desambiguación morfológica relajada
- Detección de errores morfológicos
- Detección de errores estructurales
- Proyección y desambiguación sintácticas relajadas
- Detección de errores sintácticos

5.2.2.1 Desambiguación morfológica relajada

La desambiguación morfológica relajada es similar a la de la desambiguación usual. Mientras que la desambiguación morfológica estándar ha sido diseñada con la hipótesis de un *input* correcto, en la relajada se ha previsto la posibilidad de una entrada incorrecta. Consideremos el siguiente ejemplo. Sea una configuración de un simple sintagma nominal con la secuencia X=[$_{SN}$ [$_{Det/Pronombre}$ El] [$_{Nombre}$ taula]] ('la mesa'), en la que *El* es ambiguo entre Determinante y Pronombre y *taula* es no ambigua y recibe la lectura de Nombre. Dos desambiguaciones morfológicas son posibles, una con la secuencia Determinante-Nombre y la otra Pronombre-Nombre. Ambas secuencias son incorrectas. Sin embargo, la primera es susceptible de ser el resultado de un error de concordancia de género, susceptible de ser corregido. La segunda secuencia, por supuesto del todo incorrecta, es descartada por la desambiguación morfológica relajada, por ser una configuración en la que una posible corrección resulta más difícil de prever.

Nótese además, que si hubiéramos considerado la configuración Y=[$_{SN}$ [$_{Det/Pronombre}$ El] [$_{Nombre/Verbo}$ menja]] ('lo come') en que *menja* es ambiguo entre nombre y verbo finito, la desambiguación no descartaría ninguna lectura a pesar de que hay una desambiguación posible correcta, que es la de secuencia Pronombre-Verbo. Con esto, la desambiguación morfológica de *la taula* dejaría dos posibles análisis, la de un sintagma nominal con un conflicto de concordancia, y la de un fragmento de sintagma verbal. La razón por la que se ha tomado la decisión de conservar el análisis de un sintagma nominal con concordancia errónea, se debe a que la configuración Y, podría estar seguida de otro verbo finito, que podría resultar en un posible error gramatical. Dada la robustez exhibida por el módulo de desambiguación morfológica, la relajación ha consistido pues en pequeñas modificaciones de las reglas de dicho módulo, todas realizadas manualmente. Como trabajo futuro, se contempla la posibilidad de realizar dichas modificaciones de manera automática con algoritmos de aprendizaje automático.

5.2.2.2 Detección de errores morfológicos

La detección de errores morfológicos se realiza mediante un fichero de proyección de *Constraint Grammar*, que se aplica a un fichero que previamente ha sido desambiguado morfológicamente con el módulo de desambiguación morfológica relajada. Lo que hace este módulo de mapping es proyectar etiquetas de error a posibles configuraciones incorrectas, como p. ej. un error de concordancia . Veamos un ejemplo de regla:

```
ADD (AGR-NOM-G) TARGET (DET) IF (0 MASC) (1 NOM + FEM) ;
```

Con esta regla, se proyecta la etiqueta AGR-NOM-G, que señala un posible error de concordancia nominal de género, cuando hay un determinante en género masculino seguido de un nombre de género femenino.

En el fichero de detección de errores morfológicos se ha pretendido abarcar errores de concordancia en los complejos nominales y verbales, con toda la diversidad de casos que ello implica. Nótese que con ello se consigue detectar errores clasificados previamente como de interferencia. Por ejemplo:

*El(Det-masc) dent(Nombre-fem), (*dent* corresponde a *diente*)

cuyo error es de tipo de concordancia nominal de género y que probablemente se deba a una interferencia del español, pues en esta lengua *dent* se traduce por *diente* que es de género masculino.

5.2.2.3 Detección de errores estructurales

Como en el caso de la detección de errores morfológicos, se trata de un fichero *Constraint Grammar* de proyección que se aplica al que previamente ha pasado las etapas de desambiguación morfológica relajada así la de proyección de etiquetas de errores de naturaleza morfológica. Aquí los errores que se pretende detectar son poco uniformes y difícilmente clasificables, razón por la cual, desde el punto de vista de la implementación, se ha decidido llamarlos errores estructurales. Se abarcan errores del tipo Verbo_auxiliar-Determinante, como por ejemplo *ha el*, que podría corresponder a la configuración correcta Verbo_auxiliar-Participio-Determinante, como p.ej. *ha fet el* ('ha hecho el', por omisión).

Este tipo de reglas se han obtenido de una manera semi-automática. Para ello, hemos buscado bigramas y trigramas de etiquetas que resulten en configuraciones imposibles. A este tipo de n-gramas lo llamamos n-gramas negativos (NN). Para obtener un conjunto razonable de NN, se ha procedido a buscar bigramas y trigramas de etiquetas imposibles en el corpus del IEC (Rafel 1994). Para ello, se ha seguido la estrategia propuesta por Kveton y Oliva (2002). Para cada bigrama o trigrama negativo se ha tenido que generar reglas *ad hoc* de manera automática que proyectan etiquetas de error que hacen referencia al n-grama negativo al que correspondan. Veamos un ejemplo:

Tomemos el bigrama negativo Verbo_auxiliar-Determinante del ejemplo previo, y veamos el tipo de regla de proyección de etiquetas de error correspondiente:

```
ADD (ESTRUCT_AUXVERB_DET) TARGET (DET) IF (-1 AUX) ;
```

Si por alguna razón, la forma que sigue al verbo auxiliar es ambigua entre determinante y pronombre clítico, y la lectura de determinante se pierde, entonces la regla previa no se aplica. Tres soluciones son posibles. La primera duplicar la regla y sustituir TARGET (DET) por TARGET(PRN_CLÍTICO):

ADD (ESTRUCT_AUXVERB_DET) TARGET (PRN_CLÍTICO) IF (-1 AUX) ;

Otra manera, sería sustituir el TARGET(DET) por TARGET("<el>"), ya que entonces se está proyectando la etiqueta de error directamente sobre la forma. Sin embargo, también se debería duplicar la regla para poder capturar configuraciones en que el determinante tiene género femenino.

Finalmente, veamos otro ejemplo de error estructural que puede detectar otro tipo de error de interferencia lingüística. Se trata de los verbos de régimen que en catalán sufren la "caída" de preposición en presencia de una subordinada mientras que en español no ocurre:[7]

*Em recordo de que ... (Esp: Me acuerdo de que...)

Una regla de proyección sería:

ADD (ESTRUCT_WRONG_RV) TARGET (VERB_FINITE) IF (1 PREP) (2 VERB);

Finalmente digamos que en la categoría de errores estructurales también podemos detectar aquellos de interferencia lingüística que son de naturaleza puramente léxica como por ejemplo (en este caso, se trata de un ejemplo en español):

*Hay la tele puesta...,

que en catalán correspondería a la oración correcta: 'Hi ha la tele posada...' En este caso, tenemos el NN (o error estructural) Verbo_AUX-DET. Una regla para detectar este error interferencia sería pues:

ADD (ESTRUCT_AUXVERB_DET) TARGET ("HAVER") IF (1 DET) ;

5.2.2.4 Detección de errores sintácticos

Estos módulos de *Constraint Grammar* se encargan de proyectar etiquetas funcionales tales como *Subj*, *CD*, *Atr* que corresponden respectivamente a sujeto, complemento directo y

[7] También debe incluirse los casos de adjetivos y nombre de régimen como p.ej. *es consciente de que...* que en catalán la preposición debería caer.

atributo. Siguiendo los mismos pasos que en el fichero de desambiguación morfológica, se ha procedido a relajar el proceso de proyección y desambiguación de etiquetas funcionales. Con ello, se consigue detectar errores debidos a procesos de interferencia lingüística. Un ejemplo de ello sería escribir en catalán un complemento directo animado precedido de preposición, que no es correcto:

L'equip de Manresa va derrotar *[$_{CD}$ al de Sabadell][8]

'El equipo de Manresa derrotó al de Sabadell'

El sintagma preposicional recibe, a pesar de que no sea correcto, la etiqueta funcional de complemento directo, *CD*. En un módulo posterior se podrá detectar un posible error de complemento directo erróneo, sabiendo que se ha encontrado una lectura de *CD* en un sintagma preposicional.

