

REPRESENTACIÓN SIMBÓLICA DE PREGUNTAS: EL PROYECTO RESIM

GLÒRIA VÁZQUEZ GARCÍA

gvazquez@dal.udl.es

Universitat de Lleida

IRENE CASTELLÓN MASALLES

email???

Universitat de Barcelona

ANA FERNÁNDEZ MONTRAVETA

email???

Universitat Autònoma de Barcelona

Resum. **Títol en català.** Resum en català.

Paraules clau: paraules clau en català

Abstract. **Títol en anglès.** Resum en anglès.

Key words: paraules clau en anglès

Resumen

En este artículo presentamos el proyecto ReSim: Representación simbólica de preguntas, que ha consistido en la creación de un sistema automático capaz de representar en un lenguaje formal una pregunta en lenguaje natural. Dada la complejidad del objetivo, el proyecto se centró en un ámbito temático concreto. El tipo de representación usada está basada en el modelo de las Estructuras Léxico Conceptuales de Jackendoff 1990.

1. Introducción

La respuesta a preguntas es actualmente una de las aplicaciones en las que más esfuerzo se está dedicando dentro del procesamiento de lenguaje natural. Uno de los referentes dentro de esta línea de investigación son las

conferencias MUC¹ celebradas anualmente. Los sistemas de búsqueda de respuestas suelen tener una arquitectura variada, sin embargo, a grandes rasgos, podríamos establecer una arquitectura general para estos sistemas. Suelen dividirse en tres módulos principales: el análisis de la pregunta, la localización de la respuesta mediante la selección de los textos o segmentos relevantes susceptibles de incluir dicha respuesta y, por último, la extracción de ésta (Hirshman & Gaizauskas 2001, Moldovan et al. 2002).

En el proyecto que presentamos se ha trabajado en relación con el primer módulo. En algunos de los trabajos realizados para el castellano (Vicedo 2003, Massot et al. 2003) se revela que el punto débil de estos sistemas es que, por un lado, hasta el momento, no se dispone de un analizador completo de esta lengua con resultados aceptables en el ámbito de las oraciones interrogativas y, por otro, es difícil conseguir una representación neutra respecto a la sintaxis y suficientemente amplia como para poder establecer correspondencias adecuadas con los diferentes pasajes susceptibles de considerarse respuestas.

La contribución de ReSim² en este campo se centra especialmente en la representación semántica, que va más allá de aquellas basadas únicamente en predicados simples. El tipo de representación escogida ha sido el de las Estructuras Léxico Conceptuales (ELC, Jackendoff 1990), que presenta como ventaja el hecho de que es un modelo semántico que permite establecer relaciones con la sintaxis. Además, este tipo de representación es menos dependiente léxicamente que otros, lo cual permite expresar a un nivel de abstracción más elevado los significados y, de esta manera, tratar adecuadamente la sinonimia verbal y oracional. Así, en ReSim no solamente se representan del mismo modo los predicados equivalentes, sino también aquellos fragmentos que léxica y sintácticamente difieren pero que expresan el mismo significado.

En el apartado siguiente, se expone con más detalle el tipo de representación semántica utilizada. En el apartado 3, se presenta el estudio del corpus de oraciones interrogativas realizado, perteneciente al dominio de la Lingüística Computacional. Los recursos lingüísticos utilizados y la arquitectura del sistema se presentan en los apartados 4 y 5, respectivamente. Finalmente, se exponen los resultados y las conclusiones.

¹http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html

²ReSim es un proyecto MCyT BFF2003-06456

2. Representación semántica

En ReSim, las oraciones interrogativas se representan mediante ELC, en las que se expresan los elementos léxicos de las oraciones asociados a etiquetas semánticas que forman parte de un metalenguaje de primitivos que forman una estructura jerárquica en constituyentes semánticos. Las ELC que se toman como punto de partida están asignadas a los verbos y/o a patrones sintácticos. Para la declaración de las diferentes ELC ha sido necesario adaptar y modificar las estructuras propuestas por Jackendoff.

