Anotación semiautomática con papeles temáticos de los corpus CESS-ECE

M. Antònia Martí Antonín
Centre de Llenguatge i Computació
Universitat de Barcelona
amarti@ub.edu

Mariona Taulé Delor
Centre de Llenguatge i Computació
Universitat de Barcelona
mtaule@ub.edu

Lluís Màrquez
Centre de Recerca TALP
Universitat Politècnica de Catalunya
lluism@lsi.upc.edu

Manuel Bertran
Centre de Recerca TALP
Universitat Politècnica de Catalunya
mbertran@lsi.upc.edu

Resumen: En este artículo se presenta la metodología seguida en el proceso de anotación semántica automática (estructura argumental y papeles temáticos de los predicados verbales) del corpus CESS-ECE-CAT/ESP, así como la evaluación de los resultados obtenidos. A partir de un léxico verbal (1.482 verbos) con información sobre las funciones sintácticas de cada verbo y su proyección temático-argumental, se ha anotado automáticamente el *treebank* CESS-ECE aplicando un conjunto de reglas simples sobre los árboles sintácticos. Se ha conseguido anotar automáticamente el 60% de los argumentos y papeles temáticos, con un error muy bajo (inferior al 2%). Este índice de calidad elevado permite usar la presente metodología para semi-automatizar el proceso de anotación semántica del corpus, con el consiguiente ahorro en tiempo de anotación manual. Una vez completada la anotación este corpus podrá ser utilizado como fuente de información para los sistemas de anotación automática de papeles temáticos.

Abstract: In this paper we present the methodology followed in the automatic semantic annotation (argument structure and thematic roles of the verbal predicates) of the CESS-ECE-CAT/ESP corpus. Building from a verbal lexicon (1,482 entries) with information about the syntactic functions and their projection to arguments and thematic roles, we present a set of simple rules to automatically enrich syntactic trees with semantic information. This procedure permits to automatically annotate 60% of the expected arguments and thematic roles with a fairly low error rate (below 2%). Given the high quality of the obtained results, we claim that this methodology provides substantial savings in manual annotation effort and allows a semi-automatic approach to corpus annotation. Once completed, the CESS-ECE corpus will permit researchers to develop complete systems for automatic Semantic Role Labeling of Catalan and Spanish.

1 Introducción

La anotación lingüística de corpus textuales de gran volumen es una tarea que requiere un gran esfuerzo en tiempo y recursos humanos. Cuando se pretende realizar el proceso de forma rigurosa y la anotación lingüística implica el tratamiento de información compleja (sintaxis completa con funciones sintácticas, información semántica, etc.) el esfuerzo a realizar es, si cabe, aún mayor. Por este motivo, cualquier proceso que implique la automatización parcial de los procesos de anotación va ser de una gran ayuda. Entendiendo que el proceso automático no puede sustituir en ningún caso el trabajo manual, sí puede representar una herramienta de ayuda que permita ahorrar una gran cantidad de tiempo y dinero. Para que este ahorro sea efectivo, es imprescindible que la anotación automática sea de una calidad suficiente como para que el esfuerzo de revisión manual de la anotación automática sea netamente inferior al proceso de anotación desde cero.

En este artículo se presenta la metodología seguida en el proceso de anotación semántica automática (estructura argumental y papeles temáticos) de los predicados verbales del corpus

ISSN: 1135-5948

CESS-ECE-CAT/ESP[1], en adelante simplemente CESS-ECE, así como la evaluación de los resultados obtenidos.

Para la anotación semántica automática se ha partido de dos fuentes de conocimiento: a) un léxico verbal para cada lengua elaborado manualmente a partir de ambos sub-corpus, CESS-LEX-CAT y CESS-LEX-ESP (en adelante CESS-LEX), que contiene información sintáctico-semántica; y b) un conjunto de reglas desarrolladas específicamente para la anotación automática, en las que se generalizan las correspondencias entre sintaxis y semántica especificadas en ambos léxicos.

Los léxicos verbales se han obtenido a partir de una muestra de 100.000 palabras para cada lengua de CESS-ECE (a los que llamamos *corpus-origen*) anotadas sintácticamente a nivel profundo. Estos léxicos se han utilizado para obtener las reglas de proyección sintáctico-semántica, que se han usado para la anotación automática tanto del *corpus-origen* como de otro subconjunto (85.000 palabras para el castellano y 100.000 para el catalán) a los que llamamos *corpus-prueba*. El corpus-prueba nos permitirá evaluar el grado de generalización del conjunto de reglas. En concreto, en este trabajo se analiza:

- el grado de cobertura de la anotación automática con CESS-LEX tanto en los *corpus-origen* como en los *corpus-prueba*
- la calidad de la anotación automática en los *corpus-origen* y en los *corpus-prueba* a partir de la revisión manual efectuada
- la validez del conjunto de reglas de proyección sintáctico-semántica

La evaluación de los resultados permite afirmar que en el caso de la anotación de la estructura argumental con papeles temáticos, el procesamiento automático planteado es factible y satisfactorio, puesto que: 1) se consigue anotar el 58,4% de las ocurrencias del corpus origen para el español y el 57,5% para el catalán; 2) este porcentaje mejora en ambos casos (63,4% en el español y 62,7% en el catalán) en los *corpus-prueba* (véase la sección 5.1); 3) el grado de error en la anotación automática se mantiene estable para ambas lenguas tanto en el *corpus-origen* como en el *corpus-prueba*; 4) el grado de error es realmente bajo (alrededor del 2%) en todos los corpus. Cabe destacar que para que este procedimiento sea factible, es necesario disponer de una anotación sintáctica previa de calidad, incluyendo el etiquetado de funciones sintácticas.