Mediante un fichero de mapping de Constraint Grammar, similar a los de detección de errores morfológicos, se proyectan etiquetas de error. En el caso del ejemplo anterior, se proyectaría la etiqueta *WRONG-CD*.

5.2.3 Interacción de CATSPEL con el preproceso y CGGRAM

CATSPEL se encarga de proyectar candidatos correctos con todas sus lecturas posibles a toda palabra desconocida que se encuentre durante el preproceso. CATSPEL proyecta posibles candidatos para corregir la palabra desconocida. De esta manera, en un texto del cual se espera una revisión ortográfica y gramatical, el preproceso se realiza conjuntamente con CATSPEL a fin de que antes de entrar en los sucesivos módulos de *Constraint Grammar* previamente mencionados en este apartado, se asegura que al mayor número de palabras desconocidas se le proyecten candidatos correctos con su información morfosintáctica, es decir: lema, categoría, información de género, número y persona y eventualmente subcategorización.

Una vez que el texto pasa las etapas del preproceso y CATSPEL, se procede a aplicarle secuencialmente los módulos arriba mencionados. La Figura 10 representa la arquitectura del corrector, integrada en la de la CatCG.

[8] Versión correcta:'L'equip de Manresa va derrotar [$_{CD}$ el de Sabadell]'. El complemento directo debe ser un sintagma nominal y no un SP.

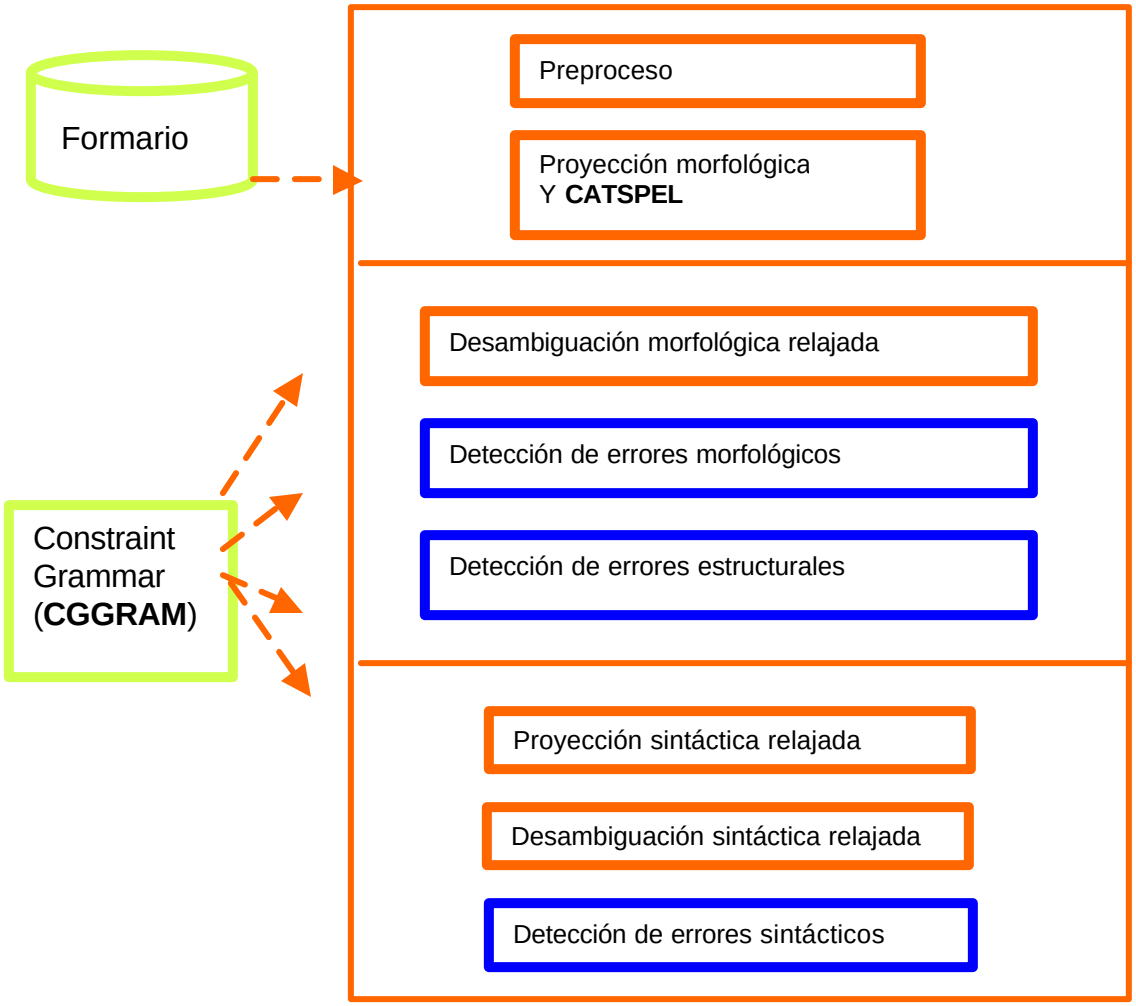

Figura 10 Arquitectura del módulo de corrección con CATSPEL y CGGRAM, integrada en la CatCG

5.2.4 Hacia la corrección gramatical

La corrección ortográfica sensible al contexto ya se realiza con **CATSPEL** Los diferentes submódulos de **CGGRAM** se encargan de la **detección** de diversos tipos de errores morfosintácticos, pero no de su corrección. Dado un texto para el cual se desea una corrección gramatical, se le debe aplicar la secuencia de módulos CGGRAM. El *output* final es un texto verticalizado con etiquetas de posibles errores morfosintácticos. Un módulo posterior diseñado en C++ está en proceso de estudio para gestionar la asignación de errores morfosintácticos. La idea es que el texto al que se le han proyectado etiquetas de errores se aproxime lo máximo posible a un texto correcto. En otras palabras, debe evitarse la sobredetección de errores.

Una vez se tiene el texto definitivo con las proyecciones de etiquetas de error, se debe dar lugar a una posible corrección gramatical. Creemos que dado el estado de la cuestión sólo se pueden proponer correcciones gramaticales en aquellos errores que sean de naturaleza de concordancia conflictiva, y de expresión errónea de una función gramatical. En el caso de los errores que conciernen a concordancia, la corrección es posible mediante búsquedas con **CATSPEL** de alternativas en el formario que estén próximas en distancia de edición y que contengan el rasgo de concordancia morfológica deseado para evitar el conflicto de concordancia. Un ejemplo: para *La noi bona* ('la chico buena'), que corresponde a la secuencia Det(fem)-Nom(masc)-Adj(fem), el uso de **CATSPEL** es requerido para buscar una entrada en el formario próxima en distancia de edición que tenga el rasgo *fem*, en este caso *noia* ('chica').

Para los llamados errores estructurales, de momento se considera que únicamente se puede mostrar un mensaje al usuario indicando que hay un posible error, pero no una sugerencia de corrección.

Finalmente, una tarea pendiente y fundamental es la de evaluar el sistema de detección de errores, es decir estimar la cobertura y precisión del sistema.

