

El proyecto SenSem: sentidos verbales y anotación de corpus

GRIAL



GRIAL

Grup de Recerca Interuniversitari en Aplicacions Lingüístiques

Grup de recerca emergent de la Generalitat de Catalunya (2005-2009)

- UAB
 - Mercè Coll: professora
 - Ana Fernández: professora
 - David Teruel: informàtic de projecte
- UB
 - Irene Castellón: professora
 - Nevena Tinkova: becària investigació
- UdL
 - Joan Antoni Capilla: personal de projecte
 - Iolanda Mateu: personal de projecte
 - Mhaela Topor: becària investigació
 - Glòria Vázquez: professora
- UNC
 - Laura Alonso: Becària Postdoctoral

GRIAL

Contenido

- Antecedentes
- El proyecto SenSem
- Desarrollo
- Situación actual

GRIAL

Contenido

- **Antecedentes**
- El proyecto SenSem
- Desarrollo
- Situación actual

GRIAL

Antecedentes

- Proyectos VOLEM I y VOLEM II
 - Xarxa temàtica interregional (ABM/acs/XTI-CTP2000-1)
 - Xarxa temàtica interregional (ABM/acs/XI 2003-12)
- Construcción de una base de datos léxica verbal
- Colaboración con otros grupos
 - (IRIT Toulouse, TALP - UPC, IXA -UPV)

GRIAL

Antecedentes

- Base de datos verbal
 - Clases verbales
 - Descripción verbal sintáctico-semántica
 - Unidad: sentido verbal
 - Conexión multilingüe
- Basado en descripción teórica (Vazquez 1999; Fernández 2000)
- Poca evidencia empírica

GRIAL

Contenido

- Antecedentes
- **El proyecto SenSem**
- Desarrollo
- Situación actual

GRIAL

El proyecto SENSEM

- **Sentence Semantics: Creación de una Base de Datos de Semántica Oracional (BFF2003-06456)**
- Duración 3 años
- Grupos implicados:
 - UAB, UdL, UB (GRIAL)
 - UOC
 - EHU-UPV

GRIAL

El proyecto SENSEM

OBJETIVO

- Creación de un banco de datos verbal
 - Construir un **corpus** de oraciones anotadas
 - Construir una **base de datos léxica**
 - **Asociar** el corpus con la base (ejemplos)

GRIAL

El proyecto SENSEM

- Mejorar la información que contenía VOLEM:
 - En cuanto a la descripción
 - En cuanto a los datos concretos
- No se asumen clases semánticas a priori
- Tener pruebas empíricas de dicha descripción
- Ampliar dicha descripción

GRIAL

Contenido

- Antecedentes
- El proyecto SenSem
- **Desarrollo**
- Situación actual

GRIAL

Contenido

- Antecedentes
- Objetivos principales
- **Desarrollo**
 - **Metodología**
 - Proceso de anotación
 - Tratamiento del error
 - Adquisición
 - Construcción de la base verbal
- Situación actual

GRIAL

Metodología

- Construcción del corpus a partir de un listado de lemas verbales
 - Selección de 250 formas verbales → alta frecuencia
 - Selección de 100 oraciones de cada forma verbal de un corpus periodístico
 - Total 25.000 frases + contexto
 - *El Periódico de Catalunya*.
 - Los ejemplos excluyen los usos perifrásticos de los verbos, así como expresiones idiomáticas y colocaciones
- Anotación
- Corrección
- Adquisición → BD léxica

GRIAL

Metodología

- Anotación: oración relativa a una forma verbal:

- [ID-Frase: 4316] abrir

El discurso del purpurado gallego, **que abrió** la **asamblea plenaria de primavera de la CEE**, tuvo un tono conciliador con el Ejecutivo salido de las elecciones del14-M.