Veamos cuál es el tipo de representación propuesta para el verbo *aparecer* en ReSim:

$$[\text{event GO}_{\text{exist}} ([\text{thing }], [\text{path TO}_{\text{exist}} ([\text{prop TRUE}])])]$$

Como se puede observar, las ELC se expresan mediante una estructura de predicados y argumentos y contienen diferentes elementos para representar el significado:

- Constituyentes semánticos: son porciones de significado marcadas a través de corchetes.
- Primitivos (GO, TO): aparecen en mayúsculas y son elementos básicos en la descomposición del significado que se quiere representar, por lo que actúan de núcleos predicativos de los constituyentes semánticos.
- Campos semánticos (exist): especifican el ámbito temático en el que deben interpretarse los primitivos.
- Tipos semánticos (event, thing, path, prop): identifican los constituyentes semánticos.
- Constantes (TRUE): representan valores conocidos.

Una de las ventajas de esta representación es que permite enlazar directamente con el nivel sintáctico. Ello nos permite asociar a aquellos constituyentes semánticos que se expresan sintácticamente información sobre su forma y su relación semántica con el verbo. Así, las ELC se han enriquecido añadiendo información más específica sobre el tipo de sintagma, la función y el papel semántico que caracteriza cada constituyente, ya que esta información será necesaria en la fase de creación de la representación. La ELC anterior ya ampliada tendría la forma siguiente:

[event GO_{exist} ([thing ** \$SN \tic\],[path TO_{exist} ([prop TRUE])])]

En esta representación se puede observar que se indica la función sintáctica de sujeto (** representa la función de sujeto), su realización (\$SN) y se expresa el tipo de papel semántico \tic\ (tema incremental de creación). Véase Fernández et al 2004 para un listado exhaustivo de los papeles usados.

Además, otra de las ventajas mencionadas respecto al uso de ELC es que permiten representar de la misma forma oraciones con diferente estructura formal pero una semántica equivalente. Como podemos observar en (1), las oraciones de (2), que en un determinado dominio tienen la misma significación, se representarán de igual forma a nivel semántico, es decir, la descomposición de los predicados lleva a una misma formulación de significados:

(1) [state BE_{ident} ([thing],[AT_{ident} ([prop])])]

(2) ¿Qué significa composicionalidad?

¿Qué es la composicionalidad?

Esta representación indica que las oraciones de (2) son estativas y que en ellas se relaciona, por tanto, un objeto con una propiedad a través del predicado BE, que aglutina el significado básico de los dos verbos (*significar* y *ser*). El esquema es el mismo que en una localización física pero con el subíndice “ident” se especifica que dicho primitivo debe interpretarse en el ámbito de la identificación. La localización se entiende, por tanto, metafóricamente, y de ahí que se utilice el primitivo AT para expresar la situación del objeto respecto al universo de propiedades existentes en el mundo.

En este caso, como se observa en (3), la sintaxis de ambas expresiones es también la misma, ya que ambos verbos (*significar* y *ser*) requieren dos SN. Además, la función semántica de estos constituyentes en las dos oraciones también es coincidente: el SN sujeto es considerado tema (tg) y el SN objeto (atributo) expresa la propiedad con la que se identifica (id):

(3) [state BE_{ident} ([thing \$SNS\tg\],[AT_{ident} ([prop \$SN2\id\)])])]

Por último, presentamos la representación resultante de las oraciones de (1) al hacer la correlación con la sintaxis:

(4) [state BE_{ident} ([thing **qué**],[AT_{ident} ([prop **composicionalidad**])])]

3. El corpus de oraciones interrogativas

Para poder llevar a cabo la construcción del motor ha sido necesaria la constitución de un corpus de preguntas, que se ha utilizado para la construcción de las distintas bases de datos y para la prueba del sistema.

El corpus inicial se constituyó a partir de 100 oraciones y se confeccionó con la colaboración de los estudiantes de la asignatura de Lingüística Computacional, impartida en la Universitat de Lleida. Este corpus se creó con las aportaciones de preguntas libres de dichos alumnos sobre dudas en relación al temario del curso, por lo que se trata de oraciones en lenguaje natural producidas por humanos sin restricciones. Consideramos que ello supone una ventaja en tanto que nos ha permitido trabajar con oraciones reales y, por lo tanto, construir un sistema más cercano a los patrones de la comunicación humana usual. No obstante, no se ha podido respetar siempre el texto original y, en ocasiones, ha sido necesario redefinir o eliminar algunas preguntas, o bien por su complejidad sintáctica o bien por su inadecuación semántica, obteniendo así un corpus de 73 oraciones.

