CENTRO INTERNACIONAL DE
ESTUDOS DE DOUTORAMENTO E
AVANZADOS DA USC

Iria
del Río Gayo

Tese de doutoramento

SpQA: un parser para análisis de preguntas en español orientado a Búsqueda de Respuestas

Santiago de Compostela, 2014

Departamento de Lingua Española. Facultade de Filoloxía

Facultad de Filología
Departamento de Lengua Española

Tesis doctoral

SpQA: un *parser* para análisis de preguntas en español orientado a Búsqueda de Respuestas

Iria del Río Gayo

Santiago de Compostela, 2013

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Facultad de Filología
Departamento de Lengua Española

Tesis doctoral

SpQA: un *parser* para análisis de preguntas en español orientado a Búsqueda de Respuestas

Iria del Río Gayo

Directora: M.ª Paula Santalla del Río

Santiago de Compostela, noviembre de 2013

Agradecimientos

Gracias a todas las personas que, de una forma u otra, han contribuido durante estos años a la realización de esta tesis. La lista es larga y todos y cada uno de vosotros estáis incluidos aquí. Disculpadme si el cansancio y el frío de noviembre me hacen olvidar alguna mención. Da igual, todos respiráis a lo largo de alguna de estas páginas.

Gracias al Departamento de Lengua Española por darme la oportunidad de realizar esta investigación mediante la concesión de una beca FPI en el 2008 y, especialmente, a mi directora por el trabajo de estos años.

Gracias a mis padres por darme siempre la libertad de elegir y su apoyo incondicional. Por esforzarse cada día y luchar por mí.

Gracias a mi hermana mayor por ser mi hermana mayor.

Gracias a todas y cada una de las personas maravillosas que he conocido en Santiago de Compostela (a alguno ya lo conocía de antes y lo había olvidado... menos mal que la sala PDI E nos volvió a reunir. Gracias por los consejos, la paciencia, lo abrazos y, como no, las cañas).

Gracias a mi familia de Santiago. Por ser eso, mi familia.

Gracias a todos los amigos y colegas que me han guiado, ayudado y animado a seguir.

Gracias a ti. Llegaste en la recta final. Pero sin tu impulso y tu fuerza no habría podido.

Y, muy especialmente, gracias a Kees Koster. Si hay alguien que ha contribuido a que esta tesis sea posible esa persona es Kees. Gracias por enseñarme tantas cosas. Por tener paciencia conmigo. Y por ser una fuerza de la naturaleza. Descansa en paz.

Santiago de Compostela, noviembre de 2013

Índice

Índice de abreviaturas

BR	Búsqueda de Respuestas
QA	*Question Answering*
RI	Recuperación de información
EI	Extracción de Información
PLN	Procesamiento del lenguaje natural
IA	Inteligencia Artificial
NLIDB	*Natural Language Interfaces to Databases*
TREC	*Text Retrieval Conference*
CLEF	*Cross-Language Evaluation Forum*
NE	*Named Entity*
NER	*Named Entity Recognition*
NEC	*Named Entity Classification*
AGFL	*Affix Grammars over a Finite Lattice*
V	Verbo
N	Nombre
PN	Nombre propio
A	Adjetivo
X	Adverbio
D	Determinante
Q	Cuantificador
ATTR	Atributo
MOD	Modificador
DET	Determinación
QUANT	Cuantificación
COMP	Comparación
AUX	Auxiliar
SUBJ	Sujeto
OBJ	Objeto directo
IOBJ	Objeto indirecto
PRED	Predicativo
CIRC	Circunstancial
PC	Complemento preposicional
QSUBJ	Sujeto interrogado
QOBJ	Objeto directo interrogado
QIOBJ	Objeto indirecto interrogado
QPRED	Predicativo interrogado
QPC	Complemento preposicional interrogado
QCIRC	Complemento circunstancial interrogado
NEG	Negación

Índice de Tablas

Resumen

Un sistema de Búsqueda de Respuestas (BR) permite a un usuario realizar una pregunta en lenguaje natural y obtener automáticamente una respuesta correcta y concisa a esa pregunta. La BR es una tarea compleja que implica la puesta en marcha de diversos procesos interdependientes desde que el usuario plantea una pregunta hasta que el sistema recupera una respuesta. Estos procesos se estructuran en tres fases o módulos:

1) Análisis y comprensión de la pregunta.
2) Análisis de información de la fuente de conocimiento y selección de fragmentos susceptibles de contener la respuesta.
3) Selección, extracción y generación de la respuesta.

Los sistemas de BR presentan distintos grados de comprensión del lenguaje natural en el que está codificada la información que manejan: desde aproximaciones más cercanas a la RI que conciben los textos como *bag of words* (BOW), a sistemas con complejas representaciones semánticas de la pregunta y sus posibles respuestas. Una serie de argumentos parecen apoyar que el procesamiento del lenguaje en BR no debe ser superficial, especialmente, en el análisis de la pregunta y en la fase de selección de la respuesta. Estos argumentos se sustentan, además, en una serie de experimentos llevados a cabo en el área.

Teniendo en cuenta estos argumentos y la compleja relación lingüística existente entre preguntas y respuestas, en este trabajo se defiende un modelo de BR en el que se maneje un conocimiento lingüístico profundo, idealmente, un modelo en el que poder utilizar representaciones semánticas del lenguaje, tanto de las preguntas como del texto del que se debe extraer la respuesta, además de poder operar con inferencias y razonamiento lógico. Este planteamiento, sin embargo, parece inviable por el momento, debido a que no es posible derivar representaciones semánticas completas de los textos (las que pueden derivarse, además, suelen ser muy costosas en términos de tiempo de procesamiento), y a que las bases de conocimiento que serían necesarias como fuente de inferencias no tienen la cobertura necesaria hoy en día. Por esta razón, en este trabajo se defiende que, al menos, los sistemas de BR deben manejar una representación lo más completa posible (desde el punto de vista lingüístico) de las preguntas. Las preguntas son estructuras cortas cuyo procesado, incluso a un nivel complejo, no debería provocar problemas de eficiencia (en términos de tiempo) para el sistema de BR. Por otra parte, aunque las preguntas son estructuras lingüísticas cortas, debido a la especial relación semántica que se da en el par pregunta-respuesta, contienen una gran cantidad de información que es clave para que el sistema de BR pueda interpretar los intereses del usuario y encontrar respuestas correctas. De hecho, la relación pregunta-respuesta provoca que la fase de análisis de la pregunta determine de forma crítica el éxito o fracaso de todo el proceso del sistema de BR: si la pregunta planteada por el usuario se interpreta de forma errónea, difícilmente el sistema podrá encontrar una respuesta correcta.

Exprimir al máximo las posibilidades informativas de las preguntas es, por lo tanto, esencial en BR.

Entre estas «posibilidades informativas» de la pregunta, las relaciones sintácticas entre los elementos son muy relevantes, especialmente en el caso de las preguntas parciales en las que tenemos un interrogativo complejo. Experimentos llevados a cabo en el área de BR demuestran que el uso de información sintáctica supone una mejora significativa en los resultados de los sistemas. De hecho, prácticamente todos los sistemas de BR que realizan un procesamiento lingüístico de cierto nivel utilizan información sintáctica. Además, muchos sistemas que manejan representaciones semánticas de las preguntas construyen estas representaciones sobre representaciones sintácticas.

La mayoría de los sistemas de BR utilizan para el análisis sintáctico *parsers* de tipo general, pese a que se ha demostrado que la eficacia de estos *parsers* disminuye al utilizarlos en dominios específicos. Esto se ha demostrado, en particular, para el análisis de preguntas, tanto en inglés como en español. Por esta razón, algunos autores defienden la necesidad de *parsers* específicos para el análisis de preguntas.

Esta es la propuesta de este trabajo: la construcción de un *parser* diseñado para el análisis de preguntas en español en un entorno de BR lingüísticamente motivado, SpQA (*Spanish Parser for Question Answering*).

Como no existen marcos teóricos que describan cómo construir un *parser*, la metodología seguida para la construcción de SpQA parte del objetivo del analizador: el análisis de preguntas en un entorno de BR lingüísticamente motivada. Por esta razón, SpQA se construye, por una parte, a partir de las necesidades del análisis de preguntas en BR y, por otra, a partir de un estudio lingüístico del funcionamiento de las preguntas y la relación pregunta-respuesta. Estos dos ámbitos confluyen en la gramática formal a partir de la que se genera el *parser* SpQA.

La gramática formal está escrita en el formalismo AGFL. Teniendo en cuenta que está orientada al análisis de preguntas, la gramática es más simple en los módulos generales (tipos de frases, tipos de cláusulas subordinadas, etc.), y se centra en el módulo de las oraciones interrogativas. En dicho módulo, se distinguen tres tipos de interrogativas, cada una con una representación diferente: totales, parciales y disyuntivas. Para las parciales, además, se formalizan en la gramática una serie de valores semánticos que ponen de manifiesto el significado de la frase interrogativa.

SpQA se genera a partir de esta gramática formal. El análisis que lleva a cabo el *parser* se representa en forma de grafo dependencial. El grafo, simple y compacto, permite la extracción de tripletes de dependencias y recoge información de tres niveles:

1) Léxico: etiquetado de la clase de palabra para cada una de las palabras en el grafo.

2) Sintáctico.

a) Identificación y etiquetado de las dependencias sintácticas de la pregunta (a nivel de la frase y de la oración).
b) Identificación de los límites de los constituyentes oracionales y de su función sintáctica.

c) Normalizaciones sintácticas («despasivización», marcado de estructuras impersonales, normalización en el formato de las fechas, etc.).

3) Semántico.

a) Identificación, a partir de la estructura sintáctica, de la variable interrogativa o incógnita presente en la pregunta (diferenciación entre preguntas totales, parciales y disyuntivas).
b) Especificación de valores semánticos concretos de la variable en las parciales, construidos a partir de información léxica y sintáctica (que permite, en ciertos casos, el tratamiento automático de la paráfrasis en la frase interrogativa).
c) Identificación de entidades nombradas, estructuras cuantificativas y estructuras temporales.
d) Posibilidad de establecer la diferenciación entre preguntas abiertas y preguntas informativas a partir del análisis semántico del interrogativo.

La evaluación intrínseca de SpQA muestra que el *parser* alcanza una eficacia aceptable en aquellos objetivos para los que ha sido diseñado, aunque el análisis de errores hace explícito que son necesarias mejoras en algunos ámbitos (modificación en la frase nominal; asignación de función sintáctica en pares como sujeto/objeto).

De cara a un futuro, se plantean varios frentes de trabajo para SpQA: subsanación de los errores detectados en la evaluación realizada, integración y evaluación en un sistema de BR, incorporación de otras estructuras lingüísticas que sirven para demandar información (peticiones de información), e integración de información semántica a la representación del grafo dependencial.

Abstract

A Question Answering (QA) system automatically retrieves brief, suitable answers to questions posed by a user in natural language. As a complex task, QA consists of a combination of several interdependent processes. These processes, which are brought into action when the user asks a question, operate together to obtain a relevant answer by developing the following steps:

1) Question analysis.
2) Analysis of the source and subsequent selection of fragments of information that may contain the answer.
3) Selection, extraction and generation of the answer.

QA systems manage different degrees of linguistic knowledge: since systems that work with approaches closer to Information Retrieval, that conceive texts as bags of words (BOW) to systems that use complex semantic representations of texts. A number of arguments nowadays support a deep processing of the natural language in QA systems, especially when it refers to question analysis and answer extraction. Recent research on QA has shown that the use of linguistic knowledge can improve the performance of QA systems.

Taking into account these previous aspects, as well as the complex linguistic relation between questions and answers, this dissertation stands up for the use of deep linguistic knowledge in QA. Starting with the premise of an ideal model of QA that would consider and implement semantic representations of language, inferences and logic reasoning – an approach that seems not likely in the short term, given the present state of the art – the purpose of this dissertation is to encourage and to justify the use of a complex representation of questions in QA systems. On the one hand, questions are short structures whose processing is not that complex and expensive (in terms of time) for a QA system. On the other hand, questions also contain a lot of information about the answer – thanks to the special semantic relation existing in question-answer pairs. As a result, this information could be crucial to improve the efficiency QA system: since question processing is the first step in the QA procedure and, a wrong understanding of the asked question will almost certainly prevent the system to find a suitable answer. That is why the exhaustive and accurate retrieval of linguistic information is essential to provide a proper analysis and interpretation of the question in QA systems.

Syntactic relations are highly relevant to the interpretation of the question, especially in WH-questions built on a complex interrogative sentence. Experiments in QA have shown that the use of syntactic information significantly improves the results of QA systems. As a matter of fact, most QA systems nowadays use syntactic representation and some of them take advantage of these representations to build semantic representations of questions.

Most QA systems use general parsers, in spite of some studies have shown that general parsers accuracy drops out of domain. This fact has been shown particularly in question analysis, in English as well as in Spanish. For this reason,

some researchers suggest the use of specific parsers designed for question analysis in QA, which is the main purpose of this dissertation: to build a parser specifically designed for question analysis in QA, SpQA (*Spanish Parser for Question Answering*).

Since there are no theoretical approaches that describe how to build a parser, the methodology used for the development of SpQA is based on the goal of the parser: question analysis in linguistic motivated QA. Consequently, SpQA is built according to the needs of question analysis in QA and considering the grammatical and semantic features of questions. These are the main two aspects that constitute SpQA.

The formal grammar of SpQA is written in AGFL formalism. Given that the grammar is oriented to question analysis, it is simpler in general modules (types of phrases, types of subordinated clauses) and it focusses in questions module. In this module, three types of questions are distinguished: yes/no questions; WH-questions and disjunctive questions. The grammar also describes semantic values for the interrogative phrase in WH-questions.

SpQA is generated from this formal grammar. The parser analysis is represented in a dependency graph. This graph is simple and compact. It allows the extraction of dependency triples and shows information of three linguistic levels:

1) **Lexical**: part of speech tagging.
2) **Syntactic**:
   a) Identification and labeling of syntactic dependencies.
   b) Identification and labeling of syntactic constituents.
   c) Syntactic normalizations (depasivization, normalization of date format, etc.).
3) **Semantic**.
   a) Question classification (yes/no question; WH-question; disjunctive question).
   b) Identification of the semantic value of the interrogative phrase in WH-questions.
   c) Named Entity, quantities and dates recognition.
   d) Classification of questions in open or informative questions using the analysis of the interrogative phrase.

An intrinsic evaluation of SpQA reveals that the parser reaches an acceptable level of accuracy in the values it was designed for. However, error analysis shows that several aspects of SpQA, such as the modifications in NP phrases or the labeling of pairs of functions (like subject/object) might need to improve.

Future research should approach challenges like the enhancement of the errors identified during the evaluation as well as the effective integration and subsequent evaluation of SpQA in a real QA system, the inclusion in the formal grammar of new linguistic structures that are used to demand information and the incorporation of semantic representations in the dependency graph.

Introducción

En la sociedad actual, la creciente cantidad de información disponible en formato electrónico, bien en la web, bien en otros medios, demanda tanto por parte del usuario medio como del profesional un acceso y manejo cada vez más sofisticados. Aunque toda esa información se guarda en formato digital, gran parte de ella está codificada en lenguaje natural. Sin embargo, las máquinas que la almacenan y tratan no entienden el lenguaje natural. Eso crea limitaciones y problemas a la hora de gestionar y acceder a esa información. Por estas razones (entre otras), existe en el mundo actual una necesidad cada vez mayor por acercar la interacción hombre-máquina al lenguaje humano. En este contexto, el objetivo de la Búsqueda de Respuestas (BR), obtener de un sistema automático respuestas concisas y claras a necesidades concretas de información, se muestra como una tarea de plena actualidad.

En la comunicación humana, una de las formas básicas mediante la cual las personas accedemos al conocimiento es haciendo preguntas. Uno hace una pregunta, generalmente, porque desea llenar un vacío informativo. La respuesta, siempre y cuando satisfaga las expectativas del que pregunta, cubrirá ese vacío. En la BR, un usuario realiza una pregunta en lenguaje natural a un sistema y este recupera de forma automática una respuesta correcta a esa pregunta. En el proceso que va desde que el usuario plantea la pregunta hasta que el sistema recupera la respuesta, hay distintas fases que plantean diferentes problemas técnicos. En esta serie de fases, la primera y una de las más cruciales es el procesamiento de la pregunta planteada.

La investigación en BR se inició en los años 60, con interfaces en lenguaje natural para bases de datos con información especializada. Desde entonces, el área ha experimentado un gran desarrollo, especialmente al amparo de las conferencias TREC. En el «modelo TREC», las preguntas son básicamente de tipo factual y el enfoque próximo al de la Recuperación de Información (RI). Este modelo ha marcado en gran parte el marco de desarrollo de la BR desde los 90 y, solo en los últimos años, otros aspectos como la interacción usuario-máquina o la semántica han empezado a cobrar importancia.

Los sistemas de BR presentan distintos grados de comprensión del lenguaje natural en el que está codificada la información que manejan: desde aproximaciones más cercanas a la RI que conciben los textos como *bag of words* (BOW), a sistemas con complejas representaciones semánticas de la pregunta y sus posibles respuestas. Una serie de argumentos parecen apoyar que el procesamiento del lenguaje en BR no debe ser superficial (cf., entre otros, Fliedner, 2007; Lavenus, Grivolla, Gillard, y Bellot, 2004; o Moldovan, Pasca, Harabagiu, y Surdeanu, 2003), especialmente, en el

análisis de la pregunta y en la fase de selección de la respuesta. Estos argumentos se sustentan, además, en una serie de experimentos (por ejemplo: Hovy, Hermjakob, Lin, 2001; Surdeanu, Ciaramita, y Zaragoza, 2008) llevados a cabo en el área.

Por otra parte, aproximaciones a la BR como la de las conferencias TREC presentan una serie de problemas y limitaciones (Sutcliffe, 2010; Marco De Boni, 2004; Erbach, 2004; Lin, Quan, Sinha, Bakshi, Huynh, Katz, y Karger, 2003), como la falta de precisión de la mayoría de los sistemas (con planteamientos de RI se recuperan muchas respuestas y rápido, pero gran parte de estas respuestas son incorrectas), la falta de interacción con el usuario o el tipo muy limitado de preguntas (factuales) al que se restringen, que provoca serias dificultades a la hora de procesar preguntas más complejas como las del denominado «BR Avanzado» (Saint-Dizier, y Moens, 2011).

Frente a aproximaciones que manejan el lenguaje de forma superficial, en este trabajo se defiende un modelo de BR basado en conocimiento lingüístico. En este modelo, lo ideal sería manejar representaciones semánticas del lenguaje, tanto de las preguntas como del texto del que se debe extraer la respuesta, además de poder operar con inferencias y razonamiento lógico (Fliedner, 2007). Este planteamiento, no obstante, parece inviable por el momento, fundamentalmente por dos razones: la primera es que, con la tecnología actual, no es posible derivar representaciones semánticas completas de los textos (las que pueden derivarse, además, suelen ser muy costosas en términos de tiempo de procesamiento); la segunda es que las bases de conocimiento que serían necesarias como fuente de inferencias no tienen la cobertura necesaria hoy en día (Fliedner, 2007).

Si bien el procesamiento lingüístico profundo de grandes cantidades de texto parece inviable por el momento, no ocurre lo mismo con el procesamiento de la pregunta. El procesamiento de la pregunta es el primer paso en el sistema de BR y una fase clave, pues implica comprender y determinar cuál es la necesidad informativa concreta del usuario, para poder así buscar una respuesta adecuada en la base de conocimiento.

Las preguntas son estructuras breves. Un análisis lingüístico profundo es, por tanto, viable. Además, las preguntas guardan una especial relación semántica con sus respuestas, de manera que, cuanto más completo sea el conocimiento del significado de la pregunta, más posibilidades tendrá el sistema de encontrar una respuesta correcta. Se ha demostrado, además, que el uso de herramientas lingüísticas mejora los resultados en el procesamiento de la pregunta en los sistemas de BR (cf., por ejemplo: Carvalho, de Matos, y Rocio, 2010).

Dentro del análisis lingüístico de la pregunta, el análisis sintáctico juega un papel necesario. El análisis sintáctico permite delimitar los constituyentes en la pregunta y sus relaciones estructurales. Esta información se muestra en muchas ocasiones como indispensable a la hora de comprender el significado de la pregunta y recuperar una respuesta correcta (cf. Fliedner, 2007; Bouma, Mur, Noord, y Groningen, 2005; o Li y Roth, 2006). Para Bouma et al. (2005), el análisis sintáctico es especialmente interesante en el procesamiento de la pregunta en relación a dos aspectos: determinar el núcleo de frases interrogativas complejas y establecer

propiedades adicionales de la pregunta. Por otra parte, la mayoría de los sistemas de BR que funcionan con representaciones semánticas construyen esas representaciones a partir de un análisis sintáctico anterior.

Pese a su importancia, el análisis sintáctico ha recibido escasa atención en el mundo de la BR. La mayoría de los sistemas utilizan *parsers* generales para realizar ese análisis, aunque se ha demostrado que los *parsers* generales presentan problemas al utilizarlos en dominios específicos (Gildea, 2001; McClosky, Charniak, y Johnson, 2006; Foster, 2010) y, más específicamente, en el análisis de preguntas (cf. Hermjakob, 2001; Petrov, Chang, Ringgaard, y Alshawi, 2010, para el inglés o Gayo, 2011a; 2011b, para el español). Por esta razón, varios especialistas han llamado la atención sobre la necesidad de utilizar *parsers* específicos para el análisis de preguntas (Fliedner, 2007; Katz y Lin, 2003).