6 Trabajos complementarios sobre léxico catalán

De la discusión en el apartado 4 se desprende que es necesario ampliar tanto la cobertura como la información presente en el léxico, tanto catalán como español. En el grupo GLiCom de la Universitat Pompeu Fabra se están desarrollando una serie de herramientas para la ampliación automática o semiautomática del léxico y de la información del mismo. En este apartado describiremos brevemente tres de las herramientas que están en desarrollo actualmente: el generador morfológico (ap. 6.1), para la ampliación de la cobertura del formario y el control de la coherencia, el sistema de adquisición de marcos de subcategorización (ap. 6.2) y el sistema de clasificación automática de adjetivos (ap. 6.3).

6.1 Generador morfológico

Los formarios, tanto para el catalán como para el español, son en principio ficheros cerrados. Un problema que se ha advertido alguna vez es que algunas formas de algunos lemas no aparecen; por ejemplo, en algunos lemas verbales faltan las formas de algunos tiempos o modos. Para remediarlo, se ha diseñado un programa implementado en C++, cuya finalidad es generar las formas deseadas que no se encuentren el formario para una forma dada.

Así, dado un lema, una categoría morfológica y la etiqueta correspondiente, el programa genera las formas correspondientes.

6.2 Adquisición de marcos de subcategorización

Actualmente, se está desarrollando un sistema que pretende adquirir automáticamente información sobre la subcategorización verbal (Mayol, en preparación). Es decir, un sistema que pueda deducir qué tipo de complementos admite o requiere un verbo a partir del comportamiento de este verbo en el corpus. Esta información puede ser de gran utilidad para un analizador sintáctico como la *Constraint Grammar*, ya que podría ayudar a eliminar ambigüedades que no se pueden resolver sin información léxica. Por ejemplo, disponiendo de información de subcategorización, es mucho más fácil distinguir entre un complemento preposicional exigido por el verbo (*creer en Dios*) y un complemento circunstancial (*comprar en el mercado*).

Para llevar a cabo la adquisición automática, el algoritmo de aprendizaje necesita información sobre los verbos. Así, se deben definir diferentes rasgos que puedan ser relevantes para clasificar distintos tipos de verbos y que se puedan extraer automáticamente de un corpus. Por ejemplo, el porcentaje de las veces que un verbo aparece en el corpus seguido o precedido de un pronombre clítico de complemento directo será mucho mayor en los verbos transitivos que en los intransitivos. En cambio, el porcentaje de veces que un verbo va seguido de signo de puntuación debería ser mayor en los verbos intransitivos que en los transitivos. Así, imaginemos que tenemos dos verbos definidos por los siguientes datos (el primer porcentaje se refiere a los pronombres clíticos de complemento directo y el segundo a la puntuación):

- 45% 12%
- 14% 30%

El primer verbo será seguramente un verbo transitivo y el segundo un verbo intransitivo. Un algoritmo de aprendizaje automático hace está misma operación de generalización, pero teniendo en cuenta más rasgos y analizando un número de objetos, de verbos, mucho mayor.

El primer objetivo de este trabajo ha sido distinguir entre verbos transitivos y intransitivos y, hasta el momento, se han hecho experimentos con dos métodos distintos de aprendizaje automático: método supervisado (inductor de reglas) y método no supervisado. El método supervisado parte de un conjunto de verbo previamente clasificado. A partir de los ejemplos, el algoritmo aprende y puede extraer reglas para predecir la clase de verbos diferentes a los que

ha utilizado para “aprender”. El método no supervisado, *clustering*, no parte de ninguna clasificación previa, sino que, en este caso, el algoritmo analiza los datos e intenta encontrar generalizaciones en los datos. Se dividen los verbos en grupos, en *clusters*, de manera que los verbos del mismo grupo sean lo más parecido posible entre ellos y lo más distinto posible de los verbos de los otros grupos.

6.3 Clasificación automática de adjetivos

Otro de los sistemas de adquisición de información léxica que se está desarrollando en nuestro grupo es el de clasificación automática de adjetivos (Boleda y Alonso 2003). El objetivo es llegar a distinguir clases *semánticas* de adjetivos (en función de si denotan atributos o relaciones con objetos o eventos) a partir de evidencia morfosintáctica o distribucional.

Para los experimentos, utilizamos un fragmento del CTILC, corpus del Institut d’Estudis Catalans (Rafel, 1994), etiquetado con la CatCG. Modelamos los adjetivos a partir de su contexto morfológico (comprobando p.ej. si el adjetivo va precedido de un adverbio de gradación, o de un verbo copulativo), con la idea de que las distintas clases semánticas de adjetivos tendrán comportamientos sintácticos diferenciados.

Para hacer la clasificación, elegimos un método no supervisado, el *clustering*, ya que no disponíamos de una clasificación previa fiable que nos permitiera usar métodos supervisados. Tal y como se ha explicado en el apartado anterior, en este método se modela cada objeto como un vector cuyos componentes corresponden a rasgos superficiales (p.ej., porcentaje de ocurrencia tras un verbo copulativo). Los distintos algoritmos de *clustering* se basan en el cómputo de las distancias entre vectores, de forma que aquellos que sean más cercanos según el criterio elegido acabarán en el mismo grupo o *cluster*.

Se procedió a elegir 100 adjetivos al azar para comparar los resultados del proceso con una clasificación por parte de tres anotadores humanos. Los anotadores recibieron instrucciones de clasificación que referían sólo a las características semánticas de los adjetivos, no a las distribucionales. Los resultados actuales indican un acuerdo medio del 90% entre la clasificación humana y la automática, es decir, indican que la tarea es factible y se puede efectivamente aplicar para la ampliación automática de la información léxica, proceso que se está llevando a cabo actualmente.

7 Utilización de las herramientas en otros proyectos

7.1 BancTrad

BancTrad es un proyecto de innovación docente financiado por la Universitat Pompeu Fabra en el marco de las subvenciones que esta Universidad destina a proyectos de innovación docente. Este proyecto se inició en septiembre de 2000 y actualmente está en fase de alimentación.

BancTrad es una plataforma de consulta de corpus (mono o multilingües) anotados vía Internet que tiene como destinatarios tanto docentes o estudiantes de traducción o lingüística como traductores profesionales, investigadores etc. La compilación de un corpus paralelo multilingüe es parte inherente de los objetivos de BancTrad. Las lenguas con las cuales trabajamos son: catalán, español, inglés, alemán y francés. Es posible realizar consultas de cualquiera de estas lenguas al español o al catalán y viceversa (pero no entre francés, alemán e inglés). Actualmente este corpus contiene un total aproximado de 1.700.000 palabras y se alimenta constantemente de fuentes diversas: traducciones realizadas en el marco de nuestra facultad, editoriales, Internet etc.

En primer lugar los textos son procesados con el fin de ser anotados extralingüísticamente y alineados con sus respectivas traducciones. Este proceso es semiautomático en el sentido de que el resultado de la alineación precisa de una revisión humana. La anotación extralingüística contempla los siguientes parámetros: nombre de la persona que introduce el texto, lenguas de origen y de llegada, referencias de original y traducción, fechas de publicación y traducción, registro, tipo de texto, ámbito temático, grado de especialización y grado de dificultad de la traducción el texto (como proceso). A partir de aquí los textos son introducidos en el servidor para ser etiquetados lingüísticamente y formateados como corpus consultables (véase Badia et al. 2002). Para esta parte del proceso, que es completamente automático, se precisan analizadores lingüísticos como los descritos en los apartados referidos a CatCG y CastCG y es en este sentido que BancTrad se distingue claramente de otros corpus paralelos (mayoritariamente no anotados) gracias en gran parte a las herramientas desarrolladas a lo largo de PrADo. Por lo que respecta a la anotación, cada lengua sigue un proceso diferente: los textos en catalán se analizan con la CatCG y los españoles serán analizados en un futuro muy próximo con la CastCG. Por otro lado, el análisis lingüístico los textos en inglés, alemán y francés se realiza mediante un *TreeTager*, un analizador superficial desarrollado en el IMS (*Institut für Maschinelle Sprachverarbeitung*, Stuttgart).