A partir de aquí se procedió a la ampliación del conjunto de preguntas objeto de estudio utilizando otro corpus, cedido por el grupo TALP de la Universitat Politècnica de Catalunya, mucho más extenso (2.000 oraciones), con el fin de incluir más patrones interrogativos y que fuera representativo de la variedad sintáctica y semántica del español. Siguiendo esta metodología se añadieron unos 200 patrones más, que posibilitaron multiplicar considerablemente el número de oraciones a analizar, ya que las combinaciones posibles entre dichos patrones son numerosas. Desde el punto de vista léxico, dicho corpus se ha ampliado también a partir de la incorporación de sinónimos verbales y nominales. El conjunto final se obtiene de todas las combinaciones posibles entre los patrones interrogativos y los léxicos.

Una vez constituido el conjunto de oraciones objeto de análisis, se procedió a elaborar un estudio sintáctico-semántico del mismo, que consistió en la clasificación de las distintas oraciones en función de los siguientes parámetros:

- El tipo de interrogativa (total, parcial).
- La presencia o ausencia de elementos interrogativos.
- La semántica básica de las interrogaciones (causa, definición, diferencia, etc.).
- La semántica de los elementos interrogativos (manera, finalidad, lugar, etc.).

- Dicha clasificación nos permitió diseñar los distintos recursos del sistema y el tipo de arquitectura.

4. Las bases de datos

Para implementar el sistema se ha desarrollado una serie de bases de datos lingüísticos que interactúan en la construcción de la representación semántica. Concretamente, se han construido cuatro bases de datos principales que contienen información léxica, sintáctica y semántica:

- Base de datos verbal. **BDV**: esta base contiene las ELC **verbales**. La unidad de representación en esta base de datos es el sentido, de forma que se agrupan las unidades verbales que comparten una misma ELC (p.e., *conseguir 1* y *obtener 1*).
- Base de datos de esquemas interrogativos indirectos. **EInd**. En esta base se han recogido fragmentos, es decir, concatenaciones de palabras que se consideran **no relevantes** para la semántica de la interrogativa (*me gustaría saber si . . . , es verdad que...*). Su función es señalar que lo que viene a continuación es una pregunta.
- Base de datos de esquemas interrogativos verbales. **EIV**. En esta base se recogen patrones que coinciden con **oraciones completas** que tienen asignada una ELC propia. Se agrupan por significados, es decir, por funciones interrogativas que se han asignado después del estudio del corpus de preguntas recogido. Las etiquetas bajo las que se han agrupado estos esquemas son del tipo: *definición, función, diferencia, verificación*, etc. A continuación se presenta un extracto del conjunto de esquemas que expresan el concepto que hemos denominado *función* y a los cuales se les ha asociado la misma representación semántica:

FUNCIÓN
patrón
cuál SER [DET] (clase de) rol de [SN**] en [SN&]
qué (tipo de) rol TENER [SN**] en [SN&]
qué (clase de) rol TENER [SN**] en [SN&]
qué (tipo de) rol JUGAR [SN**] en [SN&]
qué (clase de) rol JUGAR [SN**] en [SN&]

La LCS asociada a estos patrones es:

```
[state BE charfunct ([thing ** \tg\],
                    [place AT charfunct ([property ? \id\)]),
                    [place IN (thing & \loc\)]]
```

Como puede observarse, los sintagmas marcados como tales en los patrones se coindexan con posiciones en la LCS mediante los símbolos ‘**’ y ‘&’. Por otro parte el símbolo ‘?’ que aparece en la LCS se utiliza para indicar la posición que ocupa el sintagma interrogativo.

- Base de datos de patrones interrogativos. **PI**. Los patrones interrogativos son **fragmentos de oración** y normalmente se corresponden con el **sintagma interrogativo**. Dichos patrones están clasificados en grupos y cada grupo tiene asociada una etiqueta semántica, que coincide con el nombre de algún rol semántico de algún constituyente de la LCS. De esta manera se establece la relación entre el elemento sobre el que se interroga y la parte de la oración, expresada en términos de LCS, con la que se corresponde. Tal y como ocurría en los EIV, los PI están agrupados en sinónimos. A continuación, podemos ver un ejemplo de la declaración de los PI correspondientes al rol causa:

pi-causa
patrón
por qué (@CAUSA@)
a causa de qué
cuál (@MODAL@) SER [DET] @CAUSA@ por ([DET]) que
cuál (@MODAL@) SER [DET] @CAUSA@ de que

Los elementos marcados entre @ hacen referencia a listas de palabras especificados en otras BD. Así, además de estas 4 bases, que son básicas para la construcción de la representación semántica, se han construido también 2 bases de datos complementarias. La primera es una de verbos modales y auxiliares (p. e., *poder*, *tener que*), que se utiliza para ampliar los patrones que contienen verbos, previendo la posibilidad de que vayan acompañados de un modal o auxiliar, como se observa en los dos últimos PI del ejemplo anterior. La segunda es una base de nombres que nos permite relacionar diferentes sustantivos en función de la relación de sinonimia con el objetivo de conseguir también más cobertura en la fase de análisis. En el caso del ejemplo anterior la lista @CAUSA@ contiene las palabras *causa*, *razón* y *motivo*.