Todo lo expuesto nos lleva a la propuesta de este trabajo: la construcción de un *parser* específico para el análisis de preguntas en español en un contexto de BR, SpQA (*Spanish Parser for Question Answering*).

SpQA está diseñado para una tarea concreta: extraer de una pregunta dada toda la información sintáctica y cierta información semántica, representándola en un grafo de dependencias simple y compacto. La información sintáctica es la concerniente a las dependencias sintácticas de la pregunta y sus funciones. La información semántica consiste en definir el valor de la incógnita presente en la pregunta. La representación de SpQA es fruto de un análisis léxico sintáctico que no se vale de fuentes externas de conocimiento (como Wordnet).

No existen modelos teóricos que describan cómo abordar la construcción de una gramática formal. En la mayoría de los *parsers* no estadísticos[1], no existe un marco teórico que justifique la cobertura de la gramática (es decir: de qué construcciones se da cuenta o no y por qué). Por esta razón, a la hora de plantear la metodología de trabajo para la construcción de SpQA, lo que se ha tenido en cuenta es el objetivo de la gramática y el marco teórico en el que se encuadra dicho objetivo. El objetivo de SpQA es el análisis de preguntas en un contexto de BR, y el marco en el que se inserta dicho objetivo es el de la BR lingüísticamente motivada. Por esta razón, el *parser* se construye teniendo en cuenta las necesidades de un sistema de BR y a partir de un estudio lingüístico del funcionamiento de las preguntas y la relación pregunta-respuesta. El estudio lingüístico tiene un doble componente: una parte teórica en la que se analizan distintos aspectos concernientes a la gramática y la semántica de las preguntas y una parte de análisis de corpus en la que esos aspectos se estudian en diferentes tipos de preguntas[2]. SpQA, por lo tanto, se construye sobre tres elementos:

1) **Las necesidades de un sistema de BR en el análisis de preguntas.** Este aspecto se trata en el capítulo 1. Para ello, definimos en primer lugar la Búsqueda de Respuestas, su evolución y su funcionamiento, para, a continuación, investigar el uso de conocimiento lingüístico en el área. Concluimos que la BR debe manejar, al menos idealmente, un procesamiento del lenguaje que permita acercarse lo más

1 Los *parsers* estadísticos se construyen a partir de corpus anotados, de manera que las construcciones gramaticales que en ellos se contemplan son aquellas anotadas en dichos corpus.
2 Un planteamiento similar se utilizó para la gramática formal que se describe en Gayo (2010).

posible al significado, tanto de la pregunta como de sus posibles respuestas (cf. *supra*).

A continuación, nos centramos en el módulo de análisis de las preguntas en BR y el análisis sintáctico. En el análisis de la pregunta en BR, lo fundamental es determinar al máximo cuál es la información que se demanda, para así proceder a la búsqueda de una respuesta adecuada en la base de conocimiento. En la tradición de BR, los siguientes puntos se han mostrado como útiles para determinar esa información: identificación del tipo de respuesta esperada (generalmente, una entidad relacionada con la frase interrogativa en las preguntas parciales; *sí/no* en las totales); identificación en la pregunta de entidades (entidades nombradas, cantidades, etc.) y eventos (fechas, lugares, etc.), así como de las relaciones que estas entidades mantienen entre sí (sintácticas y/o semánticas); clasificación de la pregunta.

Teniendo en cuenta la importancia del análisis sintáctico para la comprensión de la pregunta (cf. *supra*), la mayoría de los sistemas de BR utilizan esta técnica, valiéndose de tripletes de dependencias para expresar las relaciones sintácticas. Como decíamos, además, en muchos casos, si se realiza análisis semántico, este se construye sobre el análisis sintáctico previo. En gran parte de los casos, el análisis se realiza con *parsers* generales, pese a que se ha demostrado la falta de eficacia de este tipo de analizadores generales en el análisis de preguntas (cf. *supra*). Frente a este planteamiento, se propone SpQA como *parser* específico diseñado para el análisis de preguntas en español en un contexto de BR. SpQA lleva a cabo un análisis sintáctico de tipo dependencial, del que se pueden extraer tripletes de dependencias; clasifica la pregunta de acuerdo a su variable interrogativa (como total, parcial o disyuntiva); determina un valor semántico para la frase interrogativa en las preguntas parciales; reconoce entidades nombradas y fechas.

2) **Un estudio lingüístico de las preguntas y de la relación pregunta-respuesta.** En consonancia con el planteamiento de una BR lingüísticamente motivada, SpQA se construye sobre un estudio teórico de las características gramaticales y semánticas de las preguntas (y de la relación pregunta-respuesta) en español, que se presenta a lo largo del capítulo 2. El estudio teórico se plantea desde una perspectiva general para, finalmente, recoger de él aquellas características interesantes para BR que son además susceptibles de formalización en una gramática.

En dicho estudio, en primer lugar se definen las preguntas como peticiones de información (punto de vista pragmático) que tienen la forma de oraciones interrogativas directas (punto de vista gramatical). A continuación, se presentan las características gramaticales y semánticas de las preguntas así entendidas, tratando de discernir cuáles de estas características son clave a la hora de establecer el significado de una pregunta (rol del foco en la caracterización de la incógnita presente en la pregunta; funcionamiento de las partículas interrogativas en las parciales; posibles órdenes de constituyentes; etc.). Identificadas esas características, finalmente se determina cuáles de ellas son interesantes para BR y susceptibles de formalización en la gramática sobre la que se asienta SpQA.

3) **Un estudio de corpus.** En el capítulo 3 se presenta un estudio de corpus cuyo objetivo es determinar la importancia de los rasgos lingüísticos señalados en el

capítulo 2 en el funcionamiento de preguntas reales.

Para ello, el estudio se vale de tres corpus de preguntas compilados *ad hoc* a partir de tres fuentes: preguntas de corte factual extraídas de las conferencias CLEF[3], preguntas tipo *quiz* extraídas de la web[4] y preguntas de usuarios reales extraídas del portal de preguntas y respuestas Wikirespuestas[5].

La metodología utilizada para el estudio de las preguntas varía dependiendo del rasgo lingüístico analizado: en aquellos casos en los que es posible, se analizan todas las preguntas de los corpus (por ejemplo, para estudiar la incidencia de la negación); en aquellos en los que no, el análisis se realiza sobre una selección al azar de preguntas (por ejemplo, en el estudio semántico de las partículas interrogativas).

El análisis de corpus aporta datos sobre la incidencia de los rasgos analizados en preguntas reales y permite descubrir nuevos rasgos no contemplados en la teoría (otras ordenaciones de constituyentes; otros valores semánticos de las partículas interrogativas; uso de los tipos de negación en las preguntas; etc.).

En el capítulo 4, describimos como todo este conocimiento sobre las preguntas y su tratamiento en BR, cristaliza en la gramática formal que genera el *parser* SpQA. Como se orienta al análisis de preguntas, la gramática (escrita en el formalismo AGFL) es más simple en algunos módulos generales (tipos de frases; oraciones subordinadas; etc.), y se centra en el módulo de las interrogativas.

En el módulo de las interrogativas se distinguen tres tipos de preguntas a partir del valor de la incógnita presente en la pregunta: totales (se etiquetan como *ynQ* y en ellas la incógnita afecta a toda la interrogativa y supone la afirmación o negación de lo que en ella se dice), parciales (se etiquetan como *whQ* y en ellas la incógnita afecta a la frase interrogativa) y disyuntivas (*disjQ*, la incógnita afecta a la disyunción). Como el ámbito de acción de la incógnita es distinto en cada una de estas preguntas, su representación en SpQA también lo es. En el caso de las preguntas parciales, se establecen además una serie de valores semánticos para la frase interrogativa, construidos automáticamente a partir de información léxica y sintáctica. Ciertos valores semánticos son comunes para distintas construcciones sintácticas, lo que permite además (de forma limitada) el tratamiento de la paráfrasis en SpQA.

La información codificada en la gramática se presenta en la salida del *parser* en forma de grafo que expresa relaciones de tipo dependencial, aunque también es posible derivar de él los constituyentes oracionales y sus funciones sintácticas. El análisis que muestra el grafo ofrece información sintáctica (dependencias al nivel de la frase y de la oración) y semántica (valor de la incógnita presente en la pregunta), en una representación simple y compacta. De este grafo pueden derivarse además los tripletes de dependencias presentes en la pregunta.

En el capítulo 5 se realiza una evaluación de SpQA. La evaluación es de tipo intrínseco y mide la eficacia del *parser* en relación a distintas variables concernientes al análisis sintáctico y semántico, utilizando un corpus de preguntas construido *ad hoc* y el esquema de evaluación *PARSEVAL* (Black et al., 1991). Por una parte, se mide el comportamiento del *parser* en el reconocimiento de estructuras como

3 http://www.clef-initiative.eu//
4 Preguntas extraídas el 08/05/12 de: http://platea.pntic.mec.es/jescuder/pregunta.htm
5 http://respuestas.wikia.com/wiki/WikiRespuestas

preguntas, junto a la clasificación de estas en las tres categorías que se contemplan en SpQA: preguntas totales, parciales y disyuntivas. A nivel sintáctico, se mide la eficacia de SpQA en identificación y asignación de función a constituyentes oracionales y dependencias. A nivel semántico, se evalúa la capacidad del *parser* para asignar un valor semántico a la frase interrogativa de las parciales, así como el reconocimiento de entidades nombradas y estructuras de valor temporal. Los resultados generales de la evaluación son positivos en todos los ámbitos citados. Además de la evaluación, se realiza un sucinto análisis de errores. Dicho análisis señala ciertas necesidades de mejora, tanto en el nivel de identificación de dependencias sintácticas (SpQA muestra problemas, principalmente, en el módulo de la frase nominal), como en la asignación de funciones (confusión sujeto/objeto, por ejemplo).

De cara al futuro, se abren varias vías de trabajo para SpQA. El primer paso lógico consiste en la corrección de los errores detectados en la evaluación. Solucionado este aspecto, se plantea como necesaria una evaluación del *parser* integrado en un sistema de BR. Además de estos dos aspectos, se sugieren como vías interesantes de trabajo la posibilidad de añadir a la gramática de SpQA otros tipos de peticiones de información (tras un análisis de su incidencia en el uso y de su funcionamiento) además de la incorporación de información semántica relativa al verbo y los argumentos del grafo dependencial.

Capítulo 1

Búsqueda de Respuestas y conocimiento lingüístico: SpQA

1.1 La Búsqueda de Respuestas

En esta primera sección presentamos una visión panorámica de los sistemas de Búsqueda de Respuestas (BR) con el fin de definir qué es la BR y cómo funciona. Para ello, en la sección 1.1 introduciremos el concepto de BR; a continuación, en la sección 1.2 trazaremos una cronología de desarrollo de la BR, desde sus inicios hasta la actualidad; en la sección 1.3 presentaremos los tipos de sistemas de BR; finalmente, en la sección 1.4 definiremos la arquitectura básica de los sistemas de BR así como algunas de las técnicas más utilizadas en dichos sistemas.

1.1.1 ¿Qué es la Búsqueda de Respuestas?

La Búsqueda de Respuestas es una tarea que consiste básicamente en responder de forma automática preguntas planteadas en lenguaje natural a un sistema. Por lo tanto, en términos generales, un sistema de BR es aquel que permite a un usuario realizar una pregunta en lenguaje natural y obtener automáticamente una respuesta correcta y concisa a esa pregunta (Hirschman, y Gaizauskas, 2001).

Los sistemas de BR constituyen un tipo especial dentro de los sistemas de Acceso a la Información (*Information Access Systems*). Los sistemas de Acceso a la Información se definen como estructuras que permiten el acceso a grandes cantidades de información de distintos tipos: texto, imágenes, sonido, etc. El objetivo de estos sistemas puede ser muy diverso: desde resumir automáticamente la información hasta clasificarla. A continuación presentamos sucintamente distintos tipos de sistemas de Acceso a la Información con el objetivo de componer una visión general y situar adecuadamente los sistemas de BR:

- **Sistemas de Recuperación de Información** (*Information Retrieval Systems*): los sistemas de Recuperación de Información (RI) buscan en colecciones de texto los documentos que mejor encajen con una consulta (*query*) que el usuario plantea al sistema (Baeza-Yates, y Ribeiro-Neto, 1999). Las consultas consisten en una serie de palabras clave (*keywords)*. Un ejemplo típico serían los buscadores de Internet como Google.

- **Sistemas de Extracción de Información** (*Information Extraction Systems)*: instancian patrones predeterminados que describen fragmentos (*chunks*) de información muy concretos de documentos textuales (Gaizauskas, Hepple, y Huyck, 1998). Es el caso, por ejemplo, del reconocimiento en texto de patrones de entidades nombradas (*Named Entity Recognition,* NER) para nombres de personas u organizaciones.

- **Sistemas de Búsqueda de Respuestas** (*Question Answering Systems*): los sistemas de BR buscan en bases de conocimiento respuestas concretas a las preguntas que un usuario formula en lenguaje natural (Hirschman, y Gaizauskas, 2001). La salida de un sistema de BR es una respuesta concisa a la pregunta del usuario, en lugar de un documento o un pasaje.

- **Interfaces de Bases de Datos en Lenguaje Natural** (*Natural Language Interfaces to Databases, NLIDB*): pueden considerarse un subtipo de sistema de BR. Lo que particulariza los sistemas *NLIDB* es que las bases de conocimiento que utilizan son bases de datos estructuradas.

Teniendo en cuenta lo anterior, los sistemas de BR pueden verse como una evolución de los sistemas de RI: los sistemas de RI recuperan los documentos o fragmentos de documentos que mejor se ajustan a una consulta formulada como una serie de palabras clave. Los sistemas de BR, en cambio, recuperan respuestas concisas y concretas a una pregunta del usuario generalmente formulada en lenguaje natural.

En palabras de Maybury (2004, p. 3), la BR es:

> an interactive human computer process that encompasses understanding an user information need, typically expressed in a natural language query; retrieving relevant documents, data or knowledge from selected sources; extracting, qualifying and prioritizing available answers from these sources; and presenting and explaining responses in an effective manner.

Como puede deducirse de la definición de Maybury, la BR es una tarea compleja que implica una serie de procesos interdependientes que se van poniendo en marcha desde el planteamiento de la pregunta del usuario hasta la obtención de una respuesta por parte del sistema. El flujo de trabajo de un sistema BR se estructura generalmente en tres pasos básicos:

1) Comprender la consulta del usuario y, en consecuencia, concretar al máximo cuál es su necesidad de información.
2) Analizar la base de conocimiento disponible en busca de la/s posible/s respuesta/s a esa necesidad de información.
3) Extraer y presentar al usuario la respuesta correcta, en caso de que la haya.

Como veremos en el apartado 1.4, estos tres pasos determinan la arquitectura de la mayoría de los sistemas de BR.

1.1.2 Breve repaso a la historia de la BR

Inicios

La investigación en torno a los sistemas de BR se originó en el área de la Inteligencia Artificial (AI) en los años 60. Los primeros sistemas de BR surgieron con el desarrollo de interfaces en lenguaje natural para bases de datos que contenían información especializada de un área concreta (Simmons, 1965). Se trataba por tanto de sistemas *NLIDB* que, en la mayoría de los casos, se basaban en técnicas de *pattern-matching* (cf. Androutsopoulos, Ritchie y Thanisch, 1995, para una visión general del estado de la cuestión de estos sistemas iniciales). Los dos ejemplos más famosos son BASEBALL (Green, 1961) y LUNAR (Woods, 1972). BASEBALL respondía preguntas sobre datos de la liga americana de béisbol en un período de un año, mientras que LUNAR respondía preguntas sobre el análisis geológico de piedras traídas de la luna por las misiones *Apollo*. Ambos sistemas eran muy precisos en sus dominios: se llegó a demostrar que LUNAR podía contestar un 90% de las preguntas de su dominio planteadas por usuarios no entrenados en el sistema.

Un ejemplo especial entre estos sistemas iniciales lo constituye la arquitectura desarrollada en los años 70-80 por Wendy Lehnert, QUALM (Lehnert, 1977; 1978; 1980). A diferencia de los sistemas anteriores, QUALM utilizaba razonamiento basado en reglas para responder preguntas sobre historias escritas en lenguaje natural. Lehnert utilizaba para la comprensión de las historias un tipo de representaciones basadas en los trabajos de Roger Schank sobre *scripts* y *plans* (Schank, 1977). Estas representaciones se denominaban *Conceptual Dependency (CD) networks* (Schank, 1972; 1973). En QUALM se derivaban representaciones *CD* de los textos y de las preguntas del usuario. Utilizando una tipología de preguntas que combinaba características sintácticas, semánticas y pragmáticas, las representaciones de las preguntas podían ser después interpretadas reconociendo, por ejemplo, una pregunta como *¿Tienes hora?* como una petición educada de información y no como una interrogativa total demandando una respuesta de tipo *sí/no*. La búsqueda de la respuesta se realizaba emparejando la representación de la pregunta con la representación del texto.

Los trabajos de Lehnert son especialmente interesantes porque en ellos aparecen las primeras reflexiones acerca de las características que debe cumplir un sistema de BR. En su planteamiento, estas características eran principalmente tres:

- Entender la pregunta del usuario.
- Buscar la respuesta en una base de datos de conocimiento.
- Componer la respuesta para presentarla al usuario.

Nótese que estas tres tareas configuran la arquitectura básica de cualquier sistema de BR (cf. sección 1.4). El sistema ideado por Lehnert debía integrar, por tanto, técnicas de procesamiento y comprensión del lenguaje natural, búsqueda de conocimiento y generación de lenguaje (Díaz, 2009).

BR de dominio no restringido: el modelo TREC

En los años 80 y 90, estas aproximaciones iniciales construidas sobre bases de conocimiento alcanzaron sus límites porque no podían generalizarse de un modo simple a dominios de conocimiento textual no restringidos (Engelmore, 1993). Se inicia entonces la fase de desarrollo de la BR de dominio no restringido operando sobre texto (*Open-Domain QA*).

Ya en los años 90, MURAX (Kupiec, 1993) fue uno de los primeros sistemas de BR que buscaba respuestas en una colección de documentos. MURAX utilizaba la versión electrónica de una enciclopedia para responder preguntas del juego *Trivial Pursuit*, y presentaba ya por entonces el diseño general de los sistemas de BR que describiremos en la sección 1.4.

A finales de los 90 se produce un importante hito en el desarrollo de los sistemas de BR: la investigación en BR penetra en el campo de la Recuperación de Información a través de dos vías:

- las campañas de evaluación para sistemas de BR de la *Text Retrieval Conference*[6] (TREC) para el inglés, patrocinadas por el *American National Institute* (NIST) y la *Defense Advanced Research Projects Agency* (DARPA);

- el *Cross-lingual QA track* del *Cross-Language Evaluation Forum*[7] (CLEF) para BR multilingüe.

El diseño general de las evaluaciones TREC se mantuvo más o menos constante desde 1998 hasta 2007, año en que la BR desaparece de las conferencias. A los participantes se les entregaba una colección de documentos y una lista de preguntas, y cada uno de ellos debía usar su sistema de BR para responder a las preguntas y entregar los resultados para ser puntuados comparativamente. El tipo de BR de las conferencias TREC es de dominio no restringido operando sobre colecciones de texto. El enfoque, el del área de la RI (López, Uren, Sabou, y Motta, 2011): la tarea de BR consiste en encontrar el texto que contiene la respuesta a una pregunta y extraer esa respuesta. De hecho, en un principio la BR se planteó en estas competiciones como una subtarea de RI: los sistemas de BR en TREC tenían que devolver fragmentos extraídos de los documentos que contenían la respuesta a una pregunta dada. Estos fragmentos eran en un principio de 250 bytes, más tarde de 50 bytes. Solo en el 2001 se estableció la obligación de extraer la respuesta exacta.

Las preguntas en las primeras competiciones TREC eran de tipo factual: preguntas con una respuesta breve que consistía en información fáctica, del tipo:

6 http://trec.nist.gov/
7 http://www.clef-initiative.eu//

(1a) *¿Cuál es la capital de España?*
(1b) *Madrid.*

En los años posteriores, se fueron introduciendo preguntas factuales más realistas y también «preguntas tipo lista» (*list-type questions*), preguntas cuya respuesta consiste en una serie (una lista) de ítems:

(2a) *¿Qué gases forman el aire?*
(2b) *Oxígeno, nitrógeno y argón.*

En TREC 2001 (Voorhees, 2004), se introduce el concepto de BR de tipo contextual como una subtarea del *QA track* en la que el sistema debía responder series de preguntas sobre un mismo tópico manteniendo la información contextual. En (3) tenemos un ejemplo de una de esas series de preguntas (Voorhees, 2004).

(3)
Which museum in Florence was damaged by a major bomb explosion in 1993?
On what day did this happen?
Wich galleries were involved?
How many people were involved?
Where were these people located?
How much explosive was used?

La interacción con el usuario va un paso más allá en la *TREC complex Interactive QA (cIQA) task*, parte del *TREC QA track* del 2006 (Dang, 2006). En esta tarea los sistemas tenían que ser capaces de interactuar una vez con el usuario en una interacción usuario-sistema-*feed back*, de manera que el sistema utilizara ese *feed back* para aportar al usuario una respuesta nueva y mejorada.