Una vez finalizado el proceso automático, se ha procedido a la compleción y a la revisión manual de la anotación semántica. Los corpus resultantes de dicho proceso son los que se han utilizado en la tarea 9 de evaluación de SemEval-2007 para el catalán y el castellano[2].

El resto del artículo se estructura de la siguiente forma. En la sección 2 se describen las características básicas del corpus CESS-ECE. A continuación, en las secciones 3 y 4, se presentan las diferentes fuentes de información implicadas en el proceso de anotación semántica automática. En la sección 5 se analizan los resultados obtenidos de una forma cuantitativa y cualitativa. Finalmente, en la sección 6 se presentan las conclusiones principales de este trabajo.

2 El corpus CESS-ECE

El objetivo principal del proyecto CESS-ECE es la construcción de un banco de árboles sintácticos (*TreeBank*) multilingüe (catalán, castellano y euskera) con anotación semántica[3].

En este artículo restringiremos nuestro estudio a los corpus del catalán y el castellano[4], de 500.000 palabras cada uno, procedentes de diversas fuentes, básicamente periodísticas (Véase Tabla 1). Dichos corpus se han anotado a diferentes niveles de descripción lingüística que incluyen información morfológica, sintáctica y semántica. El proceso de anotación se ha llevado a cabo de manera automática, manual o semiautomática dependiendo de la información lingüística tratada (Véase Tabla 1).

El proceso de anotación se ha llevado a cabo de manera incremental, desde los niveles más básicos de análisis, es decir empezando por el etiquetado morfosintáctico y el análisis sintáctico superficial (*chunking*) realizados automáticamente, hasta llegar a los niveles más complejos, el análisis sintáctico profundo

[1] Este corpus ha sido elaborado gracias a los proyectos CESS-ECE (HUM-2004-21127-E) y Lang2World (TIN 2006-15265-C06-06). Contiene dos partes de volumen y contenido equivalentes, una en castellano (ESP) y la otra en catalán (CAT).

[2] *Multilevel Semantic Annotation of Catalan and Spanish*, http://www.lsi.upc.edu/~nlp/semeval/msacs.html.

[3] Interfaz gráfica para consultar los corpus CESS-ECE http://www.lsi.upc.edu/~mbertran/cess-ece.

[4] El corpus del euskera se anota siguiendo otra metodología y abarca 350.000 palabras.

(proceso manual) y el análisis semántico (proceso semiautomático). Este procedimiento de anotación secuencial implica, en cada paso, la revisión manual del proceso anterior garantizando así la calidad y la consistencia interna de los datos.

Corpus	Fuentes	Anotación	Proceso
CESS-ECE-CAT	EFE (75.000) ACN (225.000) 'El Periódico'[5] (200.000)	morfosintáctica	Automático
		sintáctica superficial	Automático
		sintáctica profunda	Manual
		Papeles temáticos	Semiautomático
		Sentidos nombres (WordNet)	Manual
CESS-ECE-ESP	Lexesp [6] (85.000) EFE (225.000) 'El Periódico' (200.000)	morfosintáctica	Automático
		sintáctica superficial	Automático
		sintáctica profunda	Manual
		Papeles temáticos	Semiautomático
		Sentidos nombres (WordNet)	Manual

Tabla 1: Niveles de anotación de los sub-corpus CESS-ECE-CAT y CESS-ECE-ESP

Hasta el momento se encuentran ya disponibles las 500.000 palabras del corpus CESS-ECE anotado a nivel sintáctico superficial y profundo (constituyentes y funciones) para cada lengua y la anotación de un subconjunto de 185.000 palabras para el castellano y 200.000 para el catalán con información semántica (estructura argumental, papeles temáticos, clases semánticas y sentidos de WordNet para los nombres más frecuentes). Estos dos últimos subconjuntos son los que se analizan en este artículo. Cabe destacar que la anotación de la estructura argumental y los papeles temáticos se está realizando siguiendo un modelo inspirado en el del corpus *PropBank* para el inglés (Palmer et al., 2005). Por último, la anotación incremental de información sintáctico-semántica sobre un mismo corpus toma su modelo del proyecto *OntoNotes* (Hovy et al., 2006).

3 Fuentes de información

Como hemos dicho, la anotación semántica con estructura argumental y papeles temáticos[7] se ha realizado sobre una muestra de 185.000 palabras para el castellano (6.013 oraciones) y de 200.000 para el catalán (6.422 oraciones), y el proceso de anotación se ha llevado a cabo de manera semiautomática (Taulé et al. 2005). Estos corpus se han dividido en dos subcorpus: el *corpus-origen* (100.000 palabras) y *el corpus-prueba* (85.000 palabras para el castellano y 100.000 para el catalán). Los corpus-origen se han utilizado para la obtención de los léxicos verbales CESS-LEX-CAT y CESS-LEX-ESP, en los cuales se ha explicitado la relación entre funciones sintácticas y estructura argumental y temática de cada predicado. Los *corpus-prueba* se han utilizado para comprobar el grado de cobertura y de calidad de la anotación semántica automática. A continuación se presentan en más detalle las diferentes fuentes de información utilizadas en dicho proceso.