5. Arquitectura del sistema

El sistema se compone de dos módulos claramente diferenciados, el análisis formal y la representación semántica. En el primer módulo se realiza el análisis morfosintáctico de la pregunta. Para ello se realiza un tratamiento morfológico y sintáctico parcial (Carreras et al. 2004). Este proceso nos devuelve la oración agrupada por sintagmas, en el interior de los cuales cada palabra tiene asociada la información morfológica apropiada.

Esta gramática se enriqueció para el análisis de oraciones interrogativas con la anotación del constituyente que contiene la incógnita de la pregunta. En el análisis resultante, como puede observarse a continuación, se marca, además de la incógnita y los sintagmas relevantes para la representación semántica, el verbo, que constituye la pieza clave de nuestro análisis.

¿Por qué el corrector identifica errores inexistentes?

Análisis:

¿_Fia **grup-spq**_ [Por_sps00 snq_ [qué_pt0cs000]] **sn**_ [el_da0ms0 corrector_ncms000] **grup-verbal**_ [**identifica_vmip3s0**] **sn**_ [errores_ncmp000 inexistentes_aq0cp0] ?_Fit

Una vez se ha llevado a cabo este proceso se inicia la fase de construcción de la representación semántica (módulo de representación). Se han desarrollado cuatro posibles estrategias en la construcción de la representación semántica. Cada estrategia recoge información de una o varias de las bases descritas anteriormente. El procedimiento es el siguiente:

1. En primer lugar, el sistema etiqueta como interrogativas parciales aquellas oraciones que incluyen algún sintagma interrogativo a partir de las etiquetas obtenidas en la anotación a nivel morfosintáctico.
2. En el caso de que no exista un sintagma interrogativo en primer lugar, el sistema hipotetiza que la oración es una interrogativa indirecta, es decir, se comprueba la posible existencia de una cadena que esté recogida en la base de esquemas interrogativos indirectos (EInd). Si se detecta dicha cadena, este fragmento de la oración quedará sin representación semántica asociada, ya que está vacío de contenido. A partir de este momento se trabajará con el resto de la oración no perteneciente al esquema indirecto.

Veamos un análisis de una oración interrogativa total con esquema interrogativo indirecto inicial siguiendo las estrategias de 1 y 2:

Frase de entrada: *Me gustaría saber si existe alguna diferencia entre los sistemas de traducción asistida Deja-vu y transit*

Frase analizada: sn_[Me_pp1cs000] grup-verbal_[gustaría_vmic3s0 saber_vmn0000] si_cs grup-verbal_[existe_vmip3s0] sn_[alguna_díofs0 diferencia_ncfs000] grup-sp_[entre_sps00 sn_[los_da0mp0 sistemas_ncmp000 grup-sp-de_[de_sps00 sn_[traducción_ncfs000 asistida_aq0fsp déjã-vu_npms000 y_cc transit_npms000]]]]

No existe sintagma interrogativo en primer lugar

EIND detectado: *Me gustaría saber si*

Resto: grup-verbal_[existe_vmip3s0] sn_[alguna_díofs0 diferencia_ncfs000] grup-sp_[entre_sps00 sn_[los_da0mp0 sistemas_ncmp000 grup-sp-de_[de_sps00 sn_[traducción_ncfs000 asistida_aq0fsp déjã-vu_npms000 y_cc transit_npms000]]]]

- Posteriormente, tanto si se trata de una interrogativa directa o indirecta, se determina si la oración completa o al menos una parte importante de la misma se corresponde con algún patrón recogido en la base de EIV (esquemas interrogativos verbales). Si esto ocurre, la ELC ya está asignada al EIV correspondiente y, por lo tanto, la representación semántica final se obtiene a partir de ésta incorporando los constituyentes sintácticos en la ELC. Los constituyentes sintácticos no contemplados en la ELC, si los hay, se interpretan como adjuntos y se añaden a la representación semántica.