Los mejores sistemas de TREC-8 (Voorhees, 2004) combinaban técnicas de RI para recuperar fragmentos de texto que compartían contenido con la pregunta junto con reconocimiento de entidades (NER) para identificar y extraer de la colección de documentos el tipo de entidad interrogado en la pregunta. Con este tipo de estrategias[8], los mejores sistemas eran capaces de responder correctamente a dos de cada tres preguntas de tipo factual (Voorhees, 2004).

En la evolución de las conferencias se puede observar un progresivo intento por añadir a los sistemas conocimiento lingüístico sobre la estructura de las preguntas, si bien desde el 2001 se manifiesta la preferencia por una aproximación de procesamiento superficial basado en técnicas de RI (Verberne, 2001).

Los tres mejores sistemas de las evaluaciones TREC (Sun, Jiang, Tan, Cui, Chua, y Kan, 2005; Clifton, 2005; Cui, 2005) llegaron a responder correctamente el 60-80% de las preguntas (Harabagiu, Moldovan, Clark, Cowden, Hickl, y Wang, 2005). No obstante, la mayoría de los participantes de las conferencias se movían en un 20-35% de eficacia.

8 Una visión completa del estado de la cuestión de las técnicas de BR no restringido puede encontrarse en (Pasca, 2003).

Por su parte, las conferencias CLEF han permitido el desarrollo de sistemas de BR en otros idiomas además del inglés. Las tareas de BR en CLEF han sido, con el paso de los años, más variadas y complejas: monolingües, multilingües, operando sobre texto escrito, texto oral, etc.

En cuanto al planteamiento de los sistemas de BR evaluados en CLEF, se trata de un perfil similar al que hemos visto para las conferencias TREC: BR de tipo no restringido operando sobre colecciones de texto con preguntas principalmente de tipo factual. Actualmente[9], la BR se mantiene en las conferencias CLEF encuadrada en los *tracks QA4MRE* y *QALD-3*. El primero combina BR y *Machine Reading Evaluation*, y en él los sistemas deben extraer conocimiento de grandes volúmenes de texto y usar ese conocimiento para responder preguntas. El segundo está especializado en BR sobre *Linked Data*[10].

En 2007-2008 la BR también ocupa un lugar en NTCIR[11], en el *7th NTCIR Workshop (2007/2008) Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access.*

De 2008 a 2011 la BR estuvo presente en las conferencias INEX[12] (*Initiative for the Evaluation of XML retrieval*), en la *QA@INEX*[13]. En este foro, se pretendía evaluar una tarea compleja de BR en la que se combinaba BR, recuperación de XML/párrafos y resumen automático. Para ello se utilizaban dos tipos de preguntas: un conjunto de preguntas de tipo factual, que demandaban respuestas cortas y precisas y un conjunto de preguntas complejas que podían ser respondidas a través de varias oraciones o mediante la suma de fragmentos de texto procedentes de distintos documentos.

La presencia de la BR en los foros vistos (*TREC*, *CLEF*, *NTCIR*, *INEX*), ha permitido un gran desarrollo del área desde finales de los años 90. Motivados por las conferencias, se construyeron numerosos sistemas de BR (sobre todo vinculados al mundo de la investigación) que podían desenvolver tareas progresivamente más complejas: desde responder preguntas cortas de tipo factual, a preguntas con contexto (TREC 2001) o manejar escenarios donde era necesaria la interacción usuario-sistema (TREC 2006). Las técnicas utilizadas por los sistemas también fueron evolucionando desde perspectivas más próximas a la RI (extracción y uso de palabras clave) a planteamientos con más peso de la Extracción de Información (EI) (reconocimiento de entidades, eventos y relaciones espacio-temporales) y el Procesamiento del Lenguaje Natural (PLN) (análisis sintáctico, semántico, expansión semántica de términos, etc.).

Todos estos aspectos positivos tienen, no obstante, un contrapunto negativo: conferencias como TREC y CLEF han marcado durante muchos años la pauta en el desarrollo de la BR, limitando en cierta manera el modelo de BR a «su perfil». Algunas de las críticas que se han hecho a este perfil son (Sutcliffe, 2010):

9 CLEF 2013: http://www.clef2013.org/index.php
10 http://linkeddata.org/
11 http://research.nii.ac.jp/ntcir/ntcir-ws7/ws-en.html
12 https://inex.mmci.uni-saarland.de/
13 http://www.inex.otago.ac.nz/tracks/qa/qa.asp

- **las limitaciones del tipo de preguntas usadas**: como hemos visto, las preguntas usadas eran, mayoritariamente, preguntas simples de tipo factual;

- **las limitaciones de las fuentes de conocimiento**: únicamente colecciones de texto fijas, ignorando otras fuentes de conocimiento tan importantes hoy en día como la web, ontologías, etc.;

- **la tendencia a ignorar preguntas de usuarios reales**, utilizando en su lugar preguntas artificiales;

- **la falta de atención** prestada a factores tan importantes como el **diálogo o la interacción usuario-sistema.**

Otra de las críticas que se la hecho a este tipo de BR es la falta de marco teórico. En palabras de De Boni (2004):

> While there is agreement amongst researchers on the generic aim of QA systems (presenting an answer to a question as opposed to a set of documents associated with a query) little work has been done on clarifying the problem beyond the establishment of a standard evaluation framework for QA, the Text Retrieval Conference (TREC) QA track. [. . .] Another limiting factor has been that most current research has either aimed at solving the engineering problem of building and improving systems capable of achieving high scores within this framework without questioning the solidity of the framework, or in looking at "future directions" but without having clarified the problem setting and what directions this should lead to. (De Boni, 2004, pp. 14-15).

A estas «debilidades» pueden sumárseles también las conclusiones de estudios como los de Erbach (2004) o Lin et al. (2003). En Erbach (2004) se demuestra que usuarios experimentados son capaces de responder las preguntas de *QA@CLEF* de 2003 usando Google en una media de 70 a 80 segundos; Lin et al. (2003) concluyen que los usuarios pueden estar más interesados en recuperación de párrafos que en BR de tipo factual.

Nuevas vías: la importancia de la semántica

En los últimos años han ido surgiendo nuevos campos de investigación en el área de la BR. La gran cantidad de información disponible, bien en la web, bien en otros medios, exige tanto al usuario medio como al profesional un acceso y manejo de esa información cada vez más complejos. En este contexto, el objetivo de la BR, obtener respuestas concisas y claras a necesidades concretas de información, se vuelve cada vez más necesario. Prueba de ello es el creciente interés del sector privado por la BR; en palabras de IBM (Ferrucci, 2010):

> There is growing interest to have enterprise computer systems deeply analyze the breadth of relevant content to more precisely answer and justify answers to user's natural language questions. We believe advances in question-answering (QA) technology can help support professionals in critical and timely decision making

> in areas like compliance, health care, business integrity, business intelligence, knowledge discovery, enterprise knowledge management, security, and customer support.

Por otra parte, la web semántica (WS) (Berners, 2001) es una realidad cada vez más asentada, con bases de datos que almacenan conocimiento sobre los distintos ámbitos conceptuales del mundo. El componente semántico parece ir ganando terreno frente a las aproximaciones clásicas de la RI, basadas en el uso de palabras clave (López et al., 2011). Esta tendencia se observa, por ejemplo, en el renovado interés del mercado de los buscadores hacia la introducción de la semántica en sus sistemas (Fazzinga, 2010; Hendler, 2010; Baeza, 2010).

En resumidas cuentas: el acceso a la ingente cantidad de información disponible, sumado a los continuos avances tanto en capacidad computacional como en el desarrollo de algoritmos abren nuevas e interesantes vías de investigación en el área de BR. Dos son las tendencias más claras que se observan: el manejo cada vez más sofisticado de complejas fuentes de información (estructurada y no estructurada, entre las que se incluye la web) y la relevancia dada a la semántica (que se manifiesta, principalmente, en la explotación de fuentes de información estructuradas).

En el uso de la web como fuente de información, las aproximaciones más simples constituyen una extensión del «modelo TREC» donde sistemas de BR de este perfil utilizan la red como fuente documental sumada a, o en lugar de, colecciones fijas de documentos. Para ello, estos sistemas generan consultas con palabras clave extraídas del análisis de la pregunta que envían a buscadores como Bing[14] que rastrean la web. Un ejemplo de esta aproximación lo tenemos en Mulder (Kwok, 2001), un sistema de BR para preguntas factuales que opera sobre la web enviando múltiples consultas al buscador de Google.

Un paso más allá lo constituye la BR que busca explotar la Web Semántica, denominada por algunos autores «BR semántica» (cf. López et al., 2011, para un análisis detallado de la BR de tipo semántico). Estos sistemas de BR utilizan como fuente de información bases de conocimiento que presentan una compleja organización de tipo semántico, generalmente codificadas mediante sistemas basados en lógica formal. Uno de los objetivos de dichos sistemas es precisamente salvar la brecha entre el usuario medio y las bases de conocimiento, ofreciendo la posibilidad de utilizar preguntas en lenguaje natural para extraer información. Este planteamiento viene respaldado por estudios de «usabilidad» como los de Kaufmann, Bernstein, y Fischer (2007), que demuestran que el usuario medio prefiere utilizar el lenguaje natural para interactuar con este tipo de bases de conocimiento.

Los primeros pasos de la BR dados en este sentido los constituyen los sistemas basados en ontologías (López et al., 2011). Estos sistemas presentan una interfaz en lenguaje natural que el usuario utiliza para consultar una o varias ontologías. Una de las diferencias que se apuntan entre estos sistemas y los clásicos sistemas *NLIDB* (cf. *supra*) es el componente semántico, que provee al usuario de un contexto para resolver las posibles ambigüedades, interpretar la pregunta y responderla (López et al., 2011). El creciente desarrollo de ontologías ha propiciado

14 http://www.bing.com/

esta aproximación con el fin de explotar la riqueza de estas bases de conocimiento en contraposición a larepresentación pobre de la información en planteamientos basados en palabras clave.

En los sistemas de BR orientados a ontologías el foco se pone en la portabilidad de los sistemas y su rendimiento, sustituyéndose las técnicas complejas de PLN dependientes de un dominio concreto por técnicas superficiales independientes del dominio. Algunas de estas técnicas (López et al., 2011) serían: el uso de un amplio espectro de componentes externos al sistema, tales como colecciones de tripletes (por ejemplo Sesame 5[15]) o sistemas de recuperación de información (como Lucene 6[16]); el uso de recursos independientes de dominio como WordNet[17] o FrameNet 7[18] o el uso de *parsers* como el Stanford Parser (Klein, 2003). Ejemplos de sistemas de este tipo son: QACID (Ferrández, 2009), ORAKEL (Cimiano, 2007), e-Librarian (Linckels, 2005), GINSEN (Bernstein, 2006), NLPReduce (Kaufmann et al., 2007), Querix (Kaufmann, 2006), AquaLog López, Uren, Motta, y Pasin, 2007), PANTO (Wang, 2007) o QuestIO (Tablan, 2008) (*vid.* López et al., 2011, para un análisis detallado de cada uno de estos sistemas).

Como hemos visto, los sistemas de BR orientados a ontologías son interfaces en lenguaje natural diseñadas para la consulta de bases de datos que generalmente presentan una estructuración semántica. Esto conlleva que sistemas como los citados varíen en dos aspectos:

1) **El nivel de adaptación que requieren** (directamente relacionado con su eficacia en la recuperación de información): desde sistemas diseñados por expertos para un dominio concreto (QACID) a sistemas portables (AquaLog).

2) **El nivel de lenguaje natural que son capaces de procesar**: lenguaje natural sin restricciones, gramaticalmente correcto (Querix, PANTO), lenguaje natural controlado o guiado (GINSEN), lenguaje basado en patrones (NLPReduce).

La principal limitación de los sistemas de BR orientados a ontologías es su alcance: al tratarse de sistemas diseñados para operar sobre bases de conocimiento con una estructura compleja, se ven restringidos a operar sobre un grupo limitado de dominios. Adaptarlos a un medio abierto y heterogéneo, donde un número no limitado de dominios son tratados, constituye un problema. Se trata, por lo tanto, de una limitación en las posibilidades de portabilidad y escalabilidad (López et al., 2011).

Con el fin de superar estas limitaciones, el siguiente paso en la BR semántica se dirige hacia sistemas de dominio abierto capaces de enfrentarse al contexto actual de búsqueda de información, que puede caracterizarse como: a gran escala, heterogéneo, abierto y multilingüe (López et al., 2011). El objetivo es operar sobre la web de manera no restringida, por ejemplo, sobre *Linked Data*. Un ejemplo de esta

15 http://www.openrdf.org/
16 http://lucene.apache.org/
17 http://wordnet.princeton.edu
18 http://framenet.icsi.berkeley.edu

línea de trabajo lo constituye el Workshop de 2011 *QA over Linked Data* (Unger, 2011), en el que se presentaban sistemas de BR capaces de operar sobre este tipo de datos como FREyA (Damljanovic, 2011) o Treo (Freitas, De Oliveira, O'Riain, Curry, y Pereira Da Silva, 2011). Otro buen ejemplo de este tipo de aproximaciones lo constituye el sistema PowerAqua (López, Sabou, Uren, y Motta, 2009), evolución del sistema AquaLog (López et al. 2007). Frente a AquaLog, que operaba con una sola ontología, PowerAqua es el primer sistema que desarrolla tareas de BR sobre datos estructurados en un escenario de dominio abierto.

Otra tendencia clara en la BR más actual es la combinación entre la BR y la «Representación del Conocimiento y Razonamiento» (*Knowledge Representation Reasoning*, KRR). Este tipo de sistemas de BR se asientan sobre complejas bases de conocimiento que permiten razonamiento e inferencia. Destacados ejemplos de este tipo de sistemas (de uso profesional y privado) serían: Halo (Gunning, 2010), Cyc's Semantic Research Assistant (Lenat, 2010) o True Knowledge (Tunstall-Pedoe, 2010).

La interacción usuario-sistema ha constituido otro de los focos de atención de la BR en los últimos años. En el informe conocido como *ARDA QA Roadmap* (Burger, Cardie, Chaudhri, Gaizauskas, Harabagiu, Israel, Jacquemin, Lin, Maiorano, Miller, Moldovan, Ogden, Prager, Riloff, Singhal, Shrihari, Strzalkowski, Voorhees, y Weischedel, 2001), en el que se hace un repaso al estado de la BR del momento (2001) y a sus necesidades para el futuro, dos de los desafíos que se mencionan son la BR con contexto (*Context QA*) y la BR interactiva (*Interactive QA*, IQA). Como hemos visto, el paso de preguntas sin contexto a preguntas con él aparece ya en TREC 2001. La IQA va un paso más allá y plantea un escenario en el que el usuario está interesado en mantener un diálogo real con el sistema (Bernardi et al., 2010). El nivel de interacción, no obstante, varía mucho de unos planteamientos a otros. Hemos visto como en la *TREC complex Interactive QA (cIQA) task* del 2006 (Dang, 2006), los sistemas tenían que ser capaces de interactuar una vez con el usuario, en una interacción usuario-sistema-*feed back*, de manera que el sistema utilizara ese *feed back* para aportar al usuario una respuesta nueva y mejorada. Otras aproximaciones van más allá, como el *Interactive Question Answering Workshop* (IQA Workshop) de HLT-NAACL en 2006, donde se estableció que un sistema de BR interactivo es aquel que: establece una conversación con el usuario (Strzalkowski, 2006); entiende qué está buscando el usuario y lo que ha hecho y sabe (Kelly, 2006); es un compañero en la búsqueda.

Actualmente existen también potentes sistemas de BR en el ámbito privado, como Wolphram Alpha[19] (dominio no restringido), True Knowledge[20] (Tunstall-Pedoe, 2010) (dominio no restringido) o AskHermes (Cao, Liu, Simpson, Antieau, Andrew, Cimino, Ely, y Yu, 2011) (dominio restringido: ámbito clínico). Un ejemplo aparte lo constituye el sistema desarrollado por IBM, Watson[21] (Ferrrucci et al., 2010), como parte del proyecto DeepQA[22]. Watson es el resultado del último de los «desafíos» de la compañía IBM: desarrollar un sistema automático que pudiera

19 http://www.wolframalpha.com
20 http://www.trueknowledge.com
21 http://www-03.ibm.com/innovation/us/watson/index.html
22 http://www.research.ibm.com/deepqa/deepqa.shtml

participar con humanos en el programa de televisión estadounidense *Jeopardy!*, un concurso de preguntas y respuestas de tipo enciclopédico donde las respuestas incorrectas puntúan negativo. Debido a las características del desafío, Watson fue diseñado para poder contestar, en fracción de segundos, a preguntas sobre conocimiento enciclopédico, además de valorar en cada caso si la respuesta era lo suficientemente fiable como para emitirla. Según IBM (Ferrrucci et al., 2010), la principal premisa en el diseño de Watson era crear recursos (herramientas de PLN y EI, algoritmos de RI, etc.) no orientados a una tarea específica y fácilmente adaptables y reutilizables. Por esa razón, Watson es el resultado de una compleja combinación de múltiples algoritmos y técnicas de PLN, RI, EI y KRR. En 2011, Watson logró batir a los dos mejores concursantes de *Jeopardy!*[23]

1.1.3 Desafíos futuros de la BR

Como hemos visto en el apartado anterior, el avance y desarrollo en el área de la BR ha sido constante desde sus inicios. De interfaces en lenguaje natural que operaban sobre limitadas bases de datos relacionales se ha pasado a sistemas que interactúan con el usuario y lo guían en su búsqueda
de información en variadas fuentes de información (estructurada y no estructurada).

Pese a estos avances queda, no obstante, mucho camino por hacer. Debido a la relevancia de la BR, en los últimos años varios informes y hojas de ruta realizados por reputados investigadores del área (Burguer et al., 2001), (Maybury, 2003), (Ferrucci, 2009), (Saint-Dizier, y Moens, 2011), han tratado de situar el estado de la cuestión de la BR y, sobre todo, de fijar sus necesidades/desafíos futuros. Se recogen a continuación algunos de los puntos clave para el desarrollo futuro de la BR extraídos de Burguer et al. (2001) y Saint-Dizier y Moens (2011).

Pese a ser un informe del 2001, en Burguer et al. (2001) el comité de expertos señala una serie de puntos de investigación (que se descomponen a su vez en una serie de tareas y subtareas) que siguen teniendo plena actualidad:

1) **Tipos de preguntas**: necesidad de mejores taxonomías de preguntas.

2) **Procesamiento de las preguntas**: comprensión adecuada de las preguntas (ambigüedades, implicaturas y reformulaciones). Necesidad de un modelo semántico para la comprensión y procesamiento de las preguntas que reconozca preguntas equivalentes en su significado, más allá de las posibles variaciones de tipo léxico o sintáctico.

3) **Contexto, interacción y BR**: frente al planteamiento de preguntas aisladas de TREC, se hace patente la necesidad en los sistemas de BR de operar con un contexto que permita aclarar posibles ambigüedades planteadas, tanto en la determinación de cuál es la necesidad de información del usuario como en la fase de elección de la respuesta adecuada a esa necesidad de información.

23 http://en.wikipedia.org/wiki/Watson_%28computer%29

4) **Fuentes de datos para BR**: necesidad de fuentes de conocimiento fiable a las que los sistemas de BR puedan acudir en busca de respuestas.

5) **Extracción de la respuesta**: extracción de respuestas simples y respuestas distribuidas en distintos documentos, justificación de la respuesta y evaluación de su corrección.

6) **Formulación de la respuesta**: el resultado del trabajo del sistema de BR debe presentarse al usuario del modo más natural posible. Este objetivo se hace especialmente complejo en los casos en los que la respuesta es el resultado de la fusión de información proveniente de distintas fuentes.

7) **Búsqueda de respuestas en tiempo real**: necesidad de construir sistemas capaces de extraer respuestas correctas de grandes conjuntos de datos en pocos segundos, sea cual sea la complejidad/ambigüedad de la pregunta o el tamaño y número de las fuentes de datos.

8) **Búsqueda de respuestas multilingüe**: necesidad de construir sistemas de BR para otras lenguas además del inglés y, sobre todo, necesidad de poder buscar respuestas en fuentes de información sin tener en cuenta su idioma.

Más recientemente, en Saint-Dizier y Moens (2011) se reiteran algunos puntos ya recogidos por trabajos anteriores, al tiempo que se señalan otra serie de aspectos clave para el futuro de la BR. Destacamos:

1) El concepto de BR Avanzada (***Advanced QA)***:

> Advanced question answering refers to a situation where an understanding of the meaning of the question and the information source together with techniques for answer fusion and generation are needed. [...] Very often the correct answer is distributed across several sources and the pieces have to be combined, or the right answer is found by means of inferencing. When the answer integrates different parts, a human-readable answer needs to be generated. We further refer to this type of information search as advanced question answering.
> (Saint-Dizier, y Moens, 2011)

2) La importancia de la **identificación de relaciones temporales, espaciales y causales en los textos** para la extracción/elaboración de respuestas.

3) **La representación del contenido de las fuentes de información**: necesidad de representaciones más complejas del contenido que permitan razonamiento e inferencias (un buen ejemplo sería el lenguaje conocido como *first-order predicate logic*[24]).