3.1 Información sintáctica

Se parte de la base que la estructura argumental es el nivel de representación semántica más próxima a la representación sintáctica de la oración, en tanto que refleja la manera en que los argumentos semánticos se corresponden o relacionan con sus expresiones morfosintácticas. La estructura argumental expresa la aridad del verbo y permite establecer la relación semántica entre el predicado y sus argumentos, es decir, los papeles temáticos. Es por ello que el análisis semántico parte de la información sintáctica expresada en los corpus.

A partir de la información sintáctica codificada en el *corpus-origen* se deriva de forma automática una versión inicial de los léxicos verbales CESS-LEX, donde para cada sentido de cada verbo se explicitan todos los esquemas sintácticos en los que aparece en el corpus con sus correspondientes funciones.

En nuestra aproximación el sujeto, el verbo, los complementos del verbo y los adjuntos dependen directamente del nodo oración (S).

En la figura 1 se muestra un ejemplo de análisis sintáctico profundo del corpus CESS-ECE-ESP, representando el árbol sintáctico como una expresión parentizada. Las funciones sintácticas utilizadas en ambas lenguas son: sujeto (SUJ), objeto directo (CD) e indirecto (CI), complemento del régimen (CREG), complemento agente (CAG), predicativo

[5]El subconjunto de 200.000 palabras procedentes del 'El Periódico' son las mismas noticias en catalán y español desde enero a diciembre de 2000.

[6] Lexesp es un corpus equilibrado del español de seis millones de palabras (Sebastián et al. 2000).

[7] En este artículo no se hace referencia al proceso de anotación con synsets de WordNet.

(CPRED), atributo (ATR) y circunstanciales (CC, CCT, CCL)[8]. Estas funciones están resaltadas en negrita en el árbol de la Figura 1.

```
(S
   (sn-SUJ
     (espec.fp
       (da0fp0 Las el))
     (grup.nom.fp
       (ncfp000 reservas reserva)
       (sp
         (prep
           (sps00 en en))
         (sn
           (grup.nom.ms
             (ncms000 oro oro))))))
   (morfema.verbal-PASS
     (p0000000 se se))
   (grup.verb
     (vmip3p0 valoran valorar))
   (sp-CC
     (prep
       (sps00 en_base_a
en_base_a))
     (sn
       (espec.mp
         (Z 300 300))
       (grup.nom.mp
         (Zm dólares dólar)))
```

Figura 1: Análisis sintáctico. Fragmento de frase ejemplo: "Las reservas en oro se valoran en 300 dólares…"

3.2 Los léxicos verbales CESS-LEX

A partir de la información sintáctica se crean manualmente los léxicos CESS-LEX, donde para cada sentido verbal se especifica: a) la clase semántica (Taulé et al., 2006); b) la proyección sintáctico-semántica, es decir la correspondencia entre funciones sintácticas, argumentos semánticos y papeles temáticos; c) las alternancias de diátesis en las que puede ocurrir y d) unas frases de ejemplo extraídas del propio corpus. En la figura 2 se presenta la entrada del verbo *'valorar'* en CESS-LEX.

En la entrada léxica de la figura 2, se indica el lema (*valorar*), el número de sentido (01)[9], la clase semántica (*ELS4.2*), correspondiente a la clase de verbos transitivos agentivos, que a su vez se corresponde con el tipo ontológico de las actividades, y los dos esquemas sintácticos en los que ocurre dicho verbo en el corpus (construcción activa y pasiva) con la correspondiente relación entre funciones sintácticas, posiciones argumentales y papeles temáticos. Como se puede observar, la posición argumental y el papel temático se mantienen mientras que su función sintáctica puede variar. En la construcción activa el argumento Paciente (PAT) tiene la función sintáctica de objeto directo (CD) mientras que en la pasiva es el sujeto (SUJ). Finalmente, se incluye las frases de ejemplo.

```
valorar–01  synset: 00445269v
ELS4.2
SUJ   Arg0##AGT
CD    Arg1##PAT
CC    ArgM##ADV
EJ:   "CCOO valoró la reducción del paro…"

+PASIVA
SUJ   Arg1##PAT
CAG Arg0#por#AGT
CC    ArgM##ADV
EJ:   "Las reservas en oro se valoran
       en_base_a 300_dólares estadounidenses…"
```

Figura 2: CESS-LEX-ESP: 'valorar'

Los léxicos verbales CESS-LEX proporcionan información de las distintas alternancias diatéticas que acepta un verbo (activa, pasiva, anticausativa e impersonal). Sólo la alternancia activa-pasiva tiene marcas morfosintácticas que permiten la anotación automática de argumentos y papeles temáticos de manera directa. Sobre esta información se han elaborado algunas de las reglas de proyección.