GRIAL

Metodología

■ Anotador:

- Un único anotador anota los 100 ejemplos de un verbo en todos sus niveles
 - Estudia los diferentes sentidos
 - Coherencia
- Se realiza una corrección de la anotación

GRIAL

Recursos humanos

■ Anotación

- Anotadores: 5 (parcial): 2'5 personas /mes
- Coordinadores: 2 (parcial): 1 persona/mes
- Correctores: 3 (parcial) : 1 persona/mes
- Interfaces: 1 total: 1 persona/mes
- Adquisición y base de datos: 4 (parcial)

GRIAL

Recursos lingüísticos y herramientas

- Interfaz de anotación e interfaz de búsqueda
- Base de datos verbal VOLEM/SENSEM
 - Distinción de sentidos
 - Clase eventual léxica
 - Papeles semánticos iniciales
- Corpus de El Periódico

GRIAL

Contenido

- Antecedentes
- El proyecto Sensem
- Desarrollo
 - Metodología
 - **Proceso de anotación**
 - Establecimiento de criterios
 - Herramientas
 - Niveles de anotación
 - Corrección de errores
 - Adquisición
 - Construcción de la base verbal
- Situación actual

GRIAL

Anotación

Tipo de anotación realizada

- ❑ Desambiguación del sentido verbal
- ❑ Constituyentes
- ❑ Clase eventual
- ❑ Semántica de la construcción
- ❑ Usos metafóricos (argumentos y sentidos verbales)

GRIAL

Desambiguación de sentidos

acabar

- 1.- Finalizar algo.
- 2.- Consumir una cosa totalmente.
- 3.- Destruir, aniquilar, erradicar algo.
- 4.- Ser una cosa en su final de una forma determinada.
- 5.- Llegar a ser o a hacer algo.
- 6.- Haber ocurrido alguna cosa hace poco tiempo.
- 7.- Quedar algo o alguien en una situación, lugar o forma determinadas.

9066.- El republicano **acabó** su discurso advirtiendo en contra de la utilidad de votar a CiU o al PSC

GRIAL

Desambiguación de sentidos

acabar

ID:	1
Definición:	Finalizar algo.
RS:	[ag/caus,t-af,circ]
EE:	evento
Wordnet:	00211850v
Sinónimos:	

GRIAL

Constituyentes

- Delimitación
- Categoría
- Núcleo
- Función
- Argumento/adjunto
 - ❑ Papel semántico (distribución)

GRIAL

Clase eventual

- Objetivo inicial
 - ❑ Clase eventual de nivel léxico
 - ❑ Clase eventual del SV
 - ❑ Clase eventual oracional

GRIAL

Clase eventual

- Realidad:
 - ❑ Anotación parcial
 - ❑ Completa a nivel léxico
 - ❑ No se anota a nivel de SV
 - ❑ Algunos casos de nivel oracional

GRIAL

Semántica de la construcción

- Anticausativa (Los datos han variado, se ha quemado el estofado.)
- Antiagentiva (Se ha comprado demasiado pan.)
- Pasiva (Se siente mucho su pérdida)
- Reflexiva (Juan se lava el pelo.)
- Recíproca (Ana y Pepa se hacen regalos por reyes.)
- Causa indirecta (Juan se ha construido una casa)
- Dativo de interés (La niña no me come; Se me ha quemado el cocido)
- Impersonal (En TVE aseguran que sigue en la cadena estatal)

GRIAL

Anotación

[ID-Frase: 9566]
Verbo: saber

Evento	Módulo	Módulo
El	aprobamos	sabrá
Sigto	Obj	Circonfuncional
SN	SN	OSer
Apote	T.af	Or

GRIAL

Anotación

[ID-Frase: 9213]
Verbo: bajar

Fuente		Núcleo
Anticausativa		
El oncólogo reconoció que , a diferencia de otras enfermedades , no	bajan	las cifras de muerte por cáncer
Circunstancial	SAdjNeg	Sujeb
SP		SN
		T.af

GRIAL

Anotación

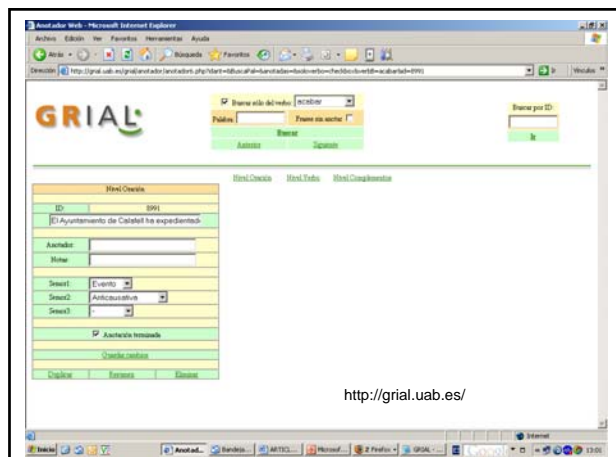
- Trabajo inicial:
 - Establecimiento de criterios
 - Anotación inicial
 - Cálculo del acuerdo entre anotadores
- Clase eventual: 77 %
- Argumento/adjunto: 73 %
- Papel semántico: 79,1 %
- Función sintáctica: 77 %
- Categoría sintáctica: 86,7 %