Es importante notar que, a pesar del uso de herramientas diferentes para la explotación de la información lingüística de nuestros textos, todas las lenguas reciben un mínimo uniforme de información: lema y etiqueta morfológica (de momento la función sintáctica solo para el catalán). De esta forma, todas las lenguas son consultables de la misma forma independientemente del etiquetador utilizado con cual se contribuye a la modularidad, ya que el procesamiento lingüístico de una cierta lengua se puede modificar sin cambiar ninguno de los otros procesos lingüísticos ni la interfaz.

Como ya hemos apuntado anteriormente, el mérito más importante de BancTrad respecto a otros corpus paralelos reside en el hecho de ser un corpus anotado. Es evidente que esto dota a este corpus paralelo de una flexibilidad y de una posibilidades de aplicación extremadamente ventajosas especialmente en el contexto de los estudios en traducción y lingüística, pero también representa un punto de partida con un alto potencial con vistas a futuros proyectos de base multilingüe.

7.2 ALLES

El proyecto ALLES (*Advanced Long-distance Language Education System*) es un proyecto RTD financiado por la Comisión Europea en el marco del Programa de IST (IST-2001-342461). La duración del proyecto es de 36 meses, de junio de 2002 a mayo de 2005, y los miembros del consorcio son: ATOS Origin Spain, el *Institute of Applied Information Sciences* de Saarbrücken, la Fundació Universitat Pompeu Fabra de Barcelona, la Universidad Europea de Madrid y la Heriot-Watt University de Edimburgo.

El proyecto aspira a crear un sistema multimedia para el autoaprendizaje de lenguas para estudiantes con un nivel avanzado y en el ámbito de la empresa y la economía (lenguaje para fines específicos). Las lenguas de trabajo son: inglés, alemán, español y catalán. Asimismo, el objetivo primordial de ALLES es demostrar que el uso de herramientas de procesamiento del lenguaje natural posibilita la capacidad de proporcionar al aprendiz comentarios acerca de cómo se desarrolla su progreso. Asimismo, como resultado de este análisis *inteligente* de los errores del aprendiz se logra, a la larga, un margen de maniobra que permite adaptar la secuencia didáctica a las necesidades del alumno.

Justamente en este último aspecto, el del procesamiento del lenguaje natural para la detección de errores, es donde ALLES y PrADo tienen mucho en común. En este sentido, la base de las arquitecturas de detección de errores usadas en ALLES es en gran parte fruto de las herramientas desarrolladas a lo largo de PrADo. Y lo son en dos sentidos, ya que el tipo de

corrección que se prevé dentro de ALLES tiene dos estrategias claramente diferenciadas: la corrección del producto lingüístico y la adecuación del mismo a la situación comunicativa.

- Herramientas para la **corrección y evaluación del producto lingüístico**: mediante este tipo de herramientas se pretende evaluar la corrección morfológica, sintáctica, semántica... en definitiva que el producto lingüístico del estudiante sea aceptable desde el punto de vista estricto de la norma lingüística.

Para esto se usa una arquitectura como la presentada en el apartado referido a la corrección de errores (5.2). Se trata de un sistema preparado para el análisis de oraciones mal formadas, es decir, suficientemente robusto para poder predecir qué estructuras incorrectas se están usando (en lugar de otras estructuras correctas).

- Herramientas para la **evaluación de la adecuación a la situación comunicativa**: mediante este tipo de herramientas se pretende evaluar la adecuación del producto del aprendiz en relación con el tipo de texto, la situación comunicativa, etc. Se puede entender como una especie de corrector de estilo en sentido amplio, porque no sólo se centra en lo que tradicionalmente se entiende por estilo, sino también en la adecuación pragmática del texto. Por ejemplo, si un estudiante utiliza la forma de tratamiento *tú* (en lugar de *usted*) para dirigirse a un posible contratador (en una carta de presentación para una solicitud de un puesto de trabajo), no está cometiendo ningún error lingüístico. Sin embargo, el receptor de la carta puede percibirlo como una falta de respeto, como un exceso de confianza, como una falta de conocimiento de las normas sociales en definitiva. Por ello, es importante que los estudiantes no sólo escriban correctamente, sino también adecuadamente.

Para este tipo de evaluación se precisan analizadores lingüísticos como los descritos en los apartados referidos a CatCG y CastCG. Se trata de herramientas que esperan un *input*, texto de entrada, que (idealmente) no presenta errores de norma lingüística. El resultado debe ser un texto analizado con información relativa a varios niveles de descripción lingüística: morfología, sintaxis, semántica y pragmática. Según el tipo de texto y ejercicio particular (determinado por el enunciado y la situación comunicativa ficticia) se dará prioridad a la presencia o ausencia de determinadas construcciones lingüísticas.

7.3 eTitle

eTitle es un proyecto europeo enmarcado en *eContent* que tiene como objetivo el subtitulado automático de material audiovisual. Para ello, se pretende integrar las últimas tecnologías en los ámbitos del reconocimiento de voz, la compresión de textos y la traducción automática. Las lenguas sobre las que se trabaja son el catalán, el español, el inglés y el checo.

El diseño del sistema es modular, de forma que la aplicación de los diferentes módulos, p.ej. traducción automática o compresión, sea opcional y adaptada a las necesidades del usuario.

En la compresión de textos se ofrece al usuario la posibilidad de escoger entre tres niveles de compresión: 1. simple edición del texto (con normalización de abreviaturas, números, fechas, ...); 2. compresión básica (sustitución de ciertos elementos por sus correspondientes paráfrasis léxicas o sintácticas); y 3. compresión avanzada (omisión o fusión de ciertos elementos).

Las herramientas desarrolladas en el marco del proyecto Prado y presentadas en este documento son de indudable utilidad para las necesidades de procesamiento lingüístico que requiere eTitle.

En primer lugar, las operaciones propias del preproceso coinciden con las funcionalidades que ofrece el nivel 1 de la compresión de texto. En el caso de la traducción automática, este mismo preproceso será también útil a la hora de preparar el texto antes de la traducción.

En segundo lugar, para implementar las funcionalidades de los niveles 2 y 3 de compresión, serán necesarios con toda probabilidad el procesamiento morfológico y sintáctico del texto. De la misma manera, para explotar de manera más eficiente los recursos de traducción basada en ejemplos, ya sea en forma de memoria de traducción (MT) o traducción automática, el procesamiento morfológico (y tal vez el sintáctico) es también imprescindible.

8 Resultados: trabajos de investigación y publicaciones

El proyecto PrADo ha constituido un marco de investigación y de formación en la investigación para los dos grupos. En este sentido podemos considerar como resultados del proyecto tanto los trabajos de investigación surgidos en su entorno, como las publicaciones resultantes de la investigación llevada a cabo. En ambos casos, estos resultados se han producido a menudo en conexión con las investigaciones de otros proyectos o con investigaciones personales de los miembros de los equipos.

A continuación, pues, listamos:

- los trabajos de investigación realizados en los grupos (leídos o en curso) y que tienen una conexión con la investigación de PrADo, y
- las comunicaciones y publicaciones surgidas directa o indirectamente de las investigaciones de PrADo.

Finalmente, indicamos los sitios web en nuestras universidades en los que se pueden ver los resultados prácticos implementados del proyecto.

Trabajos de investigación

Boleda, M. (2003) *Adquisició de classes adjectivals.* Trabajo de investigación, Doctorado en Ciencia Cognitiva y Lenguaje, Universitat Pompeu Fabra.

Bott, S. (en preparación) *The detection of Catalan web text on the basis of a word-form dictionary*. Trabajo de investigación de línea de Lingüística Computacional del doctorado en Lingüística Aplicada, Universitat Pompeu Fabra.

Gil, À. (en preparación) *El tractament de les oracions copulatives a la CatCG*. Trabajo de investigación de línea de Lingüística Computacional del doctorado en Lingüística Aplicada, Universitat Pompeu Fabra.