24 http://en.wikipedia.org/wiki/First-order_logic

4) **Tipos de preguntas**: necesidad de ampliar los tipos de preguntas manejadas por el sistema de BR. Se distinguen tres tipos de preguntas:

- **Preguntas factuales:** piden una información de tipo factual.
  (4) *¿En qué mes celebran los rusos la Revolución de Octubre?*

- **Preguntas complejas**: aquellas que pueden ser descompuestas en varias preguntas más simples.
  (5) *¿Qué edad tenía J.F.K. cuando fue asesinado?*

- **Preguntas avanzadas**: aquellas que requieren un procesamiento lingüístico y de razonamiento más complejo.
  (6) *¿Cómo convierten las plantas la luz en alimento?*

  Dentro de las preguntas de tipo avanzado se establece la siguiente tipología:

- **Preguntas de tipo procedimental**: preguntas tipo *how* cuya respuesta es un fragmento de un procedimiento (Delpech, y Saint-Dizier, 2008) (una serie de instrucciones, por ejemplo).

- **Preguntas causales**: preguntas de tipo «por qué X» cuya respuesta es en general un conjunto más o menos organizado de eventos que causan X.

- **Preguntas de tipo evaluativo y comparativo**: preguntas en las que se establece algún tipo de comparación/evaluación de términos, por ejemplo: *Is X an innovative researcher?*

- **Preguntas de opinión**: relacionadas con la minería de opinión y la argumentación probabilística. Requieren respuestas muy elaboradas donde se contrasten opiniones sobre un hecho, probablemente apoyadas por argumentos a favor o en contra de ese hecho.

Respecto a preguntas a primera vista más simples como las de tipo factual se señala además la necesidad de una mayor interacción con el usuario y en ocasiones de la fusión de múltiples respuestas en los casos en los que, pese a la aparente sencillez de la pregunta, puede que no exista una respuesta directa o tal vez existan demasiadas respuestas (cf. Saint-Dizier, y Moens, 2011, p. 903, para un ejemplo de esta situación).

5) **Evaluación**: se señala la evidente complejidad de esta tarea para sistemas de *Advanced QA*, y se apunta la necesidad de elaborar un corpus de preguntas realista, formado por preguntas planteadas por usuarios reales (frente a los clásicos sets de preguntas de TREC o CLEF). Se sugiere como una «relativamente objetiva» medida de evaluación el nivel de **satisfacción** del usuario respecto a la respuesta devuelta por el sistema (que podría establecerse por comparación con respuestas dadas por usuarios humanos en situaciones similares).

Conclusiones

Desde sus inicios a finales de los años 60 hasta la actualidad, el área de la BR ha experimentado un gran desarrollo. En sus primeros pasos dentro de la Inteligencia Artificial, la BR se limitaba a interfaces en lenguaje natural para bases de datos relacionales de dominio restringido.

El continuo crecimiento de la información disponible hizo patente la necesidad de llevar la BR hacia dominios no restringidos. Al amparo de conferencias como TREC o CLEF, la BR de dominio no restringido sobre colecciones de texto recibió un gran impulso, lo que permitió que se alcanzaran resultados muy positivos para el modelo de sistema de estas conferencias (especialmente para las preguntas de tipo factual).

En los últimos años, la disponibilidad masiva de datos, tanto en fuentes estructuradas de conocimiento (web semántica) como en fuentes no estructuradas (web, texto), la necesidad cada vez más cotidiana de acceso a esos datos por parte de los usuarios (tanto en ámbitos profesionales como no profesionales) y las limitaciones de los planteamientos tradicionales de recuperación de información basados fundamentalmente en búsqueda por palabras clave, han hecho de la BR un área de especial interés. Buena prueba de ello constituye el creciente interés de empresas como IBM. No obstante, algunos puntos básicos para el desarrollo en el área, señalados desde principios del presente siglo (Burguer et al., 2001), siguen necesitando la atención de los investigadores. Paralelamente, continúan apareciendo nuevos desafíos para la BR como la interacción con el usuario, el manejo cada vez más sofisticado del componente semántico y el acceso a las vastas y heterogéneas fuentes de información disponibles.

1.1.4 Tipos de sistemas de BR

La rápida y constante evolución en el área de la BR explica el hecho de que en la literatura especializada se engloben sistemas de características diversas bajo la etiqueta «sistemas de BR»: desde aquellos que operan con preguntas en lenguaje natural y buscan respuestas en grandes colecciones de texto, hasta aquellos que operan sobre consultas o palabras clave y buscan respuestas en la web. No obstante, como veremos en el siguiente apartado, la mayoría de estos sistemas se estructuran en torno a los tres pasos que ya hemos presentado más arriba (comprensión de la pregunta, análisis de la base de conocimiento, extracción de la respuesta) y sus diferencias vienen determinadas por factores que afectan a cada uno de esos pasos. Se presentan a continuación algunos de los factores que definen diferentes sistemas de BR (Maybury, 2003; López et al., 2011):

1) **Nivel de complejidad de la consulta del usuario/respuesta del sistema**: la consulta del usuario puede ir desde la complejidad de un diálogo con el sistema, utilizando preguntas en lenguaje natural, hasta estructuras más simples como palabras clave o frases.

A su vez, la naturaleza de la respuesta presentada al usuario puede variar desde estructuras sencillas como una entidad nombrada o una frase, hasta un párrafo

que contiene la respuesta o texto en lenguaje natural generado por el sistema, fruto de la fusión de informaciones provenientes de distintas fuentes.

2) **Fuente(s) de conocimiento para extraer la respuesta**: varía tanto el número de fuentes utilizadas, como su naturaleza: fuentes estructuradas (bases de datos relacionales, ontologías), semi-estructuradas (diccionarios, enciclopedias) o no estructuradas (colecciones de texto, la web). También varía el tipo de información codificado en esas fuentes y manejado por el sistema de BR: desde información textual (la mayoría) hasta imágenes o contenido multimedia.

3) **Dominio o ámbito sobre el que trabaja el sistema**: dominios restringidos (*Closed-domain QA*), donde el sistema opera sobre un área concreta (por ejemplo, medicina), *vs.* dominios abiertos (*Open-domain QA*).

4) **Nivel de interacción usuario-sistema**: desde sistemas en los que la interacción se limita al planteamiento de la pregunta y la obtención de la respuesta, hasta sistemas que dialogan con el usuario con el fin de obtener información que pueda aclarar ambigüedades en el proceso (principalmente en el análisis de la pregunta, aunque no solo).

5) **Técnicas utilizadas para analizar la pregunta y extraer la respuesta**: como veremos en detalle (cf. sección 2.2), las técnicas utilizadas pertenecen a áreas como la RI, la EI, o el PLN, y presentan distintos niveles de complejidad (sencillas, como la eliminación de *stop-words*, hasta complejos procesos de razonamiento espacio-temporal).

6) **Nivel de conocimiento del sistema**: además de la(s) fuente(s) de conocimiento de las que extraer la respuesta, el sistema puede utilizar otras fuentes de conocimiento en distintos procesos, como, por ejemplo, conocimiento general del mundo que permita inferencias o razonamiento lógico.

7) **El nivel de multilingüismo y *cross linguality***: sistemas que operan con un solo idioma frente a sistemas que manejan, en todas o alguna de las fases, distintas lenguas.

Podrían mencionarse también otras dimensiones como la naturaleza/expectativas de los usuarios o incluso el grado de eficacia requerido al sistema (*precision* y *recall*).

En lo sucesivo nos centraremos fundamentalmente en las características de los sistemas de BR que operan sobre texto (que es el perfil para el que está pensado SpQA), si bien también mencionaremos aspectos de otros tipos de sistemas.

1.1.5 Arquitectura básica de un sistema de BR

La arquitectura de los sistemas de BR, al margen de diferencias derivadas de dimensiones como las presentadas, es generalmente la misma y puede reducirse a

tres módulos básicos que realizan distintas tareas y que se corresponden con las tres fases ya descritas:

1) Análisis de la pregunta.
2) Análisis de información de la fuente de conocimiento y selección de fragmentos susceptibles de contener la respuesta.
3) Selección, extracción y generación de la respuesta.

Estos módulos interactúan entre sí, de manera que el éxito/fracaso de uno está directamente relacionado con el éxito/fracaso de los demás y, por ende, de todo el sistema. Así, como veremos, si, por ejemplo, el sistema de BR falla en la fase de análisis de la pregunta, no será posible obtener una respuesta al final del proceso.

A continuación se presentan en detalle los tres módulos básicos de todo sistema de BR junto con algunas de las técnicas[25] más utilizadas en cada uno de ellos[26].

1.1.5.1 Módulos de un sistema de BR

1) **Análisis de la pregunta**: se encarga del procesamiento de la pregunta planteada por el usuario. El objetivo es obtener toda aquella información de la pregunta que permita localizar, en los siguientes pasos del proceso, los fragmentos de información susceptibles de contener la respuesta.

En los sistemas de BR más básicos y próximos a la RI, la tarea fundamental de este módulo es la extracción de las palabras clave presentes en la pregunta con el fin de construir una consulta que se utilizará en los pasos siguientes.

En la mayoría de los sistemas en este paso también se lleva a cabo una clasificación de la pregunta en virtud de la respuesta esperada. La finalidad de esta clasificación es definir el tipo de entidad semántica por la que se pregunta. Por ejemplo, en un caso como

(7) *¿Cuándo fue inventado el motor de combustión interna?*

tendríamos que la entidad por la que se pregunta es una fecha. Como es lógico, este tipo de clasificación es especialmente útil en la BR de tipo factual. Este paso posibilita, por una parte, restringir la búsqueda posterior a un tipo determinado de entidades (en nuestro ejemplo, fechas). Por otro, permite que en los pasos siguientes el sistema seleccione las estrategias de búsqueda y extracción más adecuadas para esa entidad a la que la pregunta apunta.

2) **Selección y análisis de información**: a partir de la información obtenida del procesamiento de la pregunta, este módulo se encarga de seleccionar en la base de

25 En los párrafos siguientes se atenderá fundamentalmente a las técnicas utilizadas en los sistemas clásicos de BR de dominio no restringido que operan sobre texto, si bien es cierto que algunas de estas técnicas se utilizan también en otros tipos de BR como en la BR de tipo semántico.

26 Las técnicas que implican procesamiento lingüístico serán revisadas en profundidad en la siguiente sección.

conocimiento documentos o fragmentos de texto susceptibles de contener la respuesta buscada.

Las técnicas empleadas en este módulo varían dependiendo del tipo de fuente de conocimiento utilizada: colecciones de texto en los sistemas de BR de dominio no restringido, ontologías o bases de datos en la BR semántica, etc. Lo más común, no obstante, es el uso de técnicas básicas de RI: un sistema de RI (buscador) utiliza la consulta generada en el paso anterior para obtener documentos o fragmentos de texto susceptibles de contener la respuesta correcta[27].

Preprocesado de la colección de documentos

Previa a la selección de documentos, los textos de la colección suelen someterse a un preprocesado. Lo habitual en este preprocesado es eliminar las *stop-words* y realizar lematización, para posteriormente llevar a cabo una indexación de los documentos por parte del sistema de RI utilizado. Algunos sistemas, no obstante, realizan tareas más complejas de preprocesado como la etiquetación de entidades (Bouma, 2006; Neumann, 2003), o incluso la puesta en práctica de técnicas de EI para la extracción de *facts* y su almacenamiento en bases de datos que serán posteriormente utilizadas (Prager, 2000; Bouma, 2006; Cui et al., 2005; Clifton, y Teahan, 2005; Fleischman, Hovy, y Echihabi, 2003).

Selección de documentos candidatos

A partir de la consulta formulada en los pasos previos, el sistema de RI selecciona aquellos documentos susceptibles de contener la respuesta. La consulta suele enriquecerse utilizando métodos típicos de RI, fundamentalmente: técnicas lingüísticas para la expansión de la consulta (uso de sinónimos extraídos de bases de datos léxicas como Wordnet) (Qiu, y Frei, 1993) y técnicas de realimentación o *relevance feedback* (Ruthven, y Lalmas, 2003).

La mayoría de los sistemas no selecciona documentos enteros, sino fragmentos de texto (en Clarke, Cormack, Lynam, y Terra, 2006, se presenta una panorámica general de las técnicas utilizadas para la recuperación de fragmentos de texto en el área de RI).

Como hemos visto, algunos sistemas de BR no solo utilizan la colección de documentos para buscar la respuesta, sino que se valen también de Internet como fuente documental (Kaisser, y Becker, 2004; Katz, Lin, Loreto, Hildebrandt, Bilotti, Felshin, Fernandes, Marton y Mora, 2003; Katz, Bilotti, Felshin, Fernandes, Hildebrandt, Katzir, Lin, Loreto, Marton, Mora y Uzuner, 2004; Neumann, y Xu, 2003; Buchholz, y Daelemans, 2001).

Análisis de documentos/fragmentos candidatos

En este paso los documentos o fragmentos recuperados en el paso anterior son procesados. Lo habitual en este procesamiento es utilizar técnicas de EI, siendo

27 Como veremos en la siguiente sección, algunos sistemas como (Fliedner, 2007) no utilizan un sistema de RI en esta fase.

la básica el reconocimiento y marcado de entidades (NER), con el fin de identificar nombres de personas, organizaciones, fechas o cantidades. Otras técnicas más avanzadas incluyen el reconocimiento y/o extracción de eventos (Srihari, Li, y Li, 2006).

Algunos sistemas de BR incorporan a este paso un procesamiento lingüístico completo del texto recuperado en los pasos anteriores. Este procesamiento suele incluir el análisis sintáctico completo del texto, e incluso el análisis y etiquetado de *predicate-argument structures* (PAS), estructuras con un verbo y sus argumentos prototípicos junto con los roles semánticos de esos argumentos (cf. sección siguiente para más detalles).

3) **Extracción y generación de la respuesta**: este módulo utiliza la información obtenida del análisis de la pregunta y del análisis de los documentos de la colección para seleccionar una respuesta y presentarla al usuario.

Para ello, el sistema debe seleccionar una respuesta entre las posibles candidatas extraídas en el paso anterior. La selección se realiza a través de un emparejamiento o comparación entre la información extraída de la pregunta del usuario y la información presente en las respuestas candidatas. Aquella respuesta candidata que se aproxime más a «lo que busca el usuario» será la elegida. Las técnicas utilizadas para realizar ese «emparejamiento» son muy variadas; en la sección 2.2 las trataremos en profundidad.

Tras los pasos anteriores, la mejor respuesta identificada es seleccionada y presentada al usuario. La mayoría de los sistemas presentan directamente el fragmento de texto que contiene la respuesta. Pocos sistemas utilizan técnicas de generación de lenguaje para crear una respuesta (como sí se hace, por ejemplo, en Vargas-Vera, y Motta, 2004).

1.1.6 Conclusiones

En esta sección hemos presentado la Búsqueda de Respuestas, definiendo en qué consiste y mostrando su evolución y sus desafíos futuros.

Un sistema de BR es aquel que permite recuperar automáticamente una respuesta concisa a una pregunta planteada en lenguaje natural. La BR es una tarea compleja que implica diversas tareas con múltiples dimensiones; esto provoca que existan diferentes tipos de sistemas de BR, si bien todos presentan una arquitectura básica similar que se estructura en torno a tres pasos: análisis de la pregunta, análisis de la base de conocimiento y selección de respuestas candidatas y extracción de la respuesta. Hemos definido las tareas relacionadas con cada uno de estos tres pasos, así como las técnicas que se suelen emplear en cada uno de ellos. Dichas técnicas pertenecen a distintas áreas: la RI (extracción de *keywords* y generación de una consulta que se utiliza en un buscador), la EI (NER; marcado de eventos temporales, espaciales, etc.) o el PLN (análisis sintáctico y semántico).

Situados los sistemas de BR, en la segunda parte de este capítulo nos ocuparemos de todo lo relativo al manejo de información lingüística en dichos sistemas.

1.2 El uso de conocimiento lingüístico en los sistemas de BR

Hemos visto en la sección anterior que la BR es un área en la que confluyen técnicas y aproximaciones de distintos ámbitos: la Recuperación de Información (RI), la Extracción de Información (EI), la Inteligencia Artificial (IA) o el Procesamiento del Lenguaje Natural (PLN).

Frente a las aproximaciones que se construyen sobre un procesamiento superficial del lenguaje (más próximas a la RI), en este trabajo se defiende un modelo de BR motivado lingüísticamente, basado en un procesamiento lingüístico profundo, idealmente: un procesamiento semántico de pregunta y respuesta. Para conseguir ese tipo de procesamiento, se considera que el punto de partida básico es el análisis no solo semántico sino también sintáctico de pregunta y respuesta.

Por estas razones, en los apartados siguientes se profundizará en la cuestión del uso de conocimiento lingüístico en los sistemas de BR. En primer lugar, se repasarán sucintamente las distintas técnicas y recursos externos más usados en BR que manejan conocimiento lingüístico, junto a las tareas para las que son empleados. En segundo lugar, se analizarán las ventajas y desventajas del uso de conocimiento lingüístico en los sistemas de BR, y se presentarán los datos de diversos trabajos que han tratado esta cuestión.

1.2.1 Introducción: el conocimiento lingüístico en BR

Los sistemas de BR pueden manejar distintos grados de conocimiento lingüístico. Los más próximos al área de la RI apenas lo utilizan: conciben el texto que manejan (preguntas y documentos susceptibles de contener la respuesta) como *bag of words* (BOW); construyen una consulta a partir de las palabras clave de la pregunta y utilizan un sistema de RI para recuperar una respuesta. Un paso más allá en el manejo de conocimiento lingüístico lo implica el uso de técnicas de EI, que permiten el reconocimiento de «patrones» con un significado lingüístico: entidades (NER/NEC), eventos, relaciones, etc. Finalmente, las técnicas de PLN empleadas en BR presentan distintos grados de complejidad: desde la utilización de fuentes de conocimiento léxico semántico como Wordnet hasta la utilización de complejos sistemas de representación semántica de preguntas y respuestas.

Lo más habitual es que los sistemas de BR combinen el uso de técnicas de RI, EI y PLN, con mayor o menor peso de unas y otras. Como hemos visto, prácticamente todos los sistemas de BR utilizan módulos de RI (buscadores) en la fase de selección y extracción de la respuesta: primero se crea una consulta a partir del análisis de la pregunta que es utilizada por el sistema de RI para hacer una preselección de fragmentos de texto (provenientes de la colección de documentos) susceptibles de contener la respuesta. Incluso los sistemas que utilizan complejas técnicas de PLN suelen echar mano de este paso. La explicación es simple: procesar toda la colección de documentos de la base de conocimiento es una tarea que, incluso con la velocidad de los sistemas informáticos actuales, conlleva cierto tiempo. Si el tipo de procesado implica complejas representaciones lingüísticas, ese tiempo crece exponencialmente. El procesamiento que requiere un sistema de RI es, sin embargo, mucho más rápido. De ahí que la mayoría de los sistemas opten por este paso. Una

solución alternativa la ofrecen los sistemas que preprocesan *off-line* la colección de documentos, como Fliedner (2007) o Bouma (2006) (en este último se utiliza el sistema de RI como segunda opción).

La diferencia entre la mayoría de los sistemas en relación al manejo de información lingüística radica en los extremos del proceso: en el análisis de la pregunta y, especialmente, en la selección de la respuesta. En el análisis de la pregunta, la operación básica consiste en la extracción de palabras clave para la creación de la consulta que será utilizada después. A esta operación básica pueden sumarse toda una serie de procesos con más o menos implicación de conocimiento lingüístico, que resultan en algún tipo de representación de la pregunta que será utilizada en la fase de selección de la respuesta. Prácticamente todos los sistemas de BR realizan en este paso una clasificación del tipo de pregunta, siguiendo distintos criterios. Incluso sistemas que no operan con conocimiento lingüístico como Soubbotin (2001), presentan una fase de clasificación del tipo de pregunta que será la que determine los pasos a seguir por el sistema en las fases posteriores. En la fase de selección de la respuesta, se segmentan los documentos en fragmentos, buscando dentro de esos segmentos cuál de ellos encaja mejor con la consulta elaborada por el sistema. Para ello, la mayoría de los sistemas de BR **no realizan un análisis completo**, sino *keyword matching*, expresiones regulares, *part-of-speech tagging* y reconocimiento de constituyentes básicos para encontrar la respuesta. A continuación, se utilizan técnicas de clasificación para seleccionar la mejor respuesta entre las posibles obtenidas en el paso anterior. Por otra parte, en los sistemas donde se utilizan técnicas de PLN, los documentos o fragmentos seleccionados por el sistema de RI son procesados y analizados siguiendo distintas técnicas y, posteriormente, se busca en ellos el fragmento cuya representación (sintáctica, semántica, híbrida...) mejor encaje con la representación (sintáctica, semántica, híbrida...) de la pregunta obtenida en el paso inicial del proceso. Un planteamiento distinto, que se hizo popular tras la edición de TREC-10 vistos los buenos resultados del sistema que lo utilizaba, es el de Soubbotin (2001). En este sistema se utilizan patrones superficiales para la selección de la respuesta. Del análisis de la pregunta se obtienen las palabras clave para la creación de una consulta, así como una clasificación del tipo de pregunta. A partir del tipo de pregunta, se seleccionan una serie de patrones que puede presentar la posible respuesta. Esos patrones son luego utilizados en la fase de selección de la respuesta: en los documentos/fragmentos candidatos a contener la respuesta, se buscan esos patrones, y se selecciona entre las posibles respuestas aquella que presente el patrón que posee una mayor puntuación en el sistema.