GRIAL

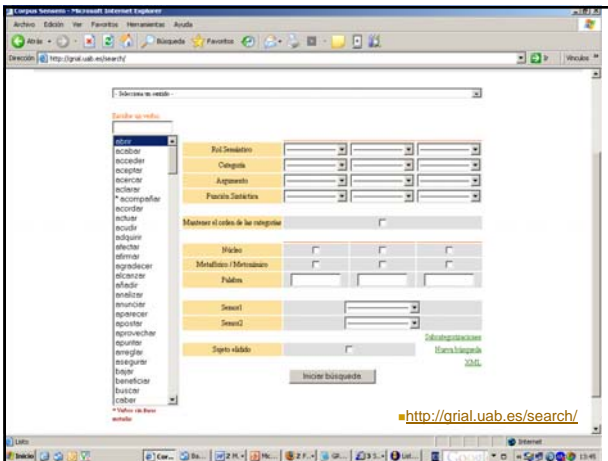
Herramientas

- <http://grial.uab.es>
- <http://grial.uab.es/search/>

GRIAL



<http://grial.uab.es/>



XML

```

<sl ID="2" semor="1" Evento="anotado" s="13" lema_verbo="celebrar" sentido="celebrar_3">
  <ctid>
    <w id="1" forma=">
      <w id="2" forma="Estaríamos">
        <w id="3" forma="encantados">
          <w id="4" forma="de">
            </ctid>
            <gtr id="1" r="Agentes" cat="SN" fs="Sujeto" Argumento="1">
              <w id="5" forma="de">
                <w id="6" forma="consejo" nucleo="1">
                  </gtr>
                  <gtr id="7" forma="celebrar" nucleo="1">
                    <gtr id="2" r="T" cat="SN" fs="Obj Directo" Argumento="1">
                      <w id="8" forma="de">
                        <w id="9" forma="junta" nucleo="1">
                          </gtr>
                          <gtr id="3" cat="SP" fs="Circunstancial">
                            <w id="10" forma="de">
                              <w id="11" forma="frases">
                                <w id="12" forma="de">
                                  <w id="13" forma="juicio" nucleo="1">
                                    </gtr>
                                    </ctid>
                                    <w id="14" forma=">
                                      <w id="15" forma="pero">
                                        <w id="16" forma="todas">
                                          <w id="17" forma="venemos">
                                            <w id="18" forma="claro">
                                              <w id="19" forma="que">
                                                <w id="20" forma="no">
                                                  <w id="21" forma="será">
                                                    <w id="22" forma="así">
                                                      <w id="23" forma=">
                                                        <w id="24" forma=">
                                                          </ctid>
                </sl>
  
```

GRIAL

Contenido

- Antecedentes
- El proyecto Sensem
- **Desarrollo**
 - Metodología
 - Proceso de anotación
 - **Tratamiento del error**
 - Adquisición
 - Construcción de la base verbal
- Situación actual

GRIAL

Tratamiento del error

- Lاپso del anotador
- Clases ambiguas
- Error en algún concepto gramatical

GRIAL

Tratamiento del error

- Realización de revisiones durante todo el proyecto:
 - Acciones en la base verbal:
 - Modificación de sentidos (eliminación, compactación, aparición)
 - Modificación de la definición (explicativa)
 - Modificación de papeles semánticos
 - Modificación o incorporación de la clase eventual

GRIAL

Tratamiento del error

- Realización de revisiones durante todo el proyecto:
 - Acciones sobre el corpus
 - Asociación del sentido
 - Función o categoría
 - Semántica oracional

GRIAL

Tratamiento del error

- **Detección automática**
 - Cadenas anómalas
 - Si SN → no finaliza con DET
 - Si SP → no finaliza con PREP
 - Cadenas incompatibles
 - Antiagentiva → no agente
 - SN → no Obj_PrepX
 - Estado → no Causativa
 - Cadenas implicadas
 - Estado → Tema
 - Agentiva → Agente
 - SN & Argumento → odirecto, sujeto