Los léxicos CESS-LEX contienen todos los verbos que aparecen en el *corpus-origen*, un total de 1.482 para el castellano y 1.052 para el catalán.

4 Reglas de proyección

Para el proceso automático de anotación semántica se ha elaborado un conjunto de reglas simples que, individualmente, describen de forma inambigua las proyecciones seguras de funciones sintácticas a argumentos y papeles temáticos. Teniendo en cuenta la dificultad de la tarea, se ha tratado de conseguir el máximo grado de cobertura minimizando el error al máximo. Distinguimos dos tipos de reglas, generales (4.1) y específicas (4.2 y 4.3).

8 Para más información y otros ejemplos se puede consultar la página web del proyecto CESS-ECE: http://www.lsi.upc.edu/~mbertran/cessece.

9 El número de sentido se asociará a uno o más *synsets* de una versión de WordNet 1.6.

4.1 Reglas generales

Incluimos bajo esta denominación un conjunto de reglas asociadas a una determinada función o propiedad morfosintáctica. En el caso de las funciones, se asignan automáticamente las posiciones argumentales correspondientes al complemento agente (CAG-Arg0-AGT), atributo (ATR-Arg2-ATR) y complemento predicativo (CPRED-Arg2-ATR)[10]. En cuanto a las posiciones adjuntas se etiquetan automáticamente los complementos circunstanciales[11] temporales (CCT-ArgM-TMP) y locativos (CCL-ArgM-LOC).

La pasiva y la pasiva refleja son un ejemplo de reglas asociadas a una propiedad morfosintáctica. En este caso la regla tiene en cuenta el tipo de verbo (es decir, si se trata del predicado 'ser'/'ésser' seguido de participio), o el rasgo morfosintáctico que lleva asociado (morfema.verbal-PASS). En ambos casos, se etiqueta de manera automática el sujeto como argumento Paciente (SUJ-Arg1-PAT) y el complemento agente como argumento Agente (CAG-Arg0-AGT), como se puede observar en la frase analizada de la figura 3. La asignación de dichos argumentos y papeles temáticos es independiente del verbo. En el caso del ejemplo, la asignación de la clase semántica D2 (verbos transitivos agentivos) al predicado 'firmar' se ha realizado también de manera automática pero a partir de la información especificada en el léxico verbal, CESS-LEX-ESP en este caso, es decir mediante la aplicación de las reglas del segundo tipo.

4.2 Reglas específicas dependientes de los léxicos CESS-LEX

Incluimos en esta sección las reglas específicas basadas en la información descrita en los léxicos verbales CESS-LEX, es decir reglas dependientes de cada predicado.

En el caso de verbos monosémicos la correspondencia entre función sintáctica y clase semántica, argumentos y papeles temáticos se realiza de manera directa. En la figura 4 se ejemplifica con el verbo 'ser' tratado monosémicamente en el corpus del castellano.

[10] CPRED-Arg3-ORI en caso de estar introducido por la preposición: "sps00 procedente_de".

[11] Son los únicos circunstanciales anotados manualmente.

En la frase analizada en la figura 4, se ha

```
(S
 (sn-SUJ-Arg1-PAT
   (espec.ms
     (da0ms0 El el))
   (grup.nom.ms
     (ncms000 acuerdo acuerdo)))
  (grup.verb-D2
   (vsif3s0 será ser)
   (vmp00sm firmado firmar))
  (sp-CAG-Arg0-AGT
   (prep
     (sps00 por por))
   (sn.co
     (snp
       (espec.fs
         (da0fs0 la el))
       (grup.nom.fs
         (ncfs000 presidenta
presidenta)))
```

etiquetado directamente el sujeto como un

Figura 3: Reglas tipo I. Ejemplo de anotación automática directa. Fragmento de frase ejemplo "El acuerdo será firmado por la presidenta…"

```
(S
  (grup.verb-C3
    (vsip3p0 Son ser))
  (sa-ATR-Arg2-ATR
    (espec
      (rg tan tanto))
    (grup.a
      (aq0mp0 raros raro)))
  (sn-SUJ-Arg1-TEM
    (Fc , ,)
    (espec.mp
      (da0mp0 los el))
    (grup.nom.mp
      (ncmp000 hombres
hombre)))
```

Figura 4: Reglas tipo II. Ejemplo de anotación automática directa. Frase ejemplo "Son tan raros, los hombres"

Arg1-TEM (Tema), el atributo como un Arg2-ATR (Atributo) y la clase semántica C3, que corresponde a verbos estativos transitivos.