Existen, por lo tanto, sistemas de BR que funcionan con muy poco conocimiento lingüístico, aplicando técnicas básicas de RI y EI, concibiendo los textos como *bag of words* y/o patrones de diversos tipos. La ventaja de estos sistemas es su planteamiento simple: su eficacia se basa en la eficacia de los algoritmos utilizados, y no necesitan de trabajo manual para la construcción de complejas arquitecturas basadas en conocimiento lingüístico.

Este tipo de sistemas ha dado buenos resultados, sobre todo, trabajando con preguntas de tipo factual donde se pregunta por una entidad. Estas preguntas suelen presentar una estructura sintáctica simple, y la mayor dificultad estriba en determinar

por qué tipo de entidad se está preguntando. Cuando el tipo de pregunta es más complejo, sin embargo, estos sistemas presentan una eficacia mucho menor. En palabras de Moldovan, Clark, Harabagiu, y Maiorano (2003):

> in TREC many systems were quite successful at providing correct answers to simpler, fact-seeking questions, but failed to answer questions that required reasoning or advanced linguistic analysis.

El mismo trabajo recoge un dato interesante: un 70% de los sistemas participantes en TREC-8 devolvieron una respuesta correcta a la pregunta Q1013:

(8) *Where is Perth?*

Sin embargo, ninguno pudo encontrar una respuesta a la pregunta Q1165, más compleja:

(9) *What is the difference between AM radio stations and FM radio stations?*

En las siguientes secciones presentaremos las técnicas y recursos que manejan conocimiento lingüístico en BR, así como las principales tareas que implican el uso de este tipo de conocimiento.

1.2.2 Tareas, técnicas y recursos que implican el manejo de conocimiento lingüístico en BR

1.2.2.1 Tareas y técnicas

Como ya se ha adelantado, las tareas y técnicas que conllevan manejo de información lingüística en los sistemas de BR son numerosas. Se consideran aquí todas las que tienen que ver con la gestión y manejo de información lingüística a cualquier nivel: desde el léxico hasta el sintáctico semántico. La consideración de «tareas y técnicas que implican el manejo de conocimiento lingüístico» es amplia, de modo que no solo se recogen las propias[28] del área de PLN, sino todas las que tienen algo que ver con procesamiento del lenguaje.

A continuación presentaremos las principales tareas y técnicas que manejan conocimiento lingüístico en BR, encuadradas en los procesos de BR que les corresponden: procesamiento de la pregunta y/o extracción de la respuesta.

28 En las secciones siguientes asociamos las distintas técnicas presentadas con aquellas áreas a las que más comúnmente se ligan: Recuperación de información (RI), Extracción de Informaicón (EI), Procesamiento del Lenguaje Natural (PLN) o Inteligencia Artifical (IA).

Procesamiento de la pregunta

Las principales tareas que se llevan a cabo son las siguientes:

1) **Extracción de palabras clave de la pregunta**: tarea típica de RI. Con estas palabras clave se elabora una consulta que será utilizada por el sistema en los pasos posteriores.

Es habitual realizar una expansión de los términos de la consulta mediante el uso de WordNet u ontologías. Para realizar esta tarea, se utilizan las siguientes técnicas propias de RI:

- **Transformación de la pregunta en una *bag of words***: conlleva pérdida de información sintáctica y jerárquica.

- **Eliminación de *stopwords***: eliminación de palabras que son consideradas poco útiles para las tareas de RI, a pesar de que estas pueden ser significativas para la interpretación de la pregunta (cf. Leveling, 2010) debido, bien a su alta frecuencia, que anula su capacidad discriminante como términos (formas verbales de *ser* o *estar*) bien a que su contenido semántico es escaso (artículos, preposiciones).

- **Lematización**: reducción de las palabras a su raíz léxica. Conlleva pérdida de información morfológica como tiempo, modo, persona, género o número.

En ocasiones también se utiliza e**tiquetado morfológico** (*Part of Speech Tagging*, POS): determinación del tipo de palabra (sustantivo, adjetivo, verbo, etc.).

2) **Identificación de entidades y eventos**: para llevar a cabo esta tarea se utilizan técnicas propias del área de la EI:

- **Reconocimiento y clasificación de entidades** (*Named Entity Recognition/Classification*, NER/NEC): reconocimiento y clasificación de estructuras lingüísticas que se refieren a entidades del mundo: personas, compañías, lugares, etc.

- **Reconocimiento y marcado de eventos**:
  Definimos un evento como cualquier tipo de situación o acontecimiento denotado por un predicado. Los eventos pueden ser *acciones*, acontecimientos llevados a cabo voluntariamente por un sujeto agente (*Los antropólogos forenses* delimitaron *el predio*); *procesos*, acontecimientos desencadenados espontáneamente (*Los árboles* están floreciendo *prematuramente por las altas temperaturas*) o acontecimientos causados por una fuerza externa al proceso (*Se supo que los fuertes vientos* derrumbaron *varios techos*); estados, situaciones que se mantienen a lo largo de un período o son permanentes (*El tránsito* está detenido *a causa de los cortes de ruta*). (Wonsever, Malcuori, y Aiala, 2012)
- **Reconocimiento y marcado de relaciones espacio-temporales**: relaciones entre eventos situadas en un tiempo y/o espacio determinados.

3) **Clasificación de la pregunta**: este es un paso muy habitual en los sistemas de BR. Las preguntas son clasificadas jerárquicamente en taxonomías (cf. Pomerantz, 2005, para un análisis de distintos tipos de taxonomías de preguntas), de manera que diferentes tipos de preguntas requieren diferentes estrategias por parte del sistema a la hora de buscar la respuesta.

Diversos trabajos en el área de la BR (Hovy, Hermjakob, y Lin, 2001; Moldovan, Clark, et al., 2003; Roth, Cumby, Li, Morie, Nagarajan, Rizzolo, Small, y Yih, 2002) han demostrado la utilidad de la clasificación de preguntas (CP) como paso previo en los sistemas de BR. Algunos estudios incluso han evaluado el impacto del uso de la CP en el resultado global del sistema de BR: en un estudio sobre errores en sistemas de dominio abierto, Moldovan, Pasca, et al. (2003) demuestra que más de un 35% de los errores se debían al módulo de CP; en Radev, Fan, Qi, Wu, y Grewal (2005), se muestra que una clasificación incorrecta de tipo de respuesta esperada utilizada como filtro anterior a la extracción de la respuesta hace que la posibilidad de encontrar una solución correcta sea 17 veces menor.

La clasificación de la pregunta puede realizarse de acuerdo con diferentes criterios. Una de las estrategias más comunes y básicas es clasificar la pregunta de acuerdo con el tipo de interrogativo que contiene (válido solo para las interrogativas parciales). Otras fórmulas establecen clases de preguntas de acuerdo con patrones léxico sintácticos, valores semánticos, etc. En la mayoría de los casos, la pregunta se clasifica teniendo en cuenta el tipo de respuesta esperada (*Expected Answer Type*, EAT). Las taxonomías de las EAT, como veremos más abajo, pueden presentar a su vez distintos grados de complejidad: desde aquellas que se establecen a partir de la información léxico-semántica del interrogativo (*quién = persona*), a elaboraciones semánticas más complejas como los *Qtargets* de la Webclopedia (Hermjakob, 2001; Hovy, Gerber, Hermjakob, Junk, y Lin, 2000; Hovy, Gerber, Hermjakob, Lin, y Ravichandran, 2001; Hovy, Hermjakob, y Lin, 2001; Hovy, Hermjakob, y Ravichandran, 2002; Ravichandran, y Hovy, 2002).

En la tarea de clasificación de la pregunta creemos que se deben distinguir dos facetas: la clasificación de la pregunta que no tiene en cuenta el tipo de respuesta esperada, a la que denominamos a continuación «categoría de la pregunta», y la clasificación de la pregunta a partir de la EAT.

Categoría de la pregunta: dependiendo del sistema, las preguntas se clasifican atendiendo a diferentes criterios: léxicos, semánticos, sintáctico-semánticos, pragmáticos... Por ejemplo, en Attardi et al. (2002), la taxonomía de preguntas se basa en las categorías de (Lehnert, 1978). Así, el tipo de pregunta para *Who invented the paper clip?* o *What did Vasco da Gama discover?* es *event-completion*. En el sistema QALC de LIMSI (Chalendar et al., 2002), el tipo de pregunta se corresponde con un determinado patrón que responde a la «forma» de esta. Por ejemplo: la pregunta *When was Rosa Park born?*, se corresponde con la categoría *WhenBePNBorn*. En Tannier y Moriceau (2009), las categorías de la pregunta responden a una tipología común en el área de BR, ya que se distingue entre preguntas factuales, de definición, *booleanas* (esperan una respuesta tipo *sí/no*), «tipo lista», etc.

Tipo de respuesta esperada (EAT): prácticamente todos los sistemas de BR clasifican la pregunta de acuerdo con el tipo de respuesta esperada. Este paso permite al sistema establecer la estrategia a seguir para encontrar la respuesta adecuada.

Las técnicas utilizadas para establecer la EAT son muy variadas, y pueden ir desde la mera identificación del interrogativo con una categoría (Litkowski, 1999) al uso de combinaciones de características léxicas, sintácticas y semánticas que definen un tipo semántico (Lavenus et. al., 2004; Hovy et al., 2000; Hovy, Gerber, Hermjakob, Lin, y Ravichandran, 2001; Hovy, Hermjakob, y Lin, 2001).

Las EAT contempladas pueden variar mucho de un sistema de BR a otro. En la mayoría de los sistemas de BR desarrollados en las conferencias TREC, estas dependen fuertemente del set de entidades que utiliza el componente de NER empleado para etiquetar la colección de documentos (Lavenus et al., 2004). A modo de ejemplo: en la edición de TREC 2002, el sistema de BR de IBM (Ittycheriah, Franz, Zhu, Ratnaparkhi, y Mammone, 2002), consideraba cinco grandes clases de entidades, que se subdividían, a su vez, en otras tantas: expresiones nominales (persona, organización, lugar, país...), expresiones temporales (fecha, tiempo...), expresiones numerales (tanto por ciento, moneda, ordinal, edad, duración...), entidades de la Tierra (tiempo atmosférico, plantas, animales...) y entidades humanas (eventos, enfermedades, cargos en compañías...). El sistema de Attardi et al. (2002) utilizaba siete categorías generales (persona, organización, localización, tiempo-fecha, cantidad, cita, lenguaje), junto a ciertas categorías específicas recogidas de la taxonomía de WordNet (Attardi, Cisternino, Formica, Simi, Tommasi, y Zavattari, 2002). Otros sistemas, sin embargo, se valían de sets mucho mayores: 48 categorías para Clarke, Cormack, Kemkes, Laszlo, Lynam, Terra, y Tilker (2002); 50 para la Universidad de Illinois (Roth, Cumby, Li, Morie, Nagarajan, Rizzolo, Small, y Yih, 2002); 54 para la Universidad de Colorado y la de Columbia (Pradhan, Krugler, Bethard, Ward, Jurafsky, Martin, Blair-Goldensohn, Schlaikjer, Filatova, Duboué, Yu, Passonneau, Hatzivassiloglou, McKeown, y Illouz, 2002).

En otros sistemas de BR el tipo de respuesta esperada no se corresponde exactamente con una entidad, sino con una categoría semántica abstracta. Es el caso de los *Qtargets* de Hovy (Hermjakob, 2001; Hovy el al. 2000; 2002; Hovy, Gerber, Hermjakob, Lin, y Ravichandran, 2001; Hovy, Hermjakob, y Lin, 2001), categorías semánticas como *why famous person* definidas mediante una combinación de características léxicas, sintácticas y semánticas de las preguntas.

Bouma et al. (2005) destaca la importancia del análisis sintáctico para determinar el tipo de respuesta demandada. En primer lugar, porque el análisis sintáctico permite establecer el núcleo de las frases interrogativas complejas. Así, ante la pregunta:

> (11) *¿Con qué organización terrorista trató de dialogar el gobierno del PSOE en 2005?*

el análisis sintáctico nos permite identificar «organización terrorista» como la entidad por la que se pregunta. En segundo lugar, porque permite establecer propiedades adicionales de la pregunta. Por ejemplo, en:

(12) *Dime el nombre de una ciudad japonesa que haya sufrido un terremoto.*

el análisis sintáctico nos indica que se pregunta por *una ciudad japonesa que haya sufrido un terremoto*, y no por cualquier *ciudad japonesa* ni mucho menos por *un terremoto*.

Como ya hemos visto, existen además herramientas específicamente diseñadas para clasificar preguntas en los sistemas de BR (bien estableciendo solo la EAT, bien combinando este aspecto con una clasificación del tipo de pregunta). Para una visión general sobre este tipo de herramientas: Díaz (2009).

4) **Representación de la pregunta**: los sistemas de BR suelen valerse de distintas estrategias para establecer una representación de la pregunta, utilizando más o menos información lingüística. Los principales tipos de representación de las preguntas que se utilizan son los siguientes:

- **Representación por medio de patrones**: estos patrones combinan información léxica, sintáctica, etc. Ya hemos mencionado el sistema de patrones superficiales para representar la pregunta de Soubbotin (2001). Otro ejemplo sería el método de Lavenus et al. (2004). En esta propuesta, las siguientes preguntas:

  (16)
  What is the collective noun for geese?
  What is the collective term for geese?
  What is a synonym for aspartame?
  What is another name for nearsightedness?
  What's another name for aspartame?
  What is the term for a group of geese?

  son consideradas equivalentes ya que pueden reducirse al patrón:

  (17) *What be [another name| a synonym| the (adj) term | noun] for GN?*

- **Representación sintáctica**: para establecer una representación sintáctica de las preguntas, los sistemas de BR se valen de parsers que realizan el análisis. La representación puede implicar distintos grados de profundidad, pero, como mínimo, conlleva el reconocimiento de frases. Generalmente, se consideran dos grandes tipos de análisis sintáctico o *parsing*:

  - **Superficial** (*shallow/light parsing*): reconocimiento de segmentos (frases o *chunks*[29]), sin establecer su estructura interna ni su rol en la oración principal. Habitual en los sistemas de BR que realizan un análisis lingüístico superficial.
  - **Complejo** (*deep parsing*): puede presentar distintos grados de complejidad, pero implica en todos los casos el reconocimiento de

29 Cf. Abney (1991).

segmentos, con su estructura interna y su rol en la oración principal. Conlleva, en algunos casos, la realización de ciertas normalizaciones sintácticas (equivalencia activa/pasiva, por ejemplo). El tipo de análisis sintáctico más utilizado por los sistemas de BR es el de tripletes de dependencias (Litkowski, 1999; 2004; Bouma et al., 2005; Bouma, 2006; Katz, y Lin, 2003a; Attardi, Cisternino, Formica, Simi, Tommasi, y Zavattari, 2002; Mollá, y Gardiner, 2004; Tannier, y Moriceau, 2009; etc.), aunque en ocasiones también se utilizan árboles sintácticos (Punyakanok, Roth, y Yih, 2004). En muchos sistemas donde se realiza también análisis semántico, este se construye a partir del análisis sintáctico, por ejemplo en Narayanan, y Harabagiu (2004b) o Salloum (2009).

- **Representación semántica**: fundamentalmente, a través del etiquetado de roles semánticos (*Semantic Role Labelling*, SRL). El análisis semántico puede tomar la forma de *predicate-argument structures* (PAS) (Narayanan, y Harabagiu, 2004; Harabagiu, 2006; Moldovan, Harabagiu Girju, Morarescu, Lacatusu, Novischi, Badulescu, y Bolohan, 2002; Moldovan, Pasca, et al., 2003; Schlaefer, Ko, Betteridge, Pathak, Nyberg, y Sautter, 2007; Ferrrucci et al., 2010; Bouma, 2006; etc.), como hemos visto, estructuras con un verbo y sus argumentos prototípicos junto con los roles semánticos de esos argumentos. Por ejemplo:

  (10) *What stimulated India's missile program?*
  PREDICATE: Stimulate
  ARG0 (role = agent)
  ARG2 (role = instrument)

  (Ejemplo tomado de: Narayanan, y Harabagiu, 2004b)

  Los PAS se construyen en ocasiones sobre un análisis sintáctico anterior como en Narayanan, y Harabagiu (2004b), aunque también existen *parsers* como ASSERT (Pradhan, Ward, Hacioglu, Martin y Jurafsky, 2004) o los descritos en Schlaefer et al. (2007) y Sun et al. (2005) que los generan directamente. En la mayoría de los casos, los PAS que manejan los sistemas se derivan de la información de PropBank (Palmer, Gildea, y Kingsbury, 2005).
  Otro ejemplo de representación semántica lo constituyen los *conceptual graphs* de Salloum (2009). En este sistema, sobre un análisis sintáctico previo se construye un análisis semántico tanto de preguntas como respuestas. Ejemplos de preguntas:

  (13) *Who invented the light bulb?*
  [Invent]-
  (Agent) -> [Person: *] ?
  (Patient) -> [Light-Bulb]

(14) *Who was assassinated on November 22, 1963, in Dallas?*
[Assassinate]-
(Agent) -> [Person: *]
(Patient) -> [Person: *] ?
(Date) -> [Date: 11/22/1963]
(Location) -> [City: Dallas]

5) **Descomposición de preguntas complejas**: otra de las tareas que se suelen llevar a cabo consiste en la descomposición sintáctica de una pregunta compleja en dos o más preguntas simples que puedan ser respondidas más fácilmente (Katz, Borchardt, y Felshin, 2005; Hickl, Wang, Lehmann, y Harabagiu, 2006; Hartrumpf, 2008; Hartrumpf, Glöckner, y Leveling, 2008; Glinos, y Gomez, 2006). De este modo, ante una pregunta compleja como:

(15) *Who is the third Republican president?*

we should first look for Republican presidents, then ask who is the third of these, rather than looking for the third president, then asking if that person is a Republican.
(Katz et al., 2005)

El análisis sintáctico es fundamental en esta tarea.

6) **Tratamiento de la paráfrasis**: ya se ha mencionado que la paráfrasis es uno de los grandes problemas a los que se enfrentan los sistemas de BR (*vid*. *infra* para más detalles).

En la fase de análisis de la pregunta, la paráfrasis se aborda generalmente en dos tareas distintas. La primera ya la hemos visto: consiste en la extensión de la consulta inicial a través del uso de términos equivalentes (extraídos, por ejemplo, de WordNet). La segunda consiste en la utilización de patrones léxico sintácticos en la representación de las preguntas (cf. *supra*) para identificar preguntas semánticamente equivalentes pero diferentes a nivel superficial (Tomuro, 2004; Bouma et al., 2005; Lavenus et al., 2004; Lin, y Pantel, 2001).

7) **Resolución de correferencia**: identificación del referente en los casos de anáfora.

8) **Desambiguación léxica** (*Word Sense Disambiguation*, WSD): concreción del significado de una palabra en casos de ambigüedad léxica.

9) **Inferencia, razonamiento lógico**: se trata de tareas que utilizan técnicas propias de la inteligencia artifical (IA):

- **Técnicas de razonamiento lógico**: utilización de técnicas que permiten la inferencia lógica.
- **Técnicas de razonamiento espacio-temporal**.

Selección y extracción de la respuesta

La mayoría de las tareas presentadas en este apartado (muchas son coincidentes con las del apartado anterior) se realizan sobre los fragmentos de texto seleccionados por el sistema de RI a partir de la consulta inicial obtenida del análisis de la pregunta (excepto en aquellos sistemas de BR que no se valen de este paso intermedio, como Fliedner (2007). El objetivo es determinar qué fragmento de texto encaja mejor con la información extraída de la pregunta (palabras clave, tipo de respuesta esperada, etc.) en la fase anterior.

Las principales tareas en la selección y extracción de la respuesta son:

1) **Etiquetado y clasificación de entidades** (NER + NEC): esta tarea se suele realizar como parte del preprocesado de la colección de documentos. Para ello, se utilizan reconocedores de entidades. El marcado y clasificación de entidades tiene sentido cuando se ha establecido en la fase anterior el tipo de entidad por el que se pregunta.

2) **Expansión semántica**: al igual que para expandir los términos de la consulta, se utilizan bases de conocimiento léxico semántico como WordNet para expandir los términos presentes en los fragmentos susceptibles de contener la respuesta.

3) **Representación sintáctica/sintáctico semántica/semántica**: representación del análisis sintáctico, sintáctico-semántico o semántico de los fragmentos susceptibles de contener la respuesta utilizando el mismo procedimiento que para el análisis de la pregunta (cf. *supra*). Para la comparación entre la representación de la pregunta y las de las posibles respuestas, se utilizan dos estrategias distintas: *matching* exacto (Katz y Lin, 2003; Litkowski, 1999; 2004), en el que las representaciones deben encajar totalmente; o utilización de métricas que calculan la distancia entre representaciones (árboles sintácticos en Punyakanok et al., 2004; tripletes de dependencias en Attardi et al., 2002; Mollá, y Gardiner 2004; y Tannier, y Moriceau, 2009).

4) **Tratamiento de la paráfrasis**: además de la expansión léxica de términos mediante WordNet, se utilizan otras estrategias como el uso de patrones léxico semánticos (cf. *supra* el ejemplo de Lavenus et al., 2004) que expresan equivalencias semánticas (Lin, y Pantel, 2001).