Gil, À. (en preparación) *Proposta d'un analitzador sintàctic per al català oral*. Proyecto de tesis del doctorado en Lingüística Aplicada, Universitat Pompeu Fabra.

Mayol, L. (en preparación) *Aprenentatge automàtic de marcs de subcategorització*. Trabajo de investigación, Doctorado en Ciencia Cognitiva y Lenguaje, Universitat Pompeu Fabra.

Quixal, M. (2003) *Theoretical basis and implementation of a linguistic-based morphosyntactic tagger for Catalan.* Trabajo de investigación, Doctorado en Ciencia Cognitiva y Lenguaje, Universitat Pompeu Fabra.

Publicaciones y comunicaciones

Aguilar, L., A-B. Avilés, J. Fontseca, C. de la Mota, Y. Rodríguez, P. Caymes, S Balari (2003) Un módulo de desambiguación morfosintáctica para el castellano basado en conocimiento lingüístico. *VI Congreso de Lingüística General*, Santiago de Compostela.

Alsina, À, T. Badia, G. Boleda, S. Bott, À. Gil, M. Quixal, O. Valentín (2002) CATCG: a general purpose parsing tool applied. En *Proceedings of the Third European Conference on Language Resources and Evaluation*, Las Palmas, mayo.

Alsina, À., T. Badia, G. Boleda, S. Bott, Á. Gil, M. Quixal, O. Valentín (2002) CATCG: un sistema de análisis morfosintáctico para el catalán. En *Actas del Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*. Valladolid, setiembre.

Ana Belén Avilés y Jordi Fontseca (2003) DeMoNiO. Módulo de Desambiguación Morfosintáctica para el Nivel Oracional. Comunicación en el *XXXIII Simposio de la Sociedad Española de Lingüística*, diciembre.

Badia, T., G. Boleda, M. Quixal, E. Bofias (2001) A modular architecture for the processing of free text. En *Proceedings of the Workshop on 'Modular Programming applied to Natural Language Processing'* en EUROLAN 2001, Iasi, Rumanía.

Badia, T., R. Saurí (2001) A note on redescription predicates. En P. Bouillon y K. Kanzaki (eds.) *Proceedings of the First International Workshop on "Generative Approaches to the Lexicon"* (GL'2001). School of Translation and Interpretation, University of Geneva, abril.

Badia, T., G. Boleda, C. Colominas, M. Garmendia, A. González, M. Quixal (2002) Eines de lingüística computacional per a la traducció: corpus paral·lels anotats. *2nd International Conference on Specialized Translation*. Barcelona, febrero-marzo.

Badia, T., G. Boleda, C. Colominas, M. Garmendia, A. González, M. Quixal (2002) BancTrad: a web interface for integrated access to parallel annotated corpora. En *Proceedings of the Workshop on Language Resources for Translation Work and Research held during the 3rd LREC Conference*, Las Palmas, mayo.

Badia, T., G. Boleda, J. Brumme, C. Colominas, M. Garmendia, M. Quixal (2002) BancTrad: un banco de corpus anotados con interficie web. En *Actas del Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Valladolid, setiembre.

Badia, T. y À. Gil, M. Quixal, O. Valentín (pendiente de publicación) NLP-enhanced error checking for Catalan unrestricted text. En *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisboa, Portugal, mayo 2004

Badia, T. y L. Díaz, M. Quixal, A. Ruggia, S. Garnier, P. Schmidt (enviado) Towards intelligent interactivity in distance language learning. *IEEE International Conferences on Advanced Learning Technologies*. Joensuu, Finlandia, agosto-septiembre 2004.

McNally, L., C. Kennedy (2001) Degree vs. Manner well: A Case Study in Selective Binding. En P. Bouillon y K. Kanzaki (eds.) *Proceedings of the First International Workshop on Generative Approaches to the Lexicon* (GL'2001). Reeditado 2002 en M.J. Arche, et al., (eds.) *Cuadernos de Lingüística IX: In Memoriam Ken Hale*, Instituto Universitario Ortega y Gasset, Madrid, 159-166.

Quixal, M., À. Alsina, T. Badia (2003) Criterios para definir las categorías gramaticales necesarias para explicar la estructura del sintagma nominal en catalán. Comunicación en el *XXXIII Simposio de la Sociedad Española de Lingüística*, diciembre.

Schmidt, P., T. Badia, L. Díaz, S. Garnier, J. Fernández, M. Quixal, C. Rico, A. Ruggia, E. Torrejón (pendiente de publicación) 'Controlled language tools' and 'information extraction tools' for CALL Applications. En *Proceedings of the InSTILL conference*, Venecia, Italia, julio 2004.

Schmidt, P., T. Badia, L. Díaz, S. Garnier, J. Fernández, M. Quixal, C. Rico, A. Ruggia, E. Torrejón (pendiente de publicación) Advanced Long-distance Language Education System (ALLES): Integrating Language Resources in ICALL. En *Proceedings of the 4th International Conference on Language Resources and* Evaluation, Lisboa, Portugal, mayo 2004.

Sitios *web*

Del proyecto PrADo en la Universitat Autònoma de Barcelona: http://prado.uab.es/

- *demo* de DeMoNiO (Desambiguador Morfosintáctico para el Nivel Oracional): http://prado.uab.es/Castellano/demo.html
- buscador de concordancias a partir de categorías morfológicas:

http://prado.uab.es/Castellano/webmorfoconcordancer.html

- otras herramientas: http://prado.uab.es/Castellano/aplicaciones.html

Del grupo GLiCom de la Universitat Pompeu Fabra: http://mutis.upf.es/glicom/

- información sobre PrADo: http://mutis.upf.es/glicom/project2.htm#proj_prado
- *demo* de CatCG: http://mutis.upf.es/cgi-bin/catcg/demo.pl

9 Conclusiones. Posibilidades de las gramáticas de bajo vs. alto nivel

La primera y principal de las conclusiones apunta, en nuestra opinión, hacia una valoración altamente positiva de los resultados del proyecto, independientemente de la medida en que estos se desvíen de sus objetivos originales. Esta valoración se justifica, por una parte, por el hecho de que, como se apunta en el apartado correspondiente de este informe final, el desarrollo de una herramienta de corrección gramatical que opere sobre la salida de una gramática de marcaje de bajo nivel es una empresa razonable y factible; un objetivo alcanzable en un período de tiempo no excesivamente largo, si se dispone de los recursos económicos y de personal adecuados. Por otra parte, el énfasis puesto en el desarrollo de las gramáticas de marcaje, en su cobertura y fiabilidad, así como en el diseño de un lenguaje de salida estándar (SGML en nuestro caso), se ha mostrado como una estrategia muy positiva, habida cuenta de que, gracias a ello, se ha visto que las gramáticas podían funcionar como base para el desarrollo de otras herramientas además de un corrector gramatical. Las gramáticas de marcaje son, por tanto, un recurso con un alto valor añadido que abre la posibilidad de su utilización en otros proyectos de investigación y desarrollo ajenos al proyecto PrADo.