En el caso de verbos polisémicos la equivalencia entre información sintáctica y semántica puede ser parcial, en el sentido de que sólo se etiqueta automáticamente aquella información que es inequívoca, es decir que no implica ambigüedad. El resultado, por lo tanto, puede ser parcial, bien porque sólo se etiqueta la posición argumental o el papel temático; bien porque sólo se etiqueta la combinatoria entre clase semántica y posición argumental o clase semántica y papel temático. En el primer caso se trata de predicados que, para sentidos o

acepciones distintas, comparten la misma función y posición argumental pero los papeles temáticos asociados son diferentes y, por lo tanto, la elección de uno u otro no es inmediata. La segunda posibilidad es el caso contrario, en las distintas acepciones del predicado se produce una coincidencia en la asignación de papeles temáticos pero no en la posición argumental de los mismos. En la figura 5 se muestra un ejemplo de anotación automática parcial.

```
(S
    (sadv-CC-ArgM-TMP
      (grup.adv
        (rg Entonces entonces)))
    (sn.e-SUJ *0*)
    (grup.verb
      (vmii3s0 movía mover))
    (sn-CD-Arg1-PAT
      (espec.fs
        (da0fs0 la el))
      (grup.nom.fs
        (ncfs000 cabeza cabeza)))
    (sadv-CC
      (grup.adv
        (rg lentamente
```

Figura 5: Reglas tipo II. Ejemplo de anotación automática parcial. Frase ejemplo "Entonces movía la cabeza lentamente".

El sujeto de 'mover' no se ha podido asignar de forma automática porque puede interpretarse distintamente en función de la acepción que se considere, en este caso un Arg0-CAU, un Arg0-AGT o un Arg1-TEM (Véase anexo 1 para consultar la entrada de 'mover'). Por la misma razón tampoco ha sido posible la asignación de la clase semántica de manera automática ya que la regla no puede decidir si pertenece a la clase semántica A1, C1 o C2.

4.3 Reglas específicas basadas en el tipo de adverbio y preposición

En este tercer conjunto consideramos las reglas basadas en el tipo de adverbio o locución adverbial o prepositiva que aparecen en un determinado constituyente. Por ejemplo, en la figura 6 se da una muestra del tipo de información considerada.

Como se puede ver, las reglas tienen en cuenta el tipo de categoría morfosintáctica, 'RG' (adverbio) o 'SPS' (locución prepositiva), la función de complemento circunstancial y el

```
a_base_de SPS00 ArgM##ADV
a_cambio_de SPS00 ArgM##ADV
a_causa_de SPS00 ArgM##CAU
a_comienzos_de SPS00 ArgM##TMP
allí_delante RG ArgM##LOC
allí_encima RG ArgM##LOC
así RG ArgM##MNR
atrás RG ArgM##LOC
aun RG ArgM##ADV
...
```

tipo concreto de adverbio o locución. Por ejemplo, si en un constituyente, con la categoría

Figura 6: Lista de adverbios y locuciones

morfosintáctica 'RG' o 'SPS' y con función CC, aparece el adverbio 'así' o la locución 'a_causa_de' se asignará de forma automática el argumento y papel temático ArgM-MNR (manera) y ArgM-CAU (causa), respectivamente.

Un subtipo de estas reglas es aquel que trata expresiones temporales representadas en la categoría morfosintáctica como W. Así, se anota como ArgM-TMP todo CC que contine

```
(snd-CC-ArgM-TMP
    (espec.ms
      (da0ms0 el el))
    (grup.nom.ms
      (W 26_de_mayo [??:26/5/??:??.??]))
  (sn.e-SUJ-Arg1-TEM *0*)
  (grup.verb-C3
    (vmii3p0 equivalían equivaler))
  (sp-CREG-Arg2-ATR
    (prep
      (sps00 a a))
    (snn
      (espec.mp
        (Z 19.100 19100))
      (grup.nom.mp
        (ncmp000 millones millón)
        (sp
          (prep
            (sps00 de de))
          (grup.nom.mp
        (Zm dólares dólar)))))))
```

un elemento del tipo W (véase figura 7):

Figura 7: Reglas de tipo III. Expresiones temporales. Frase ejemplo: "El 26 de mayo equivalían a 19.100 millones de dólares."

4.4 Aplicación de las reglas

Todas las reglas de proyección se han podido implementar de manera sencilla y eficiente,

requiriendo un esfuerzo en tiempo de programación moderado. De todas maneras, para disponer de un procesador automático que las aplique es necesario decidir qué hacer en los casos en que más de una regla sea aplicable. Heurísticamente, se ha decidido aplicar las reglas en orden descendiente de generalidad, es decir primero se aplican las reglas generales de tipo I, en segundo lugar las de tipo III y, finalmente las de tipo II.

5 Metodología y análisis de los resultados

El proceso de aplicación automática de las reglas de proyección se ha llevado a cabo en ambos corpus. Recordemos que el proceso automático se realiza separadamente sobre los *corpus-origen* y los *corpus-prueba* a partir de las fuentes de conocimiento desarrolladas, reglas y léxicos, con el fin de evaluar la validez de dichas fuentes independientemente del corpus a partir del cual se ha extraído la información. En el proceso automático se obtienen anotaciones totales con información sobre el argumento y el papel temático o bien anotaciones parciales, es decir cuando sólo se ha podido asociar a la función el argumento o el papel temático. Los resultados obtenidos en el proceso automático de anotación se presentan en la sección 5.1. Incluimos un análisis cuantitativo de la cobertura de los léxicos y de las reglas en los corpus origen y de prueba, es decir, el número de funciones sintácticas que han recibido información argumental y temática del conjunto total de funciones.