5) **Identificación de eventos, relaciones espacio-temporales, inferencia lógica**: se utilizan las técnicas de AI diseñadas para estas tareas sobre los fragmentos de texto susceptibles de contener la respuesta.

1.2.2.2 Recursos externos con conocimiento lingüístico

Numerosos sistemas de BR se sirven de recursos externos con el fin de obtener información, fundamentalmente, de tipo semántico. Esta información de tipo semántico puede abarcar desde cuestiones más propiamente lingüísticas como, por ejemplo, los roles semánticos con los que se construye un verbo (PAS), a conceptos

semánticos abstractos relacionados con lo que se ha denominado «conocimiento del mundo»: las categorías semánticas en las que se estructura el mundo, los elementos (palabras) que forman parte de esas categorías, así como las relaciones que se establecen entre esos elementos.

Presentaremos en primer lugar dos recursos externos de tipo lingüístico, relacionados con las tareas de SRL o construcción de PAS: FrameNet y PropBank. A continuación, presentaremos recursos de los que se obtiene información de tipo léxico semántico: una base de datos léxica, WordNet, y dos ontologías formales, Cyc y SUMO. Aunque este tipo de ontologías no son propiamente «recursos lingüísticos», las tratamos porque de ellas se extrae para los sistemas de BR información similar a la que se extrae de WordNet. Cuando las presentemos, abordaremos la cuestión de la diferencia entre ontologías formales y bases de datos léxicas como WordNet.

FrameNet (Ruppenhofer, Ellsworth, Petruck, Johnson, y Scheffczyk, 2006).

Recurso léxico semántico que define «marcos» (*frames*), descripciones abstractas de situaciones y objetos prototípicos, especifica roles semánticos para los participantes, recoge palabras que están asociadas con esos marcos, conecta los marcos a través de relaciones marco a marco y aporta ejemplos anotados extraídos de corpus (Ruppenhofer, Ellsworth, Petruck, Johnson, y Scheffczyk, 2006).

Los marcos de FrameNet posibilitan un alto grado de abstracción semántica, aspecto interesante para los sistemas de BR, fundamentalmente, para tratar la paráfrasis. Por esta razón, varios experimentos han tratado de determinar la utilidad de FrameNet en un sistema de BR (Chang, Narayanan, y Petruck, 2002; Narayanan, y Harabagiu, 2004a; 2004b). Algunos ejemplos de sistemas de BR que utilizan FrameNet son Fliedner (2007) o Narayanan, y Harabagiu (2004a).

PropBank (Palmer et al., 2005)

Corpus que consiste en los árboles sintácticos del Penn Treebank[30] enriquecidos con anotaciones semánticas del tipo *predicate-argument structures* (PAS), es decir: proposiciones verbales y sus argumentos (de ahí su nombre: *proposition bank*).

PropBank se utiliza en los sistemas de BR (Narayanan, y Harabagiu, 2004a) para anotación de roles semánticos en forma de PAS (cf. *supra*).

A continuación nos ocuparemos de una base de datos léxica, WordNet, y dos ontologías formales: Cyc y SUMO. Como ya se ha apuntado, este tipo de ontologías formales no son propiamente recursos lingüísticos. Las ontologías formales se consideran en el área de la Informática y la IA como repositorios de conocimiento conceptual que incluyen conceptos abstractos conectados a través de diferentes relaciones (Fliedner, 2007). En estas ontologías casi siempre hay un *mapping* de/hacia conceptos del lenguaje natural, pero ontología y lexicón en lenguaje natural son generalmente tratados como recursos diferentes, conectados a través de relaciones entre las palabras en lenguaje natural y los conceptos en la ontología (cf. por ejemplo el modelo detallado de Niremberg, y Raskin, 2004). La diferencia entre ambas entidades reside en aquello que consideran más importante, aquello que

30 http://www.cis.upenn.edu/~treebank/

consideran su foco: en las ontologías el foco se dirige a los conceptos abstractos y su relación en el mundo, mientras que en los recursos léxicos las palabras y sus relaciones lingüísticas constituyen el objetivo. Pese a esta diferencia, a menudo hay un considerable solapamiento entre el conocimiento en los recursos lingüísticos y las ontologías «técnicas», y los sistemas de BR acuden a ellos más o menos de igual forma[31] para obtener información de tipo léxico semántico. Por esta razón incluimos aquí dos ontologías formales junto a una base de datos léxica.

WordNet (Fellbaum, 1998)

Es uno de los recursos más utilizados por los sistemas de BR. WordNet es

> a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations[32].

WordNet es particularmente interesante para los sistemas de BR porque define relaciones semánticas entre palabras: sinónimos, hiperónimos, hipónimos, etc. Como veremos cuando tratemos las tareas de BR que utilizan conocimiento lingüístico (siguiente sección), esta información es muy útil en diferentes procesos como la expansión de términos de una consulta o la selección de la respuesta.

Suggested Upper Merged Ontology[33] (SUMO) (Niles, y Pease, 2001)

Se trata de una de las mayores ontologías formales de acceso público disponibles hoy en día. Es la única ontología formal que ha sido mapeada en su totalidad al lexicón de WordNet. Está codificada en el lenguaje formal SUO-KIF[34].

Un ejemplo de sistema de BR que utiliza SUMO (y otras muchas ontologías) lo encontramos en López et al. (2009).

Cyc[35] (Matuszek, Cabral, Witbrock, y Deoliveira, 2006)

Cyc es una gran ontología formal de propósito general. Se trata de una de las más grandes ontologías de este tipo, con una base de conocimiento que contiene 250.000 conceptos y más de dos millones de *facts* (reglas y aserciones; Matuszek et al. 2006).

En Lenat (2010) se describe un sistema de BR del ámbito clínico que utiliza esta ontología.

31 Las ontologías suelen presentar más problemas a la hora de ser usadas por los sistemas de BR porque generalmente están codificadas mediante algún lenguaje formal basado en lógica de primer orden (*predicate logic*).

32 Extraído de: Princeton University "About WordNet". WordNet. Princeton University. 2010. <http://wordnet.princeton.edu>

33 http://www.ontologyportal.org/

34 http://suo.ieee.org/SUO/KIF/index.html

35 http://www.cyc.com/

1.2.3 Uso de conocimiento lingüístico en BR: ventajas y desventajas

Una vez revisadas tareas, técnicas y herramientas relacionadas con la gestión de conocimiento lingüístico en BR, en este apartado nos ocuparemos de profundizar en los argumentos que sustentan el uso de ese conocimiento.

Como hemos visto ya, existen aproximaciones en BR que prácticamente no utilizan conocimiento lingüístico, ya que funcionan con planteamientos más próximos al área de la RI. La implementación de este tipo de aproximaciones suele ser más simple que la de aquellas que manejan conocimiento lingüístico. Además, como ya hemos mencionado, este tipo de planteamientos con menos procesamiento lingüístico suelen ser más eficientes en términos de tiempo (sobre todo, cuanto mayor es la colección de documentos que se debe procesar).

Pese a estas ventajas parece lógico pensar que a la hora de manejar el lenguaje aquellas aproximaciones que lo procesen en profundidad obtendrán resultados más satisfactorios. Nos interesa, en consecuencia, ir más allá de esta mera intuición y determinar cuáles son las ventajas concretas del uso de conocimiento lingüístico en los sistemas de BR. Por esta razón, en los apartados siguientes analizaremos en primer lugar ventajas y desventajas del uso de conocimiento lingüístico en los sistemas de BR. En segundo lugar, presentaremos una serie de experimentos que persiguen nuestro mismo objetivo: determinar objetivamente hasta qué punto el uso de conocimiento lingüístico es útil en BR.

1.2.3.1 Ventajas del uso de conocimiento lingüístico

1.2.3.1.1 La relación pregunta-respuesta

La compleja relación existente entre preguntas y respuestas (que analizaremos en detalle en la sección 3.1 del capítulo 2) es uno de los principales argumentos para el uso de información lingüística en los sistemas de BR.

Como ya hemos ido perfilando, el análisis de la pregunta permite determinar hasta cierto punto las características de la respuesta esperada. Por eso, el procesamiento de la pregunta es uno de los puntos clave en un sistema de BR: si la pregunta se interpreta erróneamente, disminuyen considerablemente las posibilidades de encontrar una respuesta correcta.

Las preguntas son estructuras lingüísticas muy breves. No obstante, pese a su aparente simplicidad, y dada la compleja relación que mantienen con sus respuestas, son estructuras cargadas de información relevante. En los sistemas de BR donde no hay interacción hombre-máquina, la información presente en la pregunta es todo lo que el sistema tiene para determinar las intenciones del usuario. Esta es la motivación principal del procesamiento lingüístico en los sistemas de BR, al menos, en la fase de procesamiento de la pregunta: **cuanto más se exploten las posibilidades informativas presentes en la pregunta, más probabilidades tendrá el sistema de encontrar una respuesta correcta**. De hecho, como hemos visto, la mayoría de los sistemas de BR dedican bastantes esfuerzos al procesamiento de la pregunta, y prácticamente todos presentan algún tipo de clasificación de esta, que generalmente contiene información lingüística y está relacionada con el tipo de respuesta esperada.

1.2.3.1.2 La información estructural

Las aproximaciones que no manejan información lingüística o que la manejan a un nivel muy superficial como, por ejemplo, aquellas que funcionan con BOW, presentan una serie de limitaciones. En primer lugar, este tipo de planteamientos conllevan **pérdida de información estructural y semántica**. Al tratar los textos solo como conjuntos de palabras independientes se pierde toda la información que aportan las relaciones que se establecen entre esas palabras y/o las unidades de las que forman parte. En palabras de Fliedner (2007):

> We have described above (2.2.2.5) that most QA systems do not use exact matching of questions and answers, but rather some 'loose' matching techniques. The simplest ones are based on identifying a named entity of the type expected as an answer and pattern matches. Thus, the approaches have no or only little information about linguistic structures. They can therefore not distinguish whether words in the question and in the potential answer stand in the same linguistic relation. Consequently, a potential problem of these approaches is that of low precision, i. e., they return wrong answers that seem to match the question, but in fact provide no answer.

Bouma et al. (2005) nos ofrece un ejemplo de las ventajas de usar relaciones estructurales frente al mero *matching* de palabras clave. Ante la pregunta:

(18a) *Wie is voorzitter van het Europese Parlement?*
(*¿Quién es el presidente del Parlamento Europeo?*)

un sistema que utiliza *keyword matching* recupera la respuesta:

(18b) *Karin Junkers (SPD), lid van het Europese Parlement en voorzitter van de vereniging van sociaal-democratische vrouwen in Europa...*
(*Karin Junkers (SPD), miembro del Parlamento Europeo y presidente de la sociedad de mujeres socialdemócratas de Europa...*)

El sistema recupera esta respuesta porque en ella están muy próximas las palabras clave de la pregunta (*presidente* y *Parlamento Europeo)* y una entidad que responde al tipo que le corresponde a la pregunta: Karin Junkers (*quién=persona*). Sin embargo, la respuesta es incorrecta, y es la información estructural la que nos lo indica.

El siguiente ejemplo de Li y Roth (2006) también es ilustrativo al respecto:

(19) *What is the speed hummingbirds fly ?*

El sistema de clasificación de preguntas de Li y Roth (2006) asigna la pregunta anterior a la clase *animal*, al considerar que el foco de la pregunta es *hummingbirds* y no *speed*. Como explican los investigadores, este error podría subsanarse utilizando información sintáctica en el análisis de la pregunta:

> The correct label is speed, but the classifier outputs animal. Our feature sensors fail to determine that the focus of the question is 'speed'. This example illustrates the necessity of identifying the question focus by analyzing syntactic structures.
> (Li y Roth, 2006)

En Katz y Lin (2003) se aborda hasta qué punto el análisis sintáctico puede resultar útil para mejorar la precisión en un sistema de BR. La conclusión de los autores es que el uso **selectivo** de información sintáctica de preguntas y respuestas puede ayudar a mejorar estos problemas de precisión en dos casos concretos:

1) Casos de **simetría semántica**: ocurre cuando las restricciones de selección semántica de diferentes argumentos respecto al núcleo de la estructura se superponen. En estos casos, el significado léxico es insuficiente para determinar el significado de la oración. Por ejemplo, en las siguientes preguntas:

(20) *What do frogs eat?*
(21) *What eats frogs?*

Sin la información estructural las dos preguntas parecen equivalentes, ya que, de forma aislada, las palabras no nos aportan suficiente información como para saber *quién come a quién* y por tanto, para saber qué se nos está preguntando en cada caso.

2) **Modificación ambigua**: ocurre cuando las restricciones de selección semántica de un argumento son tan laxas que el argumento puede ser modificador de más de un núcleo dado un contexto particular. Por ejemplo: en español casi todo puede ser *grande* o *bueno*, de manera que las restricciones semánticas que imprimen estos adjetivos son prácticamente inexistentes y no nos dan información sobre a qué elemento están modificando en el caso de aparecer en una oración con varios núcleos posibles. Es necesario el análisis sintáctico para obtener esa información. Katz y Lin (2003) ofrecen el siguiente ejemplo:

(22) *What is the largest volcano in the Solar System?*

El análisis sintáctico es necesario para determinar que *largest* está modificando a *volcano* y no, por ejemplo, a *Solar System*.

Hemos mencionado ya la importancia clave de la sintaxis en el procesamiento de la pregunta (Bouma et al., 2005), principalmente en dos aspectos (cf. *supra*):

- para determinar el núcleo de frases interrogativas complejas;
- para establecer propiedades adicionales de la pregunta.

1.2.3.1.3 La importancia del significado

Las preguntas son estructuras en las que se condensa mucha información en un enunciado muy breve. Por ello, el **manejo preciso del significado**, sobre todo en las preguntas pero también en las posibles respuestas, es crucial en un sistema de BR.

En Saint-Dizier, y Moens (2011), por ejemplo, se mencionan entre los posibles «problemas lingüísticos» a los que se tienen que enfrentar los sistemas de BR el manejo adecuado de la **polisemia** y la **homonimia**.

Sin entrar en aspectos tan complejos del significado, en Lavenus et al. (2004) se presentan varios ejemplos que demuestran la importancia del manejo adecuado de lo que allí se denomina *shades of meaning:*

(23) *What were the names of the* three ships *used by Columbus?*

La respuesta a esta pregunta debe incluir los nombres de los tres barcos usados por Colón (importancia tanto del numeral como de la relación sintáctica entre *three ships* y *used by Columbus*).

(24) *Name* a *female figure skater.*

En este caso, en cambio, el uso del artículo indefinido permite que se puedan dar muchas respuestas válidas a esta pregunta. Por eso, en la selección de la respuesta, este factor debe ser tenido en cuenta, ya que respuestas con una apariencia muy diferente pueden ser correctas.

(25) *Name* one of the major gods of Hinduism.

Las posibles respuestas están restringidas por el numeral *one* y la NP *major gods of Hinduism*. Solo tres respuestas son posibles: Brahma; Vishnu; Shiva.

(26) *Who was the* first *woman in space?*

Una respuesta que se refiera a *una mujer* en el espacio no sería válida, ya que se pregunta específicamente por *la primera*. Del mismo modo, en

(27) *What state has* the most *Indians ?*

La respuesta que se busca es muy concreta: el nombre del estado (americano) con el mayor número de indios.

Todos estos ejemplos muestran la importancia de manejar adecuadamente la información contenida en numerales (cardinales y ordinales) o en el grado de los adjetivos. En el siguiente ejemplo, la clave está en el procesamiento adecuado del verbo:

(28) *How large* is *Missouri's population?*

En (28) el usuario está buscando una información actual, ya que pregunta por la población en el momento presente, y no por la población de hace cinco o diez años.

Por otra parte, hemos visto que el manejo de conocimiento léxico semántico está presente en diversas tareas en los sistemas de BR. Es muy importante, por ejemplo, para establecer el tipo de respuesta esperada. El manejo de esta información permite determinar que en:

(29) *¿En qué país comenzó la epidemia de gripe de 1918?*

se está preguntando por una entidad de tipo geográfico, mientras que en

(30) *¿Qué empresa de automóviles hace the Spider?*

se pregunta por una entidad tipo «organización».

Paráfrasis

Hemos mencionado que uno de los fenómenos lingüísticos que juega un papel más relevante para la BR en cuanto a tratamiento del significado es la paráfrasis (Rinaldi et al., 2003), tanto en el procesamiento de la pregunta aislada como en la relación pregunta-respuesta. La paráfrasis afecta al procesamiento de la pregunta porque, como ya hemos visto (cf. *supra*), se da el fenómeno de que preguntas léxica y sintácticamente muy distintas son equivalentes semánticamente y, por lo tanto, apuntan a la misma respuesta. Como se menciona en Burguer et al. (2001), un sistema de BR debe ser capaz de identificar esta equivalencia, yendo más allá de las diferencias estructurales y llegando a la interpretación semántica subyacente. Las estrategias para lograr esto, son, como hemos visto, muy variadas: desde sistemas que manejan información lingüística muy elaborada (Hovy et al., 2000; 2002; Hovy, Gerber, Hermjakob, Lin, y Ravichandran, 2001; Hovy, Hermjakob, y Lin, 2001) hasta sistemas que se valen de meros patrones léxico sintácticos (Lavenus et al., 2004).

En la relación pregunta-respuesta, la paráfrasis juega un papel especialmente relevante. Como hemos visto, los sistemas de BR próximos a la RI utilizan los términos presentes en la pregunta para buscar la respuesta. Sin embargo, la relación entre pregunta y respuesta puede ser puramente semántica, y no léxica ni sintáctica. A este respecto, varios son los autores que recogen el hecho de que las probabilidades de encontrar una respuesta cuya estructura superficial sea similar a la de la pregunta aumentan de forma proporcional al aumento de tamaño de la colección de documentos de la base de conocimiento (Moldovan, Clark, y Bowden, 2007). No obstante, aún contando con una gran base de conocimiento, pueden darse casos en los que las respuestas a las preguntas planteadas por el usuario no contengan los mismos términos de la pregunta. Las diferencias léxicas y estructurales entre pregunta y respuesta solo pueden ser resueltas mediante un procesamiento lingüístico que permita acceder al significado subyacente de la pregunta.

Ya hemos visto que, a nivel léxico, una de las estrategias más comunes en los sistemas de BR es la expansión de los términos de la consulta mediante bases de conocimiento lingüístico como WordNet. En Moldovan, Pasca, et al. (2003) se lleva a cabo un experimento que trata de medir el impacto del uso de este tipo de

estrategias en un sistema de BR. En concreto, se mide el impacto de dos herramientas: WordNet para la expansión de la consulta y un sistema de NER para etiquetado de entidades. Para realizar el experimento, se procede a la desactivación de WordNet y del sistema NER por separado, no conjuntamente. La desactivación de ambas herramientas genera un descenso de la precisión general del sistema, especialmente la desactivación del sistema NER. El peso del sistema de NER se debe a que, al perder la información semántica sobre los tipos de entidades candidatas a ser las respuestas presentes en los textos, el sistema de BR se basa solamente en *keyword matching* entre pregunta-respuesta. Es interesante señalar además que la desactivación de WordNet afecta significativamente a las preguntas con *what*: la precisión general del sistema con WordNet desactivado es de 59, mientras que en las preguntas con *what* es de 37.

Hemos visto también que otro tipo de estrategias más complejas para gestionar la paráfrasis son las que operan con representaciones con un cierto nivel de abstracción semántica tanto de la pregunta como de la respuesta (Salloum, 2009; Harabagiu, Bunescu, y Maiorano, 2001; Harabagiu et al., 2005; Harabagiu, 2006; Fliedner, 2007; Moldovan, y Rus, 2001; Moldovan, Harabagiu, Girju, Morarescu, Lacatusu, Novischi, Badulescu, y Bolohan, 2002). Este tipo de representaciones intentan ir más allá de las diferencias estructurales superficiales y determinar el contenido semántico de pregunta y respuesta.

1.2.3.1.4 Pérdida de precisión

Estos son solo algunos ejemplos de la importancia del manejo preciso de la información que aportan sintaxis y semántica en los sistemas de BR. La principal consecuencia de la pérdida de esta información son los **problemas de precisión**. Para Fliedner (2007):

> We identify low precision as an important potential problem: A QA system using little or no linguistic information will often return wrong answers. We suggest that by systematically using structured linguistic information as a basis for answer searching, this problem can be tackled.

Fliedner (2007) recoge un dato ilustrativo de los problemas de precisión en los sistemas de BR: en TREC (2004) se recogen las valoraciones de todas las respuestas a preguntas factuales devueltas por los sistemas participantes (que, en su mayoría, utilizaban técnicas superficiales de procesamiento del lenguaje). De todas estas respuestas, el 77% fueron juzgadas como erróneas.

1.2.3.1.5 BR Avanzado

Todos los problemas derivados de la falta de manejo de información lingüística enumerados hasta ahora afectan a cualquier sistema de BR, opere con el tipo de preguntas que opere. No obstante, es evidente que a medida que aumenta la complejidad del tipo de preguntas planteadas al sistema, aumenta la necesidad de un

mayor manejo de conocimiento, ya no solo lingüístico, sino también conocimiento general («conocimiento del mundo»).