En relación con este punto, se puede también extraer una conclusión de orden teórico interesante. Efectivamente, nuestros resultados dan cuerpo a la idea de que es posible recorrer mucho camino en el tratamiento automático de la lengua exclusivamente dentro del marco de las llamadas técnicas de bajo nivel, con sistemas cuya capacidad computacional no excede la de los autómatas de estados finitos. Este es un resultado importante, ya que la posibilidad de expresar conocimiento lingüístico mediante gramáticas regulares abre el paso al desarrollo de sistemas de procesamiento del lenguaje eficientes y robustos que, como señalábamos en el párrafo anterior, pueden constituir la base de sofisticadas aplicaciones y herramientas que operen sobre texto irrestricto. Hasta ahora, nuestra experiencia se ha limitado exclusivamente al tratamiento de la morfología y la sintaxis, pero el presente estado de cosas invita a seguir explorando los límites de los mencionados sistemas, tanto dentro del ámbito del componente morfosintáctico como dentro del ámbito del componente semántico. En cuanto al segundo, ya algunas investigaciones realizadas por otros grupos en Europa y en los EE.UU., sugieren que puede ser posible conseguir un tratamiento satisfactorio de ciertos fenómenos también en este nivel de análisis, lo cual, de ser así, ampliaría aún más el abanico de posibles aplicaciones de los sistemas de procesamiento de bajo nivel. No olvidemos que uno de los objetivos del

proyecto PrADo era, precisamente, explorar técnicas y herramientas que permitieran, en un futuro, desarrollar correctores gramaticales más complejos con capacidad, incluso, para la corrección de estilo. En la medida que todo ello pueda mantenerse dentro de la capacidad computacional limitada de los sistemas de estados finitos, ello será siempre garantía de eficiencia y robustez.

Aun así, somos conscientes de que, con toda probabilidad, un subconjunto, posiblemente muy reducido, de fenómenos en lenguaje natural escaparán a la capacidad de cómputo de las gramáticas regulares. Sin embargo, un conocimiento preciso de estas limitaciones es lo que precisamente facilitará la aplicación de técnicas mixtas para el desarrollo de sistemas más sofisticados, tanto la inclusión de conocimiento estadístico como la utilización de gramáticas de alto nivel. Debemos recordar también que esta era una hipótesis que se apuntaba en la definición del proyecto PrADo, que consideraba la posibilidad de combinar regímenes de procesamiento de tal modo que todo aquello que escapara al poder de análisis del sistema de bajo nivel se tratara con un sistema de alto nivel que operara sobre la salida del primero. Si, como sugieren nuestros resultados y los de otros grupos que han trabajado en una línea similar, buena parte del trabajo puede quedar como responsabilidad del sistema de bajo nivel y dentro, además, de unos límites conocidos, la intervención del sistema de alto nivel, más fiable, pero también menos eficiente y robusto, podrá hacerse de una manera controlada que evitaría la pérdida en eficiencia y robustez.

Sospechamos que, a medio y corto plazo, el desarrollo de arquitecturas mixtas de procesamiento será el paso inevitable que deberá dar cualquiera que desee construir sistemas en lenguaje natural de amplia aplicabilidad (i.e., que no operen en universos del discurso muy limitados, por ejemplo), ya que los avances tecnológicos y formales no parecen indicar que la utilización de sistemas de alto nivel sea, por el momento, una alternativa viable, ni en términos de eficiencia, ni de robustez ni de fiabilidad. Por otra parte, los sistemas de bajo nivel siempre tendrán limitaciones que, por pequeñas que éstas sean, en algún momento deberemos intentar superar. No olvidemos que uno de los objetivos a largo plazo de las tecnologías del leguaje y de la comunicación es ‘naturalizar’ todo lo posible las interacciones humano-máquina, lo cual comporta modelar situaciones comunicativas cada vez más complejas en entornos multimodales y con la posible participación de más de dos agentes. Ante esta perspectiva, es importante disponer de sistemas robustos y eficientes, desde luego, pero también sistemas flexibles y capaces de tratar fenómenos de una gran complejidad y, a nuestro juicio, es este un objetivo que sólo alcanzaremos con la combinación inteligente de diferentes tecnologías.

10 Bibliografía

Alsina, À., T. Badia, G. Boleda, S. Bott, À. Gil, M. Quixal, O. Valentín (2002) CATCG: a general purpose parsing tool applied. En *Proceedings of the Third European Conference on Language Resources and Evaluation*, Las Palmas, mayo.

Badia i Margarit, A. M. (1995) *Gramàtica de la llengua catalana descriptiva, normativa, diatòpica, diastràtica*. Barcelona, Proa.

Badia, T., G. Boleda, M. Quixal y E. Bofias (2001) A modular architecture for the processing of free text. En *Proceedings of the Workshop on 'Modular Programming Applied to Natural Language Processing'* at EUROLAN 2001. Iasi, August 2001.

Boleda, M. (2003) *Adquisició de classes adjectivals.* Trabajo de investigación, Doctorado en Ciencia Cognitiva y Lenguaje, Universitat Pompeu Fabra.

Boleda, G. y L. Alonso (2003) Clustering Adjectives for Class Acquisition. En *Proceedings of the EACL'03 Student Session*, 9-16, Budapest, Hungría.

Bosque, I. y V. Demonte, dirs. (2000) *Gramática descriptiva de la lengua española.* Madrid, RAE-Espasa.

DIEC (1996) *Diccionari de la Llengua Catalana.* Institut d'Estudis Catalans.

DLC (1995) *Diccionari de la Llengua Catalana.* Enciclopèdia Catalana.

DRAE (1992) *Diccionario de la lengua española*. 21ª ed. Real Academia Española, Madrid, Espasa-Calpe.

García Yebra, V. (1982) *Teoría y práctica de la traducción*. Madrid: Gredos.

Gómez Torrego, L. (1995) *El léxico en el español actual: uso y norma*. Madrid, ArcoLibros.

Hernández García, C. (1998) Una propuesta de clasificación de la interferencia lingüística a partir de dos lenguas de contacto: el catalán y el español. *Hesperia* 1 61-79.

Karlsson, F., et. al. (1995) *Constraint Grammar: A Language-Independent Formalism for Parsing Unrestricted Text.* Mouton de Gruyter: Berlin/New York.

Kernighan, M. D., K. W. Church, W. A. Gale (1990) A spelling correction program based on a noisy channel model. En *Proceedings of COLING-90,* Helsinki.

Kveton, Pavel and Karel Oliva (2002) Detection of errors in Part-of-speech tagged corpora by bootstrapping generalized Negative n-grams. En *Proceedings of the Workshop on Linguistic Knowledge Acquisition and Representation held during the 3rd LREC Conference*, Las Palmas, junio.

Mayol, L. (en preparación) *Aprenentatge automàtic de marcs de subcategorització*. Trabajo de investigación, Doctorado en Ciencia Cognitiva y Lenguaje, Universitat Pompeu Fabra.

Prado, M. (2001) *Diccionario de falsos amigos inglés-español*. Madrid, Gredos.

Quixal, M. (2003) *Theoretical basis and implementation of a linguistic-based morphosyntactic tagger for Catalan.* Trabajo de investigación, Doctorado en Ciencia Cognitiva y Lenguaje, Universitat Pompeu Fabra.

Rafel, J. (1994) Un corpus general de referència de la llengua catalana *Caplletra*, 17.

Ruaix i Vinyet, J. (1996): *Diccionari auxiliar.* Barcelona/Moià.

Sebastián, N., M.A. Martí, M.F. Carreiras y F. Cuetos (2000) *LEXESP: Léxico informatizado del Español*. Edicions de la Universitat de Barcelona.

Solà, J. (1989) *Qüestions controvertides de sintaxi catalana.* 2a ed. Barcelona, Edicions 62.

Solà, J. (1995) *Llibre d'estil de l'Ajuntament de Barcelona*. Amb la col·laboració de Xavier Fargas, Anna Gudiol i Alba Fraser, Barcelona, Consorci per a la Normalització Lingüística.

Solà, J., et al. (dirs.) (2002) *Gramàtica del català contemporani*. Barcelona: Empúries.

Tapanainen, P. (1996) *The Constraint Grammar ParserCG-2.* Department of General Linguistics, University of Helsinki, Helsinki. Publications, 27.

Thompson, J. (1989) *Los timadores*. Júcar, Etiqueta Negra. Traducción de M. A. F. Álvarez-Nava.