Por otro lado, se ha realizado la revisión manual para corregir los errores de la anotación automática y completar las funciones que no han recibido anotación semántica tanto de los *corpus-origen* como del *corpus-prueba*. Para la anotación manual se utiliza un editor de árboles, TreeTrans (Cotton y Bird, 2000), adaptado para la anotación de argumentos y papeles temáticos. Una vez realizado el análisis cuantitativo de la calidad de la anotación automática, se presenta una valoración cualitativa de los resultados obtenidos (sección 5.2).

5.1 Análisis cuantitativo de los resultados

Los léxicos CESS_LEX se han obtenido a partir de los verbos de los corpus-origen. Por lo tanto, se plantea un problema de cobertura cuando se anota automáticamente el corpus-prueba. El léxico del castellano contiene un total de 1.482 verbos de los cuales sólo 717 aparecen en el corpus-prueba, lo que constituye el 64.1% de los 1.119 verbos del corpus de prueba. El léxico del catalán contiene un total de 1.052 verbos de los cuales sólo 664 aparecen en el corpus-prueba, lo que representa el 58,5% de los 1.134 verbos del corpus-prueba.

A pesar de que los léxicos tienen menor cobertura en los corpus-prueba, el tanto por ciento de anotación recibida mantiene el mismo nivel de cobertura que en los corpus-origen, lo que indica que las reglas de proyección independientes del léxico (tipos I y III) expresan generalizaciones que se dan también para los verbos nuevos de los *corpus-prueba*. En realidad, el porcentaje de cobertura es incluso un poco mejor en los corpus de prueba como veremos más adelante. Comentamos a continuación los resultados de la anotación de manera analítica.

El número total de funciones sintácticas que aparecen en los corpus es de 48.405 para el castellano y 48.600 para el catalán. De éstas, en el corpus del castellano, 25.645 pertenecen al corpus-origen y 22.760 al corpus-prueba; en el caso del catalán, 24.005 pertenecen al corpus-origen y 24.665 al corpus-prueba. De todas las funciones, reciben anotación semántica el sujeto (SUJ), el complemento directo (CD), indirecto (CI), de régimen (CREG), el atributo (ATR), el complemento predicativo (CPRED) y los circunstanciales (CC), que corresponden a un total de 44.499 funciones por anotar en castellano (23.587 en corpus-origen y 20.912 en corpus-prueba) y 43.531 para el catalán (21.466 en el corpus origen y 22.065 en el de prueba).

No reciben información semántica los adjuntos oracionales (AO), el vocativo (VOC), los elementos textuales (ET) y las marcas de impersonalidad, negación, pasiva e impersonal. De todas formas, esta limitación nos hace renunciar solamente a 3.906 funciones sintácticas (un 8,07% del total de funciones) en castellano y a 5.139 (un 10,5 % del total) en catalán.

En la tabla 2 se muestra de manera desglosada por funciones la anotación semántica resultante. También se incluye un desglose por cada uno de los corpus (origen/prueba). Cada celda contiene tres números que representan el número de funciones etiquetadas con las reglas automáticas, el número de

CORPUS: CESS-ECE-CAT			
Función	**C-origen**	**C-prueba**	**Total**
SUJ	3.819/7.075/53,9%	3.996 / 7.517 / 53,1%	7.815 / 14.592 / 53,5%
CD	4.099/5.055/81,08%	3.933 / 5.302 / 74,1%	8.032 / 10.357/ 77,5%
CI	406/407/99,7%	429/ 429/ 100%	835/ 836 / 99,8%
CREG	767/1.253/61,21%	646 / 1.240 / 52,0%	1.413 / 2.493 / 56,6%
ATR	903/903/100%	1.111/ 1.111/ 100%	2.014 / 2.014 / 100%
CPRED	390/390/100%	412 / 413 / 99,7%	802 / 803 / 99,8%
CAG	297/297/100%	226/ 227/ 99.5%	523/ 524 / 99,8%
CC	1.620/6.006/26,97%	3.089 / 5.826 / 53,0%	4.709 / 11.832 / 39,7%
Total	**12.301 / 21.386 / 57,5%**	**13.842 / 22.065 / 62,7%**	**26.143 / 43.451 / 60,1%**

CORPUS: CESS-ECE-CAST			
Función	**C-origen**	**C-prueba**	**Total**
SUJ	5.207 / 7.746 / 67.2%	4.631 / 6.967 / 66.5%	9.838 / 14.713 / 66,8%
CD	3.438 / 4.838 / 71.1%	3.327 / 5.018 / 66.3%	6.765 / 9.856 / 68,6%
CI	502 / 612 / 82.0%	261 / 521 / 50.1%	763 / 1.133 / 67,3%
CREG	585 / 857 / 68.3%	470 / 869 / 54.1%	1.055 / 1.726 / 61.1%
ATR	1.537 / 1.550 / 99.2%	955 / 963 / 99.2%	2.492 / 2.513 / 99,2%
CPRED	361 / 361 / 100%	288 / 288 / 100%	649 / 649 / 100%
CAG	188 / 189 / 99.5%	263 / 265 / 99.2%	451 / 454 / 99,3%
CC	1.953 / 7.434 / 26.3%	3.062 / 6.021 / 50.9%	5.015 / 13.455 / 37,2%
Total	**13.771 / 23.587 / 58.4%**	**13.257 / 20912 / 63.4%**	**27.028 / 44.499 / 60,7%**

funciones a etiquetar y el porcentaje de cobertura, respectivamente.