Hemos visto que las típicas «preguntas TREC» son preguntas de tipo factual que, en la mayoría de los casos, apuntan a una entidad. El mayor desafío de los sistemas que operan con estas preguntas es identificar por qué tipo de entidad se está preguntando y luego localizarla en los textos. Para realizar el primer paso, puede ser suficiente (a pesar de que se arrastrarían varios de los problemas mencionados ya) con estrategias próximas a la RI que funcionan con palabras clave sumadas a reglas simples que utilizan patrones para asociar cada frase interrogativa de las preguntas con un tipo de entidad en las respuestas (Litkowski, 1999). Para identificar la respuesta, se utilizarían sistemas de EI que reconocen y clasifican entidades. Hemos visto, no obstante, que incluso en el contexto de las preguntas factuales se dan muchos problemas de precisión y también casos en los que no es posible encontrar una respuesta correcta si la base de conocimiento no es lo suficientemente grande.

Todos estos problemas se multiplican en el contexto de la denominada BR Avanzada (cf. sección 1.2.1). En palabras de Saint-Dizier, y Moens (2011):

> [...] users do not only ask factual questions, which are often easy to answer, because the words of the question and of textual document sources overlap. Many interesting questions demand a more advanced understanding of the meaning of the question and of the information sources, as well as techniques of answer fusion and generation.

1.2.3.2 Experimentos sobre el uso de conocimiento lingüístico en BR

A los argumentos anteriores se pueden sumar los resultados concretos de diversos experimentos que tratan de determinar las ventajas del uso de información y recursos lingüísticos en un sistema de BR. Los trabajos que presentaremos a continuación son los siguientes: Hovy, Hermjakob, y Lin (2001); Moldovan, Pasca, et al. (2003); Quarteroni, Moschitti, Manandhar, y Basili (2007); Surdeanu, et al. (2008); Tiedemann (2005); Moreda, Llorens, Saquete, y Palomar (2011) y Carvalho et al. (2010).

Hovy, Hermjakob, y Lin (2001)

En Hovy, Hermjakob, y Lin (2001) se realiza un experimento donde se evalúa la eficacia de tres planteamientos diferentes en la fase de *matching* entre pregunta y respuesta, utilizando para ello un corpus de preguntas TREC. Los planteamientos comparados son los siguientes:

1) Uso de patrones de árboles sintácticos de pregunta y respuesta.
2) Uso de *Qtargets* (cf. *supra*) y palabras interrogativas en el árbol sintáctico de la pregunta.
3) Sistema de ventana al nivel de la palabra (*word window*).

Los planteamientos (1) y (2) utilizan información lingüística de distinto tipo: (1) información sintáctica, (2) información sintáctica y semántica. El planteamiento (3) es típico del área de RI, y funciona al nivel de la palabra. En palabras de Hovy, Hermjakob, y Lin (2001):

> Many projects employ a window-based word scoring method that rewards desirable words in the window. They move the window across the candidate answers texts/segments and return the window at the position giving the highest total score. A word is desirable if it is a content word and it is either contained in the question, or is a variant of a word contained in the question, or if it matches the words of the expected answer. Many variations of this method are possible—of the scores, of the treatment of multi-word phrases and gaps between desirable words, of the range of variations allowed, and of the computation of the expected answer words.

En el experimento, se calculan las mejoras aportadas por cada uno de los métodos a un punto de referencia (*baseline)*. Los resultados son los siguientes: en primer lugar, los patrones sintácticos de pregunta y respuesta aportan la menor contribución al *baseline*, un 5,5%, mientras que el uso de *Qtargets* y palabras interrogativas, la mayor contribución, un 26,2%. El uso de las ventanas al nivel de palabra se queda en medio de las otras dos aproximaciones al hacer una contribución a la eficacia general del *baseline* de un 10,4%.

Vemos por tanto que el planteamiento que ofrece mejores resultados es aquel que combina información sintáctica y semántica bastante elaborada (2). Por otra parte es interesante señalar el hecho de que, en este caso, el planteamiento que maneja información sintáctica (1) ofrece peores resultados que el método de RI (3).

Hovy, Hermjakob, y Lin (2001) señala los siguientes aspectos como las principales limitaciones del método de ventana al nivel de la palabra para emparejar pregunta y respuesta:

- No puede apuntar a los límites de la respuesta con precisión (por ejemplo, un nombre exacto o una frase nominal).
- Se asienta solamente sobre información al nivel de la palabra y, por lo tanto, no puede reconocer información del tipo «clase de respuesta esperada».
- No puede localizar y componer partes de la respuesta que están distribuidas en áreas mayores que la ventana.

Moldovan, Pasca, et al. (2003)

En un detallado e interesante estudio, Moldovan, Pasca, et al. (2003) evalúa el impacto de diversos factores sobre el funcionamiento general de un sistema de BR: desde el uso de herramientas y recursos lingüísticos al uso de distintos parámetros de recuperación de información. La principal conclusión, en sus propias palabras es:

> The main conclusion is that the overall performance depends on the depth of natural language processing resources and the tools used for answer finding.

De entre los múltiples análisis y experimentos que se llevan a cabo en este trabajo, dos ofrecen conclusiones especialmente relevantes.

En el primero de los experimentos se evalúa el impacto de recursos lingüísticos en la selección y extracción de respuestas. Se evalúan cuatro modelos:

1) **Sistema que no utiliza ninguna técnica de PLN**: en este modelo se seleccionan las primeras frases de los fragmentos de texto recuperados a

   partir de una búsqueda de tipo *booleano* basada en palabras clave extraídas de la pregunta.

2) **Sistema que utiliza emparejamiento léxico:** *matching* entre las palabras de la consulta inicial y las palabras en el texto.

3) **Sistema que usa conocimiento semántico**: la información semántica que se maneja es el tipo de respuesta esperada (EAT). El sistema extrae las respuestas alrededor de las entidades que encajan con la respuesta esperada, por ejemplo, nombres de persona si la EAT es *person*.

4) **Sistema que utiliza todo lo anterior y varios *feedbacks*[36].**

El peor resultado lo aporta el planteamiento que no utiliza ningún tipo de información lingüística (1), con una eficacia general del 0.028 de MRR[37]. Le sigue en eficacia el modelo de emparejamiento léxico (2), con un 0.150. La eficacia aumenta considerablemente con el uso de las EAT (3): un 0.468. Finalmente, la mayor contribución viene dada por el uso de *feedbacks* sumados a todo lo anterior (4): 0.572.

El segundo experimento ya se ha presentado (*vid*. *supra*): en él se pretende medir el impacto del uso de dos recursos lingüísticos, WordNet y un sistema de NER. Como ya se ha señalado Moldovan demuestra que el uso de WordNet y el sistema de NER (especialmente este último) mejoran considerablemente la eficacia del sistema de BR.

Manejo de información sintáctica

A continuación se presentan tres experimentos: Quarteroni et al. (2007), Surdeanu et al. (2008), y Tiedemann (2005), que investigan las aportaciones de información estructural en distintos procesos de BR. Con este fin, en los experimentos se comparan, fundamentalmente, aproximaciones que no manejan información lingüística (tipo BOW) con aproximaciones que manejan información lingüística (sintáctica, sintáctico semántica, etc.). Los tres estudios concluyen que, aunque el modelo BOW supone la mayor contribución a los resultados del sistema de BR, el uso de información lingüística, fundamentalmente estructural (información sintáctica) supone una mejora estadísticamente significativa.

36 El sistema de BR de Moldovan posee varios *feedbacks* que permiten reformular la información disponible en distintas fases del proceso.

37 *Mean Reciprocal Rank*: se trata de una medida de evaluación propia de la RI. Cf.: http://en.wikipedia.org/wiki/Mean_reciprocal_rank

Quarteroni et al. (2007)

En Quarteroni et al. (2007) se estudia el impacto de distintas aproximaciones en tres tareas de BR: clasificación de la pregunta, clasificación de la respuesta y *reranking* de la respuesta. Las aproximaciones comparadas son:

1) BOW.
2) Árboles sintácticos.
3) *Predicate-argument structures* (PAS) (*vid. supra*)
4) PAS+BOW.

Los peores resultados los ofrece el planteamiento que funciona con PAS (3); los mejores, PAS+BOW (4), que superan a la representación en árbol sintáctico (2) (*F-score* de 70.7% *vs.* 59.6%). No obstante, la mejora introducida por el uso de PAS en el modelo PAS+BOW no es muy grande, ya que el modelo BOW (1) ofrece un resultado de *F-score* 69.3% (*vs.* 70.7% PAS+BOW).

Surdeanu et al. (2008)

Surdeanu et al. (2008) investiga el problema de la clasificación de respuestas a preguntas tipo *How-to* (preguntas procedimentales). Se utiliza un corpus de 40.000 preguntas y 140.000 respuestas extraído de *Yahoo! Answers*[38] y se explora la utilidad de un amplio conjunto de características (con más o menos información lingüística) de preguntas y respuestas en la tarea de clasificación. Se concluye que los rasgos lingüísticos

> yield a small, yet statistically significant performance increase on top of the traditional BOW and n- gram representation.
> (Surdeanu et al., 2008, p. 726).

Tiedemann (2005)

Tiedemann (2005) explora el uso de información sintáctica de tipo dependencial en la recuperación de párrafos para BR factual en holandés. Para ello, indexa su corpus a diferentes niveles textuales (BOW, *part-of-speech*, relaciones dependenciales) y usa los mismos niveles en el análisis de la pregunta y la creación de la consulta. Se optimizan los parámetros de la consulta para la extracción de párrafos aplicando un algoritmo genético que asigna pesos a los términos de la consulta. El resultado muestra que los mayores pesos son asignados a las palabras clave extraídas del nivel BOW y a las palabras clave relacionadas con el tipo de respuesta esperada (como *person*). El *baseline*, usando solo el nivel BOW tiene un MRR (cf. nota 32) de 0.342. Usando las características optimizadas de RI con niveles adicionales (lingüísticos), el MRR sube hasta 0.406.

Moreda et al. (2011)

En Moreda et al. (2011) se analiza el uso de información semántica en la fase de extracción de respuestas de un sistema de BR. Para ello, se evalúan dos aspectos:

38 http://answers.yahoo.com/

1) Mejoras introducidas por la utilización de información semántica a distintos niveles: distintos tipos de etiquetado de roles semánticos (*Semantic Role Labelling*, SRL).

2) Comparación entre aproximaciones con SRL y aproximaciones tipo NER sin SRL en dos corpus de preguntas: uno con preguntas clásicas tipo entidades nombradas y otro con preguntas tipo entidades nombradas y nombres comunes.

Los tipos de SRL que se comparan, son los siguientes:

- **SRL *baseline***: módulo de extracción de respuestas basado en etiquetado semántico que usa reglas semánticas. El sistema considera como candidato para cada tipo de pregunta definida («persona», «organización», «temporal» y «localización») una serie de argumentos etiquetados con roles semánticos extraídos de PropBank. Por ejemplo, para preguntas tipo «persona», los argumentos de PropBank etiquetados como A0 y A1 son considerados respuestas candidatas.
- **SRL usando patrones semánticos**: sistema que utiliza patrones semánticos construidos automáticamente a partir de un corpus de preguntas-respuestas e información de PropBank. Los patrones semánticos están asociados a una serie de tipos de preguntas, y tienen una estructura como la siguiente:

  be, bear, |<QARG> <AM – LOC> <ARG1 >

  donde en primer lugar tenemos el verbo seguido por una serie de tipos semánticos de argumentos.

- **SRL con patrones + WordNet**: este sistema es una extensión del anterior. Consiste en añadir al patrón una lista de las clases semánticas del tipo del de la respuesta esperada, aumentando el filtrado de respuestas potenciales. Esta lista de clases semánticas se extrae de la jerarquía de hiperónimos de WordNet.
- **SRL con patrones + WordNet exacto**: sistema que funciona como el anterior pero que exige el *match* exacto del tipo de la clase semántica de WordNet.

En relación al aspecto (1), el SRL que mejor funciona en términos globales es el SRL con patrones semánticos (*F-score* 72.35% de media en los dos experimentos). En comparación, los sistemas de SRL que utilizan más información semántica (clase semántica de la respuesta esperada extraída de WordNet), muestran mayor precisión pero menor *recall*, lo que genera un menor *F-score* (por ejemplo, *F-score* 69.51% de media en los dos experimentos para SRL con patrones + WordNet exacto).

En cuanto al aspecto (2), los resultados por corpus son los siguientes:

- **Corpus de preguntas tipo NE:** el mejor resultado es para la aproximación clásica de NE sin información semántica: *F-score* 85.70% frente al 72.94% de SRL con patrones (recordemos, con el mejor resultado para los sistemas con información semántica).

- **Corpus preguntas NE + nombres comunes**: en este caso hay una inversión de resultados, y la diferencia entre el mejor sistema con SRL y el sistema con NER es considerable. La aproximación tipo NE obtiene un *F-score* general de 48,94%, mientras que el mejor de los sistemas SRL, SRL con patrones, obtiene un *F-score* de 73,83%. La diferencia aumenta si solo tenemos en cuenta las preguntas con nombres comunes y sin entidades: un 12,19% para el sistema de NE frente a un 74,73% para el sistema de SRL con patrones.

Carvalho et al. (2010)

Finalmente, en Carvalho et al. (2010) se analizan las limitaciones de un sistema de BR para el portugués, *IdSay*, comparándolo con otros sistemas. Se identifican dos aspectos como fuentes del éxito de esos otros sistemas respecto a *IdSay:*

- el uso de **fuentes de información semántica** en la fase de recuperación y validación de las respuestas;
- el uso de **herramientas de análisis lingüístico** que permiten mejoras en el procesamiento de la pregunta y en la extracción de la respuesta.

Conclusiones sobre los experimentos presentados

Una vez revisados los datos aportados por todos estos experimentos podemos concluir afirmando que todos coinciden en un hecho: el manejo de información lingüística (a distintos niveles: léxico, sintáctico y semántico), mejora la eficacia de los sistemas de BR analizados.

1.2.3.3 Desventajas del uso de conocimiento lingüístico en los sistemas de BR

Ya se han señalado dos grandes desventajas de la aplicación de conocimiento lingüístico de cierto nivel en los sistemas de BR:

1) **Pérdida de *recall*** (Katz y Lin, 2003): el uso de restricciones lingüísticas adicionales en la fase de extracción de la respuesta puede perjudicar el *recall* del sistema, al disminuir el número de respuestas correctas encontradas.

Esto puede ocurrir, por ejemplo, en los casos en los que se utilizan representaciones sintácticas de pregunta y respuesta: si las estructuras de la pregunta y la respuesta potencial difieren, la respuesta será ignorada, aunque esta pudiera ser, sin embargo, correcta.

2) **Tiempo de respuesta**: el uso de herramientas lingüísticas que requieren un procesado profundo de los textos (*parsing*, interpretación semántica, etc.) puede ampliar demasiado el tiempo necesario para obtener una respuesta, teniendo en cuenta las grandes cantidades de texto que se manejan como bases de conocimiento.

Como ya hemos visto, una de las soluciones que se han propuesto a este problema es el procesado de la base de conocimiento *off-line*, solución posible siempre y cuando se maneje una colección de documentos cerrada.

1.2.4 Conclusiones generales

En este apartado se ha analizado el manejo de información lingüística en los sistemas de BR.

En primer lugar se han presentado diversos motivos teóricos que apoyan el uso de conocimiento lingüístico, especialmente, en sistemas de BR que van más allá del modelo factual de TREC.

A continuación se han presentado los datos aportados por diferentes estudios que tratan de evaluar el impacto del uso de información lingüística en sistemas de BR. La principal conclusión que se extrae de estos datos es que el manejo de conocimiento lingüístico, especialmente en el procesamiento de la pregunta y la selección/extracción de la respuesta, mejora la eficacia general de los sistemas de BR.

Finalmente, se han presentado los dos argumentos principales en contra del uso de información lingüística en los sistemas de BR: la pérdida de *recall* y el aumento de tiempo de respuesta del sistema.

1.3 El procesamiento de la pregunta: SpQA

1.3.1 Conocimiento lingüístico en el procesamiento de la pregunta

En las secciones anteriores hemos presentado una visión general sobre los sistemas de BR y su funcionamiento, así como sobre el uso de conocimiento lingüístico en dichos sistemas.

Como hemos visto, diversos factores sugieren la conveniencia del uso de conocimiento lingüístico tanto en la fase de procesamiento de la pregunta como en la fase de extracción de la respuesta.

En Fliedner (2007) se describe un marco teórico y un sistema de BR motivado lingüísticamente. En la línea de nuestro trabajo, se exploran las contribuciones que la perspectiva lingüística puede aportar a la BR. Una de las conclusiones de Fliedner (2007) es que lo ideal en un sistema de BR sería poder utilizar un análisis semántico profundo junto con procesos de razonamiento que permitieran inferencia. Dicho análisis podría representar el significado subyacente a la pregunta y permitiría encontrar respuestas adecuadas desde el punto de vista semántico, superando posibles diferencias estructurales entre pregunta y respuesta motivadas por cuestiones léxicas y/o sintácticas. No obstante, ya hemos apuntado

que un procesamiento a ese nivel está fuera de las posibilidades de las herramientas actuales. Es por esta misma razón por la que en el sistema de Fliedner (2007) se escoge una vía intermedia entre los planteamientos con un procesamiento lingüístico superficial y los planteamientos con un procesamiento semántico profundo. En palabras del propio investigador:

> In the conclusion, we will propose that an intermediate level of abstraction, namely between shallow bag of words and deep full meaning representations, should be used for practical QA systems.
> (Fliedner, 2007, p. 97)

Este «nivel intermedio de abstracción» se construye a partir de una representación sintáctica, enriquecida semánticamente a través de información de carácter léxico[39]:

> We will show below, however, that basing QA on structured semantic representations (let alone pragmatic modelling) of texts and questions is currently beyond the scope of natural language processing systems. We will therefore suggest an approximation based mainly on syntactic and lexical semantic information.
> (Fliedner, 2007, p. 85)

En la línea de Fliedner (2007) consideramos que lo ideal en BR sería un análisis semántico profundo que manejara conocimiento del mundo y permitiera inferencia y razonamiento lógico. Sabemos que este análisis es actualmente inviable para todo el proceso de BR (especialmente para los pasos que implican el procesamiento de grandes cantidades de texto), pero sí puede plantearse para la fase de análisis de la pregunta. Las preguntas son estructuras cortas cuyo procesado, incluso a un nivel complejo, no debería provocar problemas de eficiencia (en términos de tiempo) para el sistema de BR. Por otra parte, aunque las preguntas son estructuras lingüísticas cortas, debido a la especial relación que se da en el par pregunta-respuesta, contienen una gran cantidad de información que es clave para que el sistema de BR pueda interpretar los intereses del usuario y encontrar respuestas correctas. Exprimir al máximo las posibilidades informativas de las preguntas es, por lo tanto, esencial.

1.3.2 Análisis sintáctico en el procesamiento de la pregunta

Entre estas «posibilidades informativas» de la pregunta, las relaciones estructurales (sintácticas) entre los elementos son, como ya hemos visto, muy relevantes (especialmente en el caso de las preguntas parciales en las que tenemos un interrogativo complejo). Experimentos como los de Quarteroni et al. (2007), Surdeanu (2008) o Tiedemann (2005) demuestran que el uso de información sintáctica supone una mejora significativa en los resultados de los sistemas de BR.

39 Como veremos en el capítulo 4, nuestra propuesta para el análisis de las preguntas es similar a esta.

De hecho, prácticamente todos los sistemas de BR que realizan un procesamiento lingüístico de cierto nivel utilizan información sintáctica. Además, muchos sistemas con SRL funcionan sobre un análisis sintáctico previo.

Todos estos argumentos nos llevan a considerar que el análisis sintáctico debe jugar un papel en ese procesamiento lingüístico profundo de las preguntas en los sistemas de BR. La correcta identificación de constituyentes, así como de sus funciones y relaciones de dependencia, es un paso básico para una interpretación acertada de la pregunta, más si se tiene en cuenta que sobre este análisis sintáctico se pueden construir representaciones más complejas de tipo semántico.

Pese a su importancia, sobre todo en la fase de procesamiento de la pregunta (Hermjakob, 2001; Moldovan et al., 2002), el análisis sintáctico no ha recibido demasiada atención en el mundo de la BR. La mayoría de los sistemas que manejan información sintáctica utilizan *parsers* generales y de tipo estadístico, a pesar de que se ha demostrado que la eficacia de los *parsers* de tipo general disminuye al utilizarlos en dominios específicos (Gildea, 2001; McClosky, et al., 2006; Foster, 2010). Concretamente, varios estudios han demostrado que esta pérdida de eficacia se da en la utilización de *parsers* estadísticos para el análisis de preguntas (Hermjakob, 2001; Petrov et al., 2010). Esto se debe a que, en la mayoría de los casos, los *parsers* estadísticos no han sido entrenados con corpus de preguntas sino con corpus generales. En palabras de Fliedner (2007):

> Note that the issue of question parsing must not be underestimated: many available parsers for natural language lack support for interrogative structures. Statistical parsers, for example, often do not handle them due to their scarcity in general corpus.