Veamos un análisis de una oración interrogativa parcial siguiendo los pasos mencionados hasta ahora:

Frase de entrada: *¿Qué son las memorias de traducción?*

Frase analizada: snq_[Qué_pt0cs000] grup-verbal_[vser_[son_vsip3p0]] sn_[las_da0fp0 memorias_ncfp000 grup-sp-de_[de_sps00 sn_[traducción_ncfs000]]]

Sintagma interrogativo detectado en primer lugar ->oración interrogativa parcial

EIND no detectado

EIV detectado: qué SER [SN&]

ELC asociada: [state BE_{ident} ([thing&],[AT_{ident} ([prop]])]

Roles semánticos asociados: \tg\id\

ELC resultante:

[state BE_{ident} ([thing LAS MEMORIAS DE TRADUCCIÓN \tg\],[AT_{ident} ([prop QUÉ \id\)])]

- En el caso de que no se reconozca un EIV en la oración, se busca si el patrón marcado como sintagma interrogativo está descrito en la base de patrones interrogativos (PI). Si se ha detectado un PI, es necesario marcar el fin de este patrón en la oración que se analiza para adjudicarle la interpretación adecuada en términos de papeles semánticos.

Frase de entrada: *¿Para qué se utiliza la interlingua?*
 Frase analizada: grup-spq_[para_sps00 snq_[qué_pt0cs000]] grup-verbal-pron
 _[se_p0300000 utiliza_vmip3s0] sn_[la_da0fs0 interlingua_ncfs000]
 Sintagma interrogativo detectado en primer lugar ->oración interrogativa parcial
 EIND no encontrado
 EIV no encontrado
 PI detectado: *Para qué* ->PI fin

5. El próximo paso es identificar el verbo principal y consultar en la base de verbos (BDV) la representación semántica (ELC) que tiene asignada éste. Siguiendo con el ejemplo anterior, esta acción se ejecutaría así:

Verbo: *utiliza*
 Lema: *utilizar*
 ELC *utilizar 01*³
 [event ACT ([thing \$SNS],[thing \$SN2],[event FOR (event \$SP)])]
 Roles semánticos asociados:
 SNS:inic
 SN2:fin
 SP(para):instr

A partir de aquí, se procede a la construcción de la interpretación semántica de la oración propiamente dicha, identificando los sintagmas y relacionándolos con las posiciones argumentales. Cuando el esquema es pronominal y el lema no lo es el sistema marca con NULL la posición correspondiente al sujeto.

ELC resultante:
 [event ACT ([thing NULL],
 [thing LA INTERLINGUA\instr\],
 [event FOR (event PARA QUÉ?\fin\)])]

6. En el caso de fallar los tres procedimientos anteriores (EIND, EIV y PI) el sistema asigna a la oración la clase interrogativa total. Como antes, se detecta el verbo y se recupera su representación en la BDV. En el caso de que se trate de un conjunto verbal formado por un verbo conjugado y otro en forma impersonal (infinitivo, gerundio o participio) formando una perífrasis, como en el ejemplo siguiente, se tendrá que buscar la información asociada al verbo en forma impersonal.

Frase de entrada: *¿Se puede crear un traductor infalible?*

³El sistema no realiza ninguna desambiguación semántica.

```

Frase analizada: grup-verbal-pron_[ se_p0300000 puede_v mip3s0 crear_v mn0000
] sn_[ un_di0ms0 traductor_ncms000 infalible_aq0cs0 ]
No existe sintagma interrogativo en primer lugar
EIND no encontrado
EIV no encontrado
Pi no encontrado ->oración interrogativa total
Verbo: crear
Lema: crear
ELC crear 01
[event CAUSE ([human $SNS],
[event GOexist ([thing $SN2],[place TOexist EXIST]])])
Roles semánticos asociados:
SNS:inic
SN2:tic
ELC resultante:
[event CAUSE ([human NULL],[event GOexist ([thing UN TRADUCTOR
INFALIBLE\tic\],[place TOexist EXIST]])])

```

6. Resultados

Una vez construido el sistema a partir del corpus inicial, éste se ha evaluado con un subconjunto de las oraciones que constituyen el corpus de RESIM, concretamente, con 73 oraciones interrogativas. En general el sistema es capaz de construir la interpretación semántica de la oración en un 79,4% de los casos, dejando fuera de cobertura un 20,6%. Respecto a la precisión, el 67,3% de los casos procesados es correcto mientras que el 32,7% fue considerado incorrecto.