Tabla 2: Resultados de la anotación automática desglosada por sub-corpus, funciones e idioma

Podemos ver como del total de funciones sintácticas que pueden recibir anotación semántica, 44.499 para el castellano y 43.531 para el catalán, se han anotado automática-mente un total de 27.028 y 26.143 respectivamente, lo que corresponde a un 60'7% para la primera lengua y a un 60,1% para la segunda: el 58'4% del corpus-origen y el 63'4% del corpus-prueba para el castellano y el 57,5% y el 62,73% respectivamente para los corpus del catalán. Como se puede observar la cobertura de anotación es superior en el corpus de prueba debido fundamentalmente al etiquetado de los CC. Esto se explica porque, por razones de criterios de anotación de los corpus, la anotación sintáctica del CC es más rica en el corpus de prueba que en el corpus de origen[12].

En lo que se refiere a la cobertura, como se puede observar en la tabla 2, las funciones sintácticas ATR, CAG, CPRED en ambas lenguas y el CI, en el caso del catalán, reciben de manera casi unívoca una determinada anotación semántica y su grado de cobertura alcanza casi el 100%. En el caso del CD y del CREG en ambas lenguas y del CI en el caso del castellano, el grado de cobertura de la anotación es inferior en el corpus-prueba debido a que en este corpus la cobertura del léxico es inferior. La razón estriba en el hecho de que la asignación de argumento y papel temático, en estos casos, depende exclusivamente de las reglas de tipo II (léxicas). La función sujeto recibe menos anotaciones semánticas en catalán (7.759) que en castellano (9.690) porque el verbo ser -el más frecuente en el corpus- se ha tratado de forma monosémica en castellano, mientras que en catalán es polisémico.

Las posiciones no argumentales, que corresponden a los CC sintácticos, son las que dan peores resultados debido a la amplia tipología de argumentos (ArgM, Arg2, Arg3, Arg4) y papeles temáticos que pueden recibir (LOC, TMP, MNR, etc.).

Finalmente, en la tabla 3 se muestra el número total y el porcentaje de anotaciones totales y parciales. En la fila de los "Totales" se incluye el número total de funciones

[12] En el corpus de prueba los complementos circunstanciales de tiempo y lugar reciben una etiqueta específica, CCT y CCL respectivamente, que facilita la asignación semántica.

etiquetadas y el porcentaje que representa este número con respecto al total de anotaciones realizadas automáticamente y al total de funciones sintácticas por anotar. Se puede observar como, en casi todos los casos, la mayoría de anotaciones son completas (argumento + papel temático). El alto número de anotaciones parciales en el caso del complemento de régimen (CREG) se debe a que se ha decidido, de momento, no asignarle papel temático por la problemática que entraña. Este tema está en fase de estudio y se resolverá en una fase posterior. Los complementos con mayor ambigüedad y, por lo tanto, con mayor número de anotaciones parciales, son el sujeto (SUJ) y el complemento directo (CD): la posición argumental es deducible, pero no así el papel temático.

CORPUS: CESS-ECE-CAT				
Función	**Arg+PT**	**Arg**	**PT**	**Total**
SUJ	7.759	52	4	7.815
CD	7.854	178	---	8.032
CI	835	---	---	835
CREG	157	1.256	---	1.413
ATR	2.014	---	---	2.014
CPRED	802	---	---	802
CAG	523	---	---	523
CC	4.608	101	---	4.709
Total	**24.552**	**1.587**	**4**	**26.143**
	93,9%	6,0%	0,01%	100%
	56,5%	3,6%	0,0%	**60,1**%

CORPUS: CESS-ECE-CAST				
Función	**Arg+PT**	**Arg**	**PT**	**Total**
SUJ	9.690	144	4	9.838
CD	6.600	129	36	6.765
CI	703	---	60	763
CREG	57	998	---	1.055
ATR	2.492	---	---	2.492
CPRED	649	---	---	649
CAG	451	---	---	451
CC	4.859	156	---	5.015
Total	**25.501**	**1.427**	**100**	**27.028**
	94,3%	5,3%	0,4%	100%
	57,3%	3,2%	0,2%	**60,7**%

Tabla 3: Anotación total vs. parcial en el corpus origen+prueba del catalán y castellano

5.2 Análisis cualitativo de los resultados

Actualmente se ha completado la anotación semántica manual del corpus-origen y se ha revisado y validado el etiquetado obtenido automáticamente para ambas lenguas. En cuanto al corpus-prueba, se ha revisado y completado el 10% de las oraciones de ambos corpus. En este proceso de validación se ha computado el número de errores para un subconjunto de 500 funciones en cada corpus de ambas lenguas. El resultado obtenido es de un 2,1% de asignaciones incorrectas en el corpus del castellano y de 1,9% en el corpus del catalán, fundamentalmente en los CC.