GRIAL

Tratamiento del error

- **Funciones :**
 - Objetos preposicionales 1 y 2
 - Objetos indirectos
- **Categorías :**
 - Subordinadas adverbiales
 - Sintagmas preposicionales
 - Construcciones de gerundio
- **Clases eventuales**
 - Estados / Eventos
- **Papeles semánticos**
 - Iniciador
 - Temás

GRIAL

Tratamiento del error

- Errores detectados sobre todas las categorías a anotar **24,5 %**
- Oraciones con error **17 %**
- Datos provisionales, previos a la corrección

GRIAL

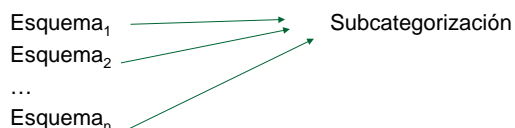
Contenido

- Antecedentes
- El proyecto Sensem
- **Desarrollo**
 - Metodología
 - Proceso de anotación
 - Tratamiento del error
 - **Adquisición**
 - Construcción de la base verbal
- Situación actual

GRIAL

Adquisición

- Proceso de compactación de los esquemas en los que participa el verbo



GRIAL

Adquisición

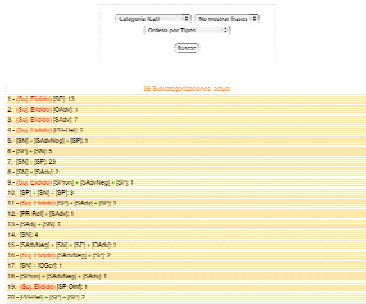
Extracción de esquemas :

- Categoría
- Rol semántico
- Funciones
- Argumentos y adjuntos
- Número de ocurrencias
- Orden diferente

GRIAL

Extracción de esquemas

■ <http://grial.uab.es/search>



GRIAL

Adquisición

■ Compactación

- Orden
- Argumentos
- Categorías
 - Con similar distribución
 - Con igual función

GRIAL

Adquisición



GRIAL

Contenido

- Antecedentes
- El proyecto Sensem
- Desarrollo
 - Metodología
 - Proceso de anotación
 - Tratamiento del error
 - Adquisición
 - **Construcción de la base verbal**
- Situación actual

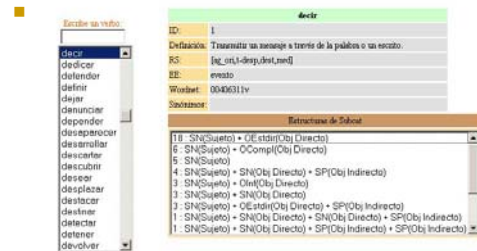
GRIAL

La base de datos verbal

- Unidad: sentido
- Contenido
 - Definición
 - Papeles semánticos
 - Clase eventual léxica
 - Sinónimos
 - Asociación a EuroWordNet (1.5)
 - Subcategorización
 - Semántica oracional
 - Restricciones selectivas
 - Ejemplos asociados (corpus anotado)

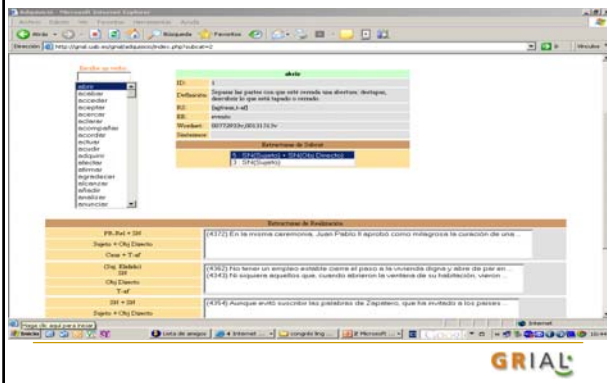
GRIAL

La base de datos verbal



GRIAL

Base de datos verbal



Contenido

- Antecedentes
- El proyecto Sensem
- Desarrollo
- **Situación actual**

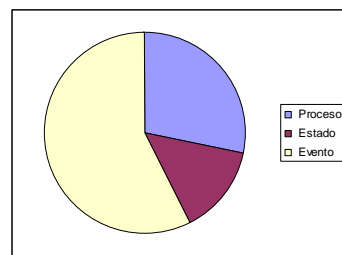


Situación actual

- el 99'6% del corpus está anotado
 - 249 verbos,
 - 25.064 oraciones
 - 376.907 palabras anotadas
 - 836.573 palabras corpus total (anotación + contexto)
- El 42% revisado
- En el momento de la finalización del proyecto : 60% del corpus revisado