Por otro lado, en algunas oraciones, el sistema proporciona más de un resultado, ya que no se ha llevado a cabo un proceso de desambiguación semántica verbal, por lo que se proponen todas las ELC asociadas a un determinado lema. El porcentaje de estos casos sobre el total de las oraciones a las que el sistema da respuesta es del 5%

Con respecto a la falta de cobertura (20,6%), podemos observar que el sistema no proporciona respuesta por diversos motivos que enumeramos a continuación. En algunas ocasiones no se ha asociado un patrón del tipo EIV. Esto es así, debido a que entre las cadenas que esperaba el sistema se encuentra alguna palabra o algún sintagma que no se había previsto y, por tanto, el algoritmo de unificación no tiene éxito. Otro tipo de error viene dado por el fallo en el análisis sintáctico, ya que los sintagmas que resultan de este análisis son incorrectos y, por tanto, no son los que espera el sistema. En

último lugar, la unificación puede fallar porque el sistema no ha seleccionado correctamente el verbo principal de la oración y por tanto la LCS que obtiene para unificar no presenta el mismo número de argumentos que la frase que se está analizando.

Respecto a la precisión, se da la misma casuística de errores que en la cobertura. En este caso, aunque se produce una LCS completa con elementos oracionales, el resultado obtenido no es el correcto. El error más frecuente que se ha detectado es la unificación errónea de complementos y adjuntos, ya que en algunos casos coinciden las especificaciones de ambos por lo que se refiere al tipo de sintagma y/o a la preposición, lo cual dificulta la elección.

7. Conclusiones

Hemos presentado un sistema automático que, a partir de preguntas en lenguaje natural, produce una representación semántica en términos de LCS⁴. Aunque los resultados son prometedores, queda aún trabajo por realizar para que el sistema supere sus limitaciones y pueda ser incorporado a un sistema de pregunta-respuesta.

En primer lugar, es necesario refinar la gramática de oraciones interrogativas para contemplar casos que no han sido previstos y subsanar, en la medida de lo posible, los errores detectados debidos a un análisis sintáctico erróneo.

Por otro lado, el sistema utiliza patrones basados en léxico, morfología y sintaxis, que se deben declarar a priori. Para mejorar la definición de estos patrones sería útil ampliar los mismos y mejorar su representación utilizando operadores que nos proporcionen más flexibilidad en la declaración. Ahora bien, en cuanto a la construcción de patrones, creemos que puede ser más interesante superar las limitaciones que supone la intervención manual en su creación. En este sentido, una posible mejora consistiría en la detección automática de dichos patrones mediante técnicas de análisis superficial. De esta forma, el sistema sería capaz de aprender patrones y, mediante una fase de validación humana, aplicarlos en los análisis. Creemos que esta es una línea interesante de investigación que ya hemos aplicado a otros ámbitos (Alonso et al. 2002) con resultados aceptables.

⁴<http://grial.uab.es/fproj.php?id=5>

Referencias

- [Alonso et al. 2002] L. Alonso et al. X-Tractor: a tool for Discourse Marker Extraction. *Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data, LREC'02*, pàgines???. Las Palmas de Gran Canaria, 2002.
- [Carreras et al. 2004] X. Carreras et al. FreeLing: An Open-Source Suite of Language Analyzers. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pàgines???. seu editorial???, 2004.
- [Hirshman & Gaizauskas 2001] L. Hirshman & R. Gaizauskas. *Natural Language Questions Answering: The view from here*. Cambridge University Press, Cambridge, UK, 2001.
- [Jackendoff 1990] R. Jackendoff. *Semantic Structures*. The MIT Press, Cambridge, MA, 1990.
- [Massot et al. 2003] M. Massot et al. QA UdG-UPC System at TREC-12. *Twelfth Retrieval Text Conference*, 500–255, 762. NIST Gaithersburg, USA, 2003.
- [Moldovan et al. 2002] D. Moldovan et al. LCC Tools for Question Answering. *Eleventh Text Retrieval Conference*, 500–251, pàgines???. NIST special Publication . Gaithersburg, USA, 2002.
- [Vicedo 2003] J. L. Vicedo. La búsqueda de respuestas: Estado actual y Perspectivas de Futuro. *Revista Iberoamericana de Inteligencia artificial*, VIII(22):37–56, 2003. Monográfico Acceso a la Información Multilingüe.