Tuells, T. (1998) Constructing and Updating the Lexicon of a Two-Level Morphological Analyzer from a Machine-Readable Dictionary. En *Proceedings of the First InternationalConference on Language Resources & Evaluation*, Granada.

A Etiquetario catalán

Categorías principales

C conjunction	L locution, idiom
D Adverb.	N noun
E determiner	P preposition
H adjective / past participle	R pronoun
I interjection	V verb
J adjective	W non-linguistic item

Atributos pertenecientes a varias categorías

Person	**Number**	**Gender**
1 first person	S singular	M masculine
2 second person	P plural	F feminine
3 third person	6 underspecified (plural or singular)	6 underspecified (feminine or masculine)
6 underspecified (1st or 3rd person)		

Atributos pertenecientes a alguna categoría en concreto

Verbs	**Nouns**
Non-finite forms	*Subclasses of nouns*
C past participle	4 proper noun
G present participle	5 common noun
I infinitivo	**Determiners**
Mode	*Subclasses of determiners*
7 indicative/subjunctive/ imperative	6 cardinal/indefinite
8 indicative/imperative	0 non-head
9 imperative/imperative	A article
D indicative	C cardinal
J imperative	D demonstrative
R imperative	N indefinite
Tense	P possessive
R present	Q wh-word
A imperfect	T quantitative
P perfect	
U future	
C conditional	

Subclasses of pronouns	**Personal pronouns**
6 passive / pronominal-impersonal / personal, weak, unspecified case/number/ gender, third person	E weak personal pronouns
D demonstrative	O strong personal pronouns
E personal	Case in weak pronouns
F quantitative	6 accusative/dative
N indefinite	7 accusative / oblique
R relative	8 dative/oblique
	9 nominative/nominative
	B oblique
	C accusative

Adjectives	**Adverbs**
Subclasses of adjectives	4 documented adverb
O ordinal	5 documented derivational adverb
Q qualitative	6 non-documented adverb
P possessive	
R relative	
Conjunctions	**Locutions**
C coordinative	C conjunctive
S subordinate	D adverbial
	P prepositive

B Detalle de errores y ambigüedades restantes en la sintaxis del catalán

ADJETIVO

RESUMEN

Aciertos	1475	92,0%		
Errores	43	2,7%	Precisión	0,97
Ambigüedades	60	3,7%	Cobertura	0,96
Errores morfológicos	25	1,6%	F-score(alfa=0.5)	0,97
total	*1603*			

ACIERTOS

+<CN/<CN	1017	68,9%
+CN>/CN>	379	25,7%
+ATR/ATR	63	4,3%
+PRED/PRED	16	1,1%
Total	*1475*	

AMBIGÜEDADES			**ERRORES**		
<CN_CN>/+CN>	34	56,7%	<CN/+?	13	30,2%
<CN_CN>/+<CN	10	16,7%	<CN/+PRED	8	18,6%
<CN_PRED/+<CN	6	10,0%	<CN/+CN>	6	14,0%
ATR_PRED/+ATR	4	6,7%	<CN/+ADVL	4	9,3%
<CN_ATR_PRED/+<CN	2	3,3%	CN>/+<CN	4	9,3%
<CN_ATR/+<CN	2	3,3%	<CN/+ATR	3	7,0%
<CN_ATR/+ATR	1	1,7%	CN>/+PRED	2	4,7%
ATR_CN>/+ATR	1	1,7%	PRED/+?	1	2,3%
Total	*60*		CN>/+ATR	1	2,3%
			ATR/+<CN	1	2,3%
			total	*43*	

ADVERBIO

RESUMEN

Aciertos	975	96,7%		
Errores	21	2,1%	Precisión	0,98
Ambigüedades	2	0,2%	Cobertura	1,00
Errores morfológicos	10	1,0%	F-score(alfa=0.5)	0,99
total	*1008*			

ACIERTOS

+ADVL/ADVL	791	81,1%
+CA>/CA>	174	17,8%
+<P/<P	10	1,0%
total	*975*	

AMBIGÜEDADES

ADVL_CA>/+ADVL	1
ADVL_CA>/+CA>	1
total	2

ERRORES

ADVL/+?	6	28,6%
CA>/+ADVL	5	23,8%
ADVL/+CA>	3	14,3%
<P/+ADVL	2	9,5%
CA>/+?	2	9,5%
/+ADVL	1	4,8%
ADVL/+<P	1	
ADVL/+	1	
total	*21*	

CONJUNCIÓN

RESUMEN

Aciertos	1105	90,2%		
Errores	9	0,7%	Precisión	0,95
Errores2	51	4,2%	Cobertura	0,95
Ambigüedades	56	4,6%	F-score(alfa=0.5)	0,95
Errores morfológicos	4	0,3%		
total	*1225*			

ACIERTOS

+CONJ/CONJ	963	87,1%
+ADVL/ADVL	70	6,3%
+CD/CD	43	3,9%
+<P/<P	16	1,4%
+ATR/ATR	13	1,2%
total	*1105*	

AMBIGÜEDADES			**ERRORES**		
<CN_ADVL_SUBJ/+ADVL	31	55,4%	<CN_ADVL_SUBJ/+CD	21	35,0%
<CN_ADVL_CD_SUBJ/+CD	6	10,7%	<CN_ADVL_SUBJ/+?	12	20,0%
<CN_ADVL_ATR_SUBJ/+ADVL	5	8,9%	<CN_ADVL/+CD	4	6,7%
<CN_ADVL_CD/+CD	3	5,4%	<CN_ADVL_ATR_SUBJ/+CD	4	6,7%
<CN_ADVL_ATR_SUBJ/+SUBJ	3	5,4%	ATR/+SUBJ	3	5,0%
Otros	8	14,3%	<CN_ADVL_SUBJ/+<P	16	26,7%
Total	*56*		*total*	*60*	

DETERMINANTE

RESUMEN

Aciertos	1467	90,5%		
Errores	46	2,8%	Precisión	0,96
Errores2	11	0,7%	Cobertura	0,95
Ambigüedades	84	5,2%	F-score(alfa=0.5)	0,95
Errores morfológicos	13	0,8%		
total	*1621*			

ACIERTOS

+DN>/DN>	1408	96,0%
+<P/<P	50	3,4%
+CD/CD	4	0,3%
+SUBJ/SUBJ	4	0,3%
+ATR/ATR	1	0,1%
total	*1467*	

AMBIGÜEDADES			**ERRORES**		
CD_DN>_SUBJ/+DN>	31	33,7%	<P/+DN>	15	26,3%
CD_DN>/+DN>	20	21,7%	DN>/+<P	7	12,3%
CD_SUBJ/+SUBJ	8	8,7%	CD/+?	5	8,8%
DN>_SUBJ/+DN>	8	8,7%	DN>/+?	4	7,0%
CD_SUBJ/+CD	4	4,3%	CD_SUBJ/+?	4	7,0%
ATR_SUBJ/+SUBJ	4	4,3%	CD/+DN>	4	7,0%
otros	17	18,5%	otros	18	31,6%
total	*92*		*total*	*57*	

NOMBRE

RESUMEN

Aciertos	1437	59,2%		
Errores	216	8,9%	Precisión	0,74
Errores2	284	11,7%	Cobertura	0,76
Ambigüedades	446	18,4%	F-score(alfa=0.5)	0,75
Errores morfológicos	46	1,9%		
total	*2429*			

ACIERTOS

+<P/<P	1149	80,0%
+CD/CD	127	8,8%
+SUBJ/SUBJ	88	6,1%
+<NN/<NN	39	2,7%
+ATR/ATR	27	1,9%
+AP/AP	4	0,3%
+ADVL/ADVL	3	0,2%
total	*1437*	