La precisión de la anotación automática depende básicamente de dos factores: las reglas y el léxico. Las reglas generales son aplicables a cualquier corpus y el resultado debería tener un nivel de calidad equivalente. No ocurre lo mismo con las reglas dependientes de la información contenida en el léxico, ya que este está constituido sobre el corpus de origen. Los problemas en este caso se circunscriben a la cobertura de verbos y sentidos.

De los resultados obtenidos en la revisión manual se puede concluir, por lo tanto, que las reglas de tipo 1 y 3 dan resultados satisfactorios y son aplicables a cualquier otro corpus. Un análisis más profundo del complemento circunstancial y la proyección del mismo en la anotación semántica, mejoraría los resultados de las reglas de tipo 3.[13] Los errores detectados se refieren fundamentalmente a la asignación de papeles temáticos a los CC, debido a la ambigüedad y variedad de los mismos y por la presencia de sentidos en los corpus-prueba que no aparecen en los corpus origen.

En cuanto a las reglas de tipo 2, en tanto que dependen del léxico, serían mejorables aumentando la cobertura de los mismos y si trataran las preposiciones que rigen los verbos en determinados complementos circunstanciales.

A la vista de los resultados obtenidos tanto en la cobertura de anotación (una media el 60,4%) como en la alta calidad de ésta (alrededor de un 98% de aciertos) es innegable que la metodología propuesta supone un ahorro importante y resuelve en gran medida la tarea de la anotación de corpus con argumentos y papeles temáticos. Ello se debe, entre otras razones, a la base lingüística incorporada en las

[13] Esta ampliación se está ya aplicando para la anotación semántica automática del corpus CESS-ECE-CAT.

reglas y a la calidad de los procesos de anotación previos (morfosintáctico y sintáctico). Para hacernos una idea, se ha cuantificado que el coste de anotación manual del 40% no cubierto por el proceso automático junto con la revisión manual del 100% del texto de los dos idiomas es de 1.655 horas persona. El coste de haber etiquetado manualmente el 60% cubierto por las reglas automáticas se estima comparable a esta cantidad, mientras que el desarrollo y la implementación de las reglas no ha superado las 100 horas persona, dejando el ahorro neto en una cantidad superior a 1.500 horas persona.

6 Conclusiones

En este artículo se ha presentado la metodología seguida en la anotación automática del corpus CESS-ECE con estructura argumental y papeles temáticos. Se trata de un proceso automático previo a la anotación manual completa. Las reglas que se han aplicado, a la vista de los resultados cualitativos obtenidos, tienen una precisión próxima al 100%, aunque se ha sacrificado la cobertura ya que sólo se anota un 60% del corpus. No se trata en ningún caso de un sistema de anotación automática de roles temáticos, sino de un pre-proceso automático de ayuda en la anotación y revisión manuales de un corpus con esta información. Una vez completado, el corpus posibilitará, entre otras cosas, estudios lingüísticos empíricos y también la aplicación de técnicas de aprendizaje automático para desarrollar herramientas automáticas de análisis de los nuevos niveles semánticos incorporados.

Cabe destacar, en primer lugar, que el grado de cobertura alcanzado tanto en el corpus-origen como en el corpus-prueba es prácticamente el mismo (algo superior en este último). En segundo lugar, la alta calidad de la anotación automática (98%). Finalmente, resulta de especial interés el conjunto de reglas que se ha elaborado tanto por su alta resolución como por las generalizaciones lingüísticas que recogen.

Bibliografía

Cotton, S. y S. Bird. An Integrated Framework for Treebanks and Multilayer annotations. En *Proceedings of the 2nd International Conference on Language Resources and Evaluation, LREC-2000*. Atenas, 2000.

Hovy, E.H., M. Marcus, M. Palmer, S. Pradhan, L. Ramshaw, and R. Weischedel. OntoNotes: The 90% Solution. En *Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL 2006)*. New York, NY, 2006.

Palmer, M., D. Gildea, y P. Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles, *Computational Linguistics*, 31(1), 2005.

Sebastián, N., M. A. Martí, M. F. Carreiras y F. Cuetos. LEXESP: Léxico Informatizado del Español, Ediciones de la Universidad de Barcelona, Barcelona, 2000.

Taulé, M., J. Aparicio, J. Castellví, y M. A. Martí. Mapping syntactic functions into semantic roles. En *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*. Universitat de Barcelona, Barcelona, 2005.

Taulé, M., M. A. Martí y J. Castellví. Semantic Classes in CESS-LEX: Semantic Annotation of CESS-ECE. En *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT-2006)*. Prague, Czech Republic, 2006.

Anexo 1: Entrada verbal 'mover'

Este anexo presenta la información de la entrada verbal 'mover' en CESS-LEX, que tiene tres sentidos diferenciados.

mover–01 synset: **01249365v**
C1
SUJ Arg1##TEM
CC Arg4#hacia#DES
CC ArgM##MNR
EJ: "que se movía lentamente"

mover–02 synset: **01263706v**
C2
SUJ Arg0##AGT
CD Arg1##PAT
CC Arg2#por#LOC
CC ArgMcon#ADV
EJ: "las naves que movía el viento por aguas y mares con estimable eficiencia "

mover–03 synset: **01133437v**
A1
SUJ Arg0##CAU
CREG Arg1#a#
EJ: "resultó ser de los que mueven a la reflexión"