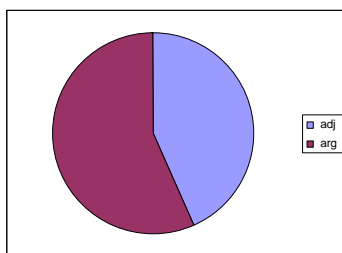
Estructura eventual



Proceso 28'3%
Estado 14'3%
Evento 57'4%



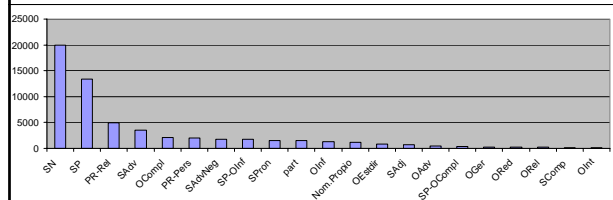
Argumentos/adjuntos



Adjuntos 43'4%
Argumentos 56'6%



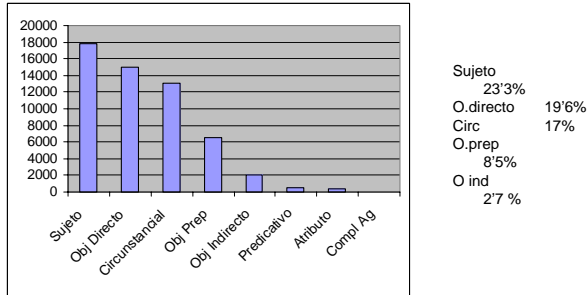
Categorías



SN 26%
Sp 17'5%
Prel 6'5%
Sadv 4'6%

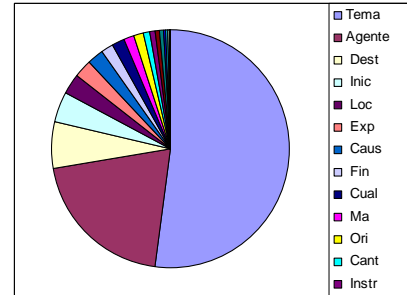


Funciones



GRIAL

Papeles semánticos



GRIAL

Situación actual

Semántica de la construcción

- Antiagentiva 6'68%
- Anticausativa 2'6%
- Pasiva 1'8%
- Habitual 0'4%
- Impersonal 0'9%

- Reflexiva 0'4%
- Recíproca 0'6%

GRIAL

Situación actual

Tareas actuales

- Finalización de la anotación
- Corrección
- Adquisición subcategorización
- Desarrollo de la interfaz de acceso a la base de datos verbal

GRIAL

Trabajo futuro

Anotación

- Estructura eventual
 - Sintagma verbal
 - Oracional

Corrección

Adquisición

- Semántica oracional
- Restricciones selectivas
- Preposiciones

GRIAL

Algunas publicaciones

- Vázquez, G., A. Fernández, L. Alonso (2005) Description of the Guidelines for the Syntactico-semantic Annotations of a Corpus in Spanish". Angelova, G., K. Bontcheva, R. Mitkov, N. Nicolov (ed.), /International Conference Recent Advances in NaturalLanguage/, Shoumen (Bulgaria); p. 603-607.
- Alonso, L., J.A. Capilla, I. Castellón, A. Fernández, G. Vázquez (2005). "The Sensem Project: Syntactico-Semantic Annotation of Sentences in Spanish", Proceedings of the International Conference RANLP, p. 39-46. Borovets, Bulgaria.
- Castellón, I., A. Fernández, G. Vázquez (2005). "La semántica oracional del español: perspectiva desde el léxico". G. Wojtak, J. Cantero (ed.), Entre semántica léxica, teoría del léxico y sintaxis. Frankfurt/Leipzig. Peter Lang, Europaischer Verlag der Wissenschaften, p. 113-122.
- Alonso, L., I. Castellón, N. Tincheva (2006) "Detección automática de errores en el corpus Sensem" RESLA (en prensa).
- Fernández, A., G. Vázquez, D. Teruel (2006) "Interfaz de explotación del corpus SenSem" RESLA (en prensa).

GRIAL

El proyecto SenSem



<http://grial.uab.es/proyectos/sensem>