AMBIGÜEDADES			**ERRORES**		
CD_SUBJ/+SUBJ	112	25,1%	CD_SUBJ/+?	98	19,6%
CD_SUBJ/+CD	106	23,8%	AP_CD_SUBJ/+?	46	9,2%
ATR_SUBJ/+SUBJ	33	7,4%	CD/+?	28	5,6%
AP_CD_SUBJ/+SUBJ	32	7,2%	<NN/+?	25	5,0%
AP_CD/+AP	22	4,9%	AP_CD/+?	22	4,4%
AP_CD_SUBJ/+CD	19	4,3%	SUBJ/+?	18	3,6%
AP_CD_SUBJ/+AP	16	3,6%	<P/+SUBJ	16	3,2%
AP_CD/+CD	13	2,9%	<P/+CD	14	2,8%
ATR_SUBJ/+ATR	8	1,8%	AP_CD/+<P	14	2,8%
ADVL_SUBJ/+ADVL	8	1,8%	CD_SUBJ/+<NN	13	2,6%
ADVL_CD_SUBJ/+ADVL	7	1,6%	SUBJ/+CD	13	2,6%
<P_CD_SUBJ/+<P	6	1,3%	AP_CD_SUBJ/+<NN	13	2,6%
AP_ATR/+AP	6	1,3%	CD/+SUBJ	12	2,4%
<P_AP_CD_SUBJ/+<P	6	1,3%	CD/+<P	11	2,2%
otros	52	11,7%	CD_SUBJ/+<P	9	1,8%
total	*446*		AP_ATR/+SUBJ	9	1,8%
			ATR/+?	7	1,4%
			ATR/+SUBJ	7	1,4%
			ATR/+AP	7	1,4%
			CD_SUBJ/+ADVL	6	1,2%
			otros	112	22,4%
			total	*500*	

PREPOSICIÓN

RESUMEN

Aciertos	754	52,9%		
Errores	30	2,1%	Precisión	0,93
Errores2	29	2,0%	Cobertura	0,55
Ambigüedades	609	42,8%	F-score(alfa=0.5)	0,74
Errores morfológicos	2	0,1%		
total	*1424*			

ACIERTOS

+<CN/<CN	423	56,1%
+ADVL/ADVL	299	39,7%
+<P/<P	32	4,2%
total	*754*	

AMBIGÜEDADES			**ERRORES**		
<CN_ADVL/+ADVL	293	48,1%	ADVL/+<CN	10	16,9%
<CN_ADVL/+<CN	236	38,8%	<CN_ADVL/+?	8	13,6%
ADVL_P-CD/+ADVL	17	2,8%	<CN_<P_ADVL/+?	4	6,8%
<CN_ADVL_P-CD/+ADVL	12	2,0%	ADVL/+?	3	5,1%
<CN_ADVL_P-CD/+<CN	10	1,6%	<CN_ADVL/+<P	3	5,1%
<CN_<P_ADVL/+<P	9	1,5%	ADVL/+P-ATR	3	5,1%
ADVL_P-ATR/+ADVL	7	1,1%	<CN/+ADVL	3	5,1%
otros	25	4,1%	<CN_ADVL/+P-CD	3	5,1%
total	*609*		ADVL/+P-CD	3	5,1%
			<CN/+?	19	32,2%
			total	*59*	

PRONOMBRE

RESUMEN

Aciertos	711	61,7%		
Errores	58	5,0%	Precisión	0,90
Errores2	22	1,9%	Cobertura	0,69
Ambigüedades	322	28,0%	F-score(alfa=0.5)	0,79
Errores morfológicos	39	3,4%		
total	*1152*			

ACIERTOS

+PR-REFLEX/PR-REFLEX	302	42,5%
+SUBJ/SUBJ	107	15,0%
+CD-CLT/CD-CLT	91	12,8%
+ADVL/ADVL	86	12,1%
+<P/<P	49	6,9%
+CI-CLT/CI-CLT	40	5,6%
+CD/CD	18	2,5%
+ATR/ATR	10	1,4%
+CN>/CN>	8	1,1%
total	*711*	

AMBIGÜEDADES

CD_SUBJ/+SUBJ	95	29,5%
CD_SUBJ/+CD	52	16,1%
CD-CLT_CI-CLT_PR-REFLEX/+CI-CLT	35	10,9%
CD-CLT_CI-CLT/+CI-CLT	23	7,1%
CD-CLT_CI-CLT_PR-REFLEX/+PR-REFLEX	22	6,8%
CD-CLT_CI-CLT_PR-REFLEX/+CD-CLT	12	3,7%
CD-CLT_CI-CLT/+CD-CLT	10	3,1%
ATR_CD-CLT/+CD-CLT	9	2,8%
ADVL_SUBJ/+SUBJ	8	2,5%
ADVL_CD_SUBJ/+SUBJ	6	1,9%
ATR_SUBJ/+SUBJ	6	1,9%
ATR_CD-CLT_CI-CLT/+CD-CLT	44	13,7%
total	*322*	

ERRORES

SUBJ/+CD	9	11,3%
SUBJ/+?	7	8,8%
CD-CLT/+CI-CLT	7	8,8%
CD/+SUBJ	4	5,0%
CD-CLT_CI-CLT/+?	4	5,0%
ATR/+SUBJ	3	3,8%
CN>/+ADVL	3	3,8%
CD-CLT/+ADVL	3	3,8%
otros	40	50,0%
total	*80*	

VERBO

RESUMEN

Aciertos	1132	58,5%		
Errores	116	6,0%	Precisión	0,87
Errores2	51	2,6%	Cobertura	0,66
Ambigüedades	576	29,8%	F-score(alfa=0.5)	0,77
Errores morfológicos	60	3,1%		
total	*1935*			

ACIERTOS

+VPRIN/VPRIN	309	27,3%
+<CN/<CN	238	21,0%
+VAUX>/VAUX>	186	16,4%
+<P/<P	126	11,1%
+<C/<C	124	11,0%
+CD/CD	65	5,7%
+ATR/ATR	26	2,3%
+ADVL/ADVL	26	2,3%
+PRED/PRED	19	1,7%
+SUBJ/SUBJ	11	1,0%
+CN>/CN>	2	0,2%
total	*1132*	

AMBIGÜEDADES

<C_VPRIN/+VPRIN	241	41,8%
<C_VPRIN/+<C	24	4,2%
<C_<P_ADVL_ATR_CD_CN>_SUBJ_VPRIN/+VPRIN	17	3,0%
ADVL_PRED/+ADVL	17	3,0%
<P_ATR_CD_SUBJ/+<P	15	2,6%
<C_<CN_VPRIN/+VPRIN	13	2,3%
<C_<CN_<P_ADVL_ATR_CD_CN>_SUBJ_VPRIN/+VPRIN	11	1,9%
<C_<CN_<P_CD/+<CN	11	1,9%
<C_<CN_VPRIN/+<CN	10	1,7%
Otros	217	37,7%
Total	*576*	

ERRORES

<CN/+<C	19	11,4%
<CN/+?	11	6,6%
<CN/+VPRIN	10	6,0%
<C_VPRIN/+<CN	7	4,2%
VPRIN/+<CN	6	3,6%
VPRIN/+<C	6	3,6%
otros	108	64,7%
total	*167*	

C Lista de investigadores

Personal de la Universitat Autònoma de Barcelona

Investigador Principal

Sergio Balari

Equipo Investigador

Lourdes Aguilar

Yolanda Rodríguez

Carme de la Mota

Teresa Vallverdú

Estudiantes de segundo ciclo

Anna-Belén Avilés

Jordi Fontseca

Personal de la Universitat Pompeu Fabra

Investigador Principal

Toni Badia

Equipo Investigador

Àlex Alsina

Gemma Boleda

Stefan Bott

Jenny Brumme

Carme Colominas

Anna Espunya

Josep Fontana

Àngel Gil

Carmen Hernández

Louise McNally

Martí Quixal

Oriol Valentín

Enric Vallduví