Los *parsers* no estadísticos de tipo general, también presentan ciertos problemas a la hora de analizar preguntas, como muestra Gayo (2011a) y Gayo (2011b) para el español. Al respecto, Katz y Lin (2003) apuntan que:

> However, deriving relations using off-the-shelf parsers, while convenient for current experimental purposes, might not be the ideal situation in the longer term. A custom-built lightweight parser specifically designed to extract relations might be faster and more accurate.
> (Katz y Lin, 2003)

La propuesta de este trabajo va en la línea de las palabras de (Katz y Lin, 2003): la construcción de un *parser*, SpQA, específico para el análisis (sintáctico y, en menor medida, semántico) de preguntas, robusto y con cobertura, capaz de extraer la máxima cantidad de información relevante de una pregunta para la obtención de una respuesta correcta. El principal objetivo del *parser* es el análisis sintáctico, entendido como la identificación y etiquetado de los constituyentes de la pregunta, así como de sus relaciones de dependencia. No obstante, al ser un *parser* específico para preguntas y diseñado para un entorno de BR, SpQA también consta de un componente semántico (construido, como en Fliedner, 2007, a partir de información léxico-semántica y sintáctica). El nivel semántico está relacionado con la identificación y clasificación de la variable o incógnita presente en la pregunta (dato especialmente relevante en el análisis de los interrogativos en las preguntas parciales

como veremos en detalle en el capítulo 4), y con el reconocimiento de entidades clave en BR, tales como entidades nombradas, cantidades y fechas.

El marco teórico en el que se inserta SpQA es el de la BR lingüísticamente motivada, en la línea del sistema de Fliedner (2007). Por esta razón, la construcción del *parser* se asienta sobre tres pilares:

1) Las necesidades que presenta la fase de análisis de la pregunta en los sistemas de BR, que hemos tratado en este primer capítulo, y que se resumen en una comprensión lo más completa posible del significado de la pregunta. Podríamos además señalar como claves para esa comprensión los siguientes puntos:

   - Reconocimiento de entidades presentes en la pregunta (verbos, entidades nombradas, fechas, cantidades, etc.) y delimitación de las entidades clave (entidades en el constituyente interrogativo en las parciales, por ejemplo).

   - Relaciones (idealmente, semánticas, pero también sintácticas) que se establecen entre esas entidades, con el fin de establecer el valor semántico general de la pregunta.

2) El estudio teórico de la estructura de las preguntas en español, así como de la relación pregunta-respuesta (que trataremos en el capítulo 2). A partir de este estudio lingüístico se pretenden extraer características que ayuden a una mayor comprensión del funcionamiento de las preguntas (y de la relación pregunta-respuesta) y de su significado.

3) El análisis de preguntas, tanto de usuarios reales como de otro tipo, con el fin de obtener datos reales sobre diversos aspectos de su estructura (trabajo de corpus que presentaremos en el capítulo 3). Se espera que estos datos reales puedan enriquecer los extraídos del análisis lingüístico teórico (punto anterior).

Estos tres pilares sustentan y dan lugar a SpQA, de cuya descripción daremos cuenta en el capítulo 4 y cuya eficacia evaluaremos en el capítulo 5.

1.4 Conclusiones generales del capítulo

En este primer capítulo se han presentado los sistemas de BR, se ha analizado el uso de conocimiento lingüístico en dichos sistemas y se ha presentado el modelo de BR lingüísticamente motivado en el que se integra SpQA.

En la primera sección se han caracterizado los sistemas de BR y se ha trazado su evolución. Hemos visto que la BR es una tarea compleja que implica diversos procesos interrelacionados, estructurados en tres módulos: análisis de la pregunta búsqueda y selección de la respuesta, extracción y presentación de la respuesta. Este trabajo se centra en el módulo de análisis de la pregunta, que es donde se integraría SpQA. En el análisis de la pregunta en BR, son comunes tareas como: la extracción

de palabras clave; la representación sintáctica, semántica o sintáctico semántica de la pregunta, el establecimiento del tipo de respuesta esperada (EAT) o la clasificación de la pregunta. Para llevar estas tareas a cabo se utilizan técnicas propias de diferentes áreas: RI, EI, PLN o IA.

En la segunda sección se ha profundizado en el uso de conocimiento lingüístico en los sistemas de BR. La principal conclusión es que el uso de conocimiento lingüístico mejora en general la eficacia de estos sistemas. Los principales problemas que se han señalado para el uso de conocimiento lingüístico de cierto nivel en BR son dos: aumento del tiempo de respuesta del sistema y pérdida de *recall*.

En la última sección hemos defendido un modelo de BR lingüísticamente motivado. En este marco se sitúa SpQA. Nuestro modelo ideal sería aquel que permitiera un procesado semántico profundo de pregunta y respuesta. Creemos que actualmente este planteamiento es viable, al menos, para el análisis de la pregunta. Para llegar a él, consideramos que un paso previo básico es el análisis sintáctico semántico de las preguntas. Este análisis sintáctico semántico de las preguntas en un entorno de BR es el objetivo de SpQA. Los pilares sobre los que SpQA se asienta son tres: las necesidades del módulo de procesamiento de la pregunta de un sistema de BR, el conocimiento lingüístico sobre preguntas y la relación pregunta-respuesta, y el conocimiento de corpus sobre el funcionamiento de las preguntas. El primer pilar lo hemos tratado en este primer capítulo, donde hemos defendido la importancia de un análisis lingüístico de la pregunta lo más rico posible (idealmente, un análisis semántico que represente su significado). Este análisis debe implicar, al menos: el reconocimiento de las entidades presentes en la pregunta (entidades nombradas, fechas, cantidades, etc.), las relaciones (semánticas y sintácticas) que se establecen entre esas entidades, la delimitación de las palabras clave de la pregunta, la clasificación de esta, etc. En el siguiente capítulo nos ocuparemos del estudio lingüístico de las preguntas y la relación pregunta-respuesta.

Capítulo 2

Las preguntas: descripción lingüística

En el capítulo anterior se ha establecido como uno de los pilares sobre los que se construye SpQA el estudio de la estructura y funcionamiento de las preguntas en español, así como de la relación pregunta-respuesta.

La información sobre el funcionamiento de las preguntas en español aportada por los estudios teóricos tiene un doble objetivo: por una parte, nos permite situar desde el punto de vista teórico nuestro objeto de análisis (las preguntas). Por otra, más importante para los fines prácticos de este trabajo, nos permite definir características de las preguntas que pueden ser pertinentes para su formalización y posterior análisis en un entorno de BR.

Por estas razones, en el presente capítulo nos ocuparemos de la descripción teórica de las preguntas desde la perspectiva lingüística, con el fin de definir **qué es una pregunta y cuáles son sus características más relevantes.** Para ello nos valdremos de una serie estudios lingüísticos[40] que tratan el funcionamiento de las preguntas[41], tanto a nivel gramatical como semántico.

Nuestra descripción pretende ser amplia y exhaustiva en lo que se refiere al funcionamiento de las preguntas y la relación pregunta-respuesta, con el objetivo de formalizar en la gramática de SpQA toda la información relevante para dar cuenta de ese funcionamiento. Con este objetivo, se describirán aquellos aspectos lingüísticos que juegan un papel tanto en la estructura (orden de constituyentes, características de las partículas interrogativas, etc.) como en el significado (foco, relación pregunta-respuesta, etc.) de las preguntas en español.

Como veremos a lo largo de la descripción, no todos estos fenómenos son interesantes para el perfil típico de BR factual (cf. capítulo 1), o no son formalizables en una gramática como la de SpQA[42]. Por esta razón, al final del capítulo, retomaremos y reexaminaremos los aspectos más importantes para nuestra formalización de las interrogativas en la gramática de SpQA (cf. capítulo 4).

---

40 Los estudios lingüísticos, tanto gramaticales como semánticos, centrados en las preguntas en español, no son muy abundantes.

41 Siempre que sea posible, se utilizarán estudios centrados en la descripción de las preguntas en español. Sobre todo en las cuestiones semánticas, los trabajos utilizados serán de tipo general.

42 Pero sí pueden ser útiles para otros tipos de sistemas de BR (con diálogo, por ejemplo) o para otras formalizaciones. La descripción, por tanto, sirve a la formalización en SpQA, pero no está totalmente limitada por ella.

La información teórica pertinente se estructurará en tres bloques:
1) ¿Qué es una pregunta?
2) Gramática de las preguntas.
3) Semántica y pragmática de las preguntas.

2.1 ¿Qué es una pregunta?

En el área de BR el término «pregunta» se suele utilizar de un modo laxo, de modo que engloba tanto estructuras que se consideran prototípicamente preguntas (1) como otro tipo de estructuras (2):

(1) *¿Cómo murió Bob Marley?*
(2) *Enumera todos los presidentes del gobierno español desde 1980.*

Por esta razón, consideramos importante definir en primer lugar a qué nos referimos en este trabajo cuando hablamos de «preguntas».

Observando los ejemplos anteriores, se puede deducir que lo que equipara (1) y (2) es **la finalidad que estas estructuras persiguen**: **una demanda de información** por parte del hablante, que es, sin embargo, expresada a través de **formulaciones diferentes**. En Lingüística, la finalidad de un acto de habla es un factor de tipo pragmático que se define a partir de la intención del hablante (una categoría ilocutiva).

En la tradición de los sistemas de BR este factor de tipo pragmático ha sido el preponderante a la hora de definir qué se consideraba «pregunta»: estructuras gramaticales diferentes que, en un lenguaje más o menos formal, representan una demanda de información. Hirschman, y Gaizauskas (2001) lo exponen claramente (el destacado es nuestro):

> Next we can distinguish **different kinds of questions**: yes/no questions, "wh" questions (who was the first president, how much does a killer whale weigh), indirect requests (I would like you to list ...), and commands (Name all the presidents...). **All of these should be treated as questions.**
> (Hirschman, y Gaizauskas, 2001, p. 278)

En este trabajo no manejaremos un concepto de pregunta tan poco restrictivo. Para nosotros una pregunta también será una demanda de información, pero su estructura gramatical se corresponderá con la de una oración interrogativa directa[43] y no con cualquier otro tipo de estructura. Para todo aquello que sea una demanda de información pero que no sea una oración interrogativa directa utilizaremos la formulación genérica «petición de información» (Escandell, 1999, p. 3974).

Otra confusión habitual es la de equiparar «oración interrogativa» y «demanda de información»[44]. Esto ocurre porque la forma prototípica que se asocia a

43 No consideramos las oraciones interrogativas indirectas como preguntas porque definimos estas últimas como enunciados autónomos. Volveremos sobre esta cuestión más adelante.

44 En los estudios gramaticales tradicionales se da habitualmente una confusión entre pregunta y oración interrogativa, equiparándose ambas nociones. En ámbitos como el anglosajón, sin embargo, es común distinguir entre una categoría sintáctica, denominada *interrogative* y una

las demandas de información es la de las oraciones interrogativas (al igual que la forma gramatical que se asocia a las órdenes son las oraciones exhortativas). Sin embargo, sabemos que no todas las oraciones interrogativas representan una demanda de información.

Imaginemos a dos personas sentadas a una mesa, A y B. La mesa está servida para el desayuno y, entre otras cosas, hay té. B está más próximo al té, de manera que A le dice:

(3a) —*¿Podrías pasarme el té?*

Si B respondiese a (5) lo siguiente:

(3b) —*Sí, podría.*

y no hiciese nada, A se extrañaría y, probablemente, se enfadaría. Lo esperable sería que B contestase algo como

(3c) —*Claro.*

seguido de la acción de pasarle el té a A. Esto ocurre porque cualquier hablante de español sabe que con el enunciado de (3a) A no está demandando la información de si «B está capacitado para pasarle o no pasarle el té», sino que lo que está haciendo es realizar una petición de forma cortés.

Las oraciones interrogativas son un tipo de estructura gramatical complejo que presenta variaciones tanto en su forma como en sus contextos de uso (Escandell, 1999). Es por ello por lo que debe quedar claro que no todas las oraciones interrogativas son preguntas: solo un determinado tipo de oraciones interrogativas, cuyas características detallaremos a continuación, se utilizan en español para demandar información y, por lo tanto, como preguntas.

Por lo tanto, en este trabajo el término «pregunta» se utiliza para un enunciado que:

- representa una demanda de información por parte del hablante (nivel pragmático semántico);
- tiene la forma de una oración interrogativa directa[45] (nivel sintáctico semántico), pero no la de cualquier oración interrogativa directa, sino solo la de aquellas que se utilizan para demandar información.

Nuestras preguntas constituyen, por tanto, la intersección entre la categoría pragmática «demanda de información» y la categoría gramatical «oración interrogativa directa».

categoría ilocutiva de demanda de información, denominada *inquiry* o *asking* (Escandell, 1999).

45 Como veremos en la siguiente sección, nuestro concepto de «oración interrogativa directa» se basa fundamentalmente en (Escandell, 1999) y coincide en términos generales con el de la Nueva Gramática de la Academia (Real Academia Española, 2009, 42.6).

En nuestro estudio no abordamos las peticiones de información. La razón es que consideramos que en un sistema de BR sin interacción las preguntas tal y como las hemos definido constituirían la fórmula prototípica para demandar información[46] y, por lo tanto, la fórmula a la que se debe dar prioridad en SpQA. Lograr una adecuada cobertura en su análisis es un primer paso necesario y básico en el módulo de procesamiento de la pregunta. Por esta razón, centraremos nuestro trabajo en las preguntas[47].

En los apartados siguientes nos ocuparemos de la descripción lingüística de las preguntas. En primer lugar, abordaremos las características de tipo gramatical; por tanto, nos ocuparemos de la descripción de las oraciones interrogativas directas que se utilizan para demandar información. A continuación, profundizaremos en aspectos semánticos y pragmáticos del funcionamiento de las preguntas tal como la relación pregunta-respuesta.

2.2 Aspectos gramaticales de las preguntas: las oraciones interrogativas directas[48]

En esta sección nos ocuparemos de la caracterización gramatical de las preguntas. Abordaremos, por tanto, la descripción de las oraciones interrogativas directas.

En primer lugar, definiremos las oraciones interrogativas y situaremos en esta categoría las oraciones interrogativas directas. A continuación, nos centraremos en las características gramaticales específicas de las oraciones interrogativas directas que se utilizan para demandar información (nuestras preguntas), tratando solo aquellas oraciones interrogativas (y sus características) que se utilizan con otras finalidades cuando sea útil a la descripción.

2.2.1 Las oraciones interrogativas como incógnita

El primer paso a la hora de definir gramaticalmente nuestro objeto de estudio es la definición de la categoría a la que pertenece, es decir: la definición de las oraciones interrogativas.

Siguiendo a Escandell (1999), consideramos que aquello que define las oraciones interrogativas, sea cual sea su finalidad pragmática o su forma gramatical, es el hecho de contener una incógnita, una variable cuyo contenido no está resuelto. Las oraciones interrogativas son, por lo tanto, estructuras abiertas, incompletas[49].

46 En el capítulo 3 veremos como las preguntas así entendidas constituyen la principal fórmula documentada en los tres corpus con los que se ha trabajado.

47 Esto no significa que no se contemplen las peticiones de información como posibles fórmulas para preguntar (de hecho, su posible integración es uno de los pasos futuros que se plantean para SpQA, cf. Conclusiones), significa que las preguntas tal y como las entendemos aquí se consideran la fórmula primaria para demandar información, por lo que, al menos en una primera fase de desarrollo, su correcto procesamiento es el objetivo de SpQA.

48 Esta sección se nutre, fundamentalmente, de la descripción de las oraciones interrogativas de (Escandell, 1999).

49 Profundizaremos en este aspecto cuando tratemos la semántica de las preguntas (cf. sección 3 de este capítulo).

Esto no significa, no obstante, que todas las oraciones interrogativas supongan un vacío informativo que deba ser «llenado» por un interlocutor. Ya hemos visto que las oraciones interrogativas se utilizan con otras finalidades además de la demanda de información. En palabras de Escandell (1999) (el resaltado es nuestro):

> Las razones por las que un emisor decide utilizar una fórmula abierta son muy variadas: manifestar desconocimiento real, expresar una duda, avanzar una hipótesis, insinuar sin afirmar explícitamente, presentar un contenido que no comparte, etc.: en ausencia de un contexto y una situación determinados, **emitir una oración interrogativa equivale simplemente a expresar una función proposicional abierta: el objetivo con que se haga y las circunstancias que lo rodeen constituyen ya aspectos pragmáticos del significado.**
> (Escandell, 1999, p. 3934)

La variable o incógnita presente en las oraciones interrogativas puede ser de tres tipos:

- Puede estar definida por una partícula interrogativa:
  (4a) Interrogativa directa: *¿**Qué** es el Mossad?*
  (4b) Interrogativa indirecta[50]: *Me gustaría saber **qué** es el Mossad.*

  SIGNIFICADO BÁSICO: El Mossad es + incógnita (= ¿qué?)

- Puede corresponder al carácter afirmativo o negativo de lo que se predica:
  (5a) Interrogativa directa: *¿Es verdad que se pueden ver de noche los satélites como estrellas?*
  (5b) Interrogativa indirecta: *Quisiera saber si es verdad que se pueden ver de noche los satélites como estrellas.*

  SIGNIFICADO BÁSICO: Se pueden ver de noche los satélites como estrellas + incógnita (= ¿sí?/¿no?)

- Puede corresponder a un elemento de una serie presente en la pregunta:
  (6a) Interrogativa directa: *¿Llega el martes o el miércoles?*
  (6b) Interrogativa indirecta: *Me pregunto si llega el martes o el miércoles.*

  SIGNIFICADO BÁSICO: LlegaB + incógnita (=¿el martes?/¿el miércoles?)

Los tres tipos de variable, junto a una serie de características gramaticales concretas, definen tres tipos de oraciones interrogativas: parciales (4), totales (5) y disyuntivas (6).

Como veremos en detalle, en el caso de las parciales las características de la incógnita vienen determinadas por la partícula interrogativa, que en unos casos puede restringir mucho el ámbito de la variable (*quién*, por ejemplo, siempre apunta una

50 Pese a que en este trabajo no nos ocupamos de las interrogativas indirectas, en este apartado se tratan también porque se describen las oraciones interrogativas en general.

entidad animada de tipo persona), y en otros no tanto (*cuál* puede apuntar tanto a entidades animadas como inanimadas). En las totales y las disyuntivas, el ámbito de la variable queda totalmente definido en la pregunta:

- En las totales corresponde a la afirmación (*sí*) o a la negación (*no*) de lo que se predica en la oración.
- En las disyuntivas corresponde a uno de los elementos de la estructura disyuntiva (*martes*/*miércoles* en (6)).

Por lo tanto, al menos en un principio, la oración interrogativa predetermina con su forma el tipo y la categoría del elemento que puede cerrar la proposición planteada por el que pregunta: en las parciales debe ser un elemento permitido por la partícula interrogativa; en las totales, *sí* o *no*; en las disyuntivas, uno de los elementos que aparecen en la disyunción. Como veremos cuando tratemos la semántica de las preguntas (cf. sección 3 de este capítulo), esta es la razón por la que el significado de una pregunta se ha identificado a veces con el conjunto de respuestas que formalmente predetermina (sus respuestas posibles) (Escandell, 1999). Este es el planteamiento que se defiende en este trabajo. Escandell (1999) recoge que este mismo planteamiento se encuentra, entre otros, en Stahl (1956), Belnap (1966; 1983), Belnap y Steel (1976), o Jacques (1981), y cita, para otras perspectivas, a Bäuerle (1979), Wunderlich (1981) o Diller (1984).

2.2.1.1 Totales y disyuntivas

Respecto a la clasificación de las interrogativas que acabamos de presentar, algunos autores han planteado que las interrogativas totales son en realidad un tipo de disyuntivas, ya que presentan dos alternativas entre las que se debe elegir: *sí* o *no*. De hecho, en la Nueva Gramática (NGRALE), la Academia agrupa totales y disyuntivas bajo una misma clase, «totales», distinguiendo entre «polares» (nuestras totales) y «alternativas» (nuestras disyuntivas) (Real Academia Española, 2009, 42.6a y ss.). Para NGRALE

> [las disyuntivas] no dejan de ser preguntas totales porque contienen la totalidad de la información que se presenta como pertinente.
> (Real Academia Española, 2009, 42.7a).

Por nuestra parte, siguiendo a Escandell (1999), consideramos que existen argumentos para definir totales y disyuntivas como estructuras distintas. Por ejemplo, Bolinger (1978) señala que totales y disyuntivas parecen obedecer a condiciones de uso diferentes, de manera que no son intercambiables en todos los contextos y situaciones. Además de las condiciones de uso discursivo, existen restricciones de tipo sintáctico, como las relacionadas con la aparición de términos de polaridad negativa (Escandell, 1999, p. 3933). Por otra parte, aunque la «totalidad de la información se presente como pertinente» en ambas estructuras, como dice la Academia, el valor de la variable es distinto en totales y disyuntivas.

Por estas razones, en nuestra descripción teórica trataremos totales y disyuntivas como entidades independientes.

2.2.2 Clasificación de las oraciones interrogativas: interrogativas directas *vs.* interrogativas indirectas

Las oraciones interrogativas pueden clasificarse de distintas maneras dependiendo de cuáles sean las características que se tengan en cuenta para realizar esa clasificación. De esta manera, por ejemplo, en Escandell (1999) podemos encontrar distintas clasificaciones de las interrogativas: Hudson (1975), Goody (1978), Escandell (1988; 1993), Selting (1992), Athanasiadou (1991; 1994), Freed (1994), Huddleston (1994) o Fava (1995).

Atendiendo a cuestiones gramaticales, las oraciones interrogativas se han dividido tradicionalmente en dos grupos: interrogativas directas (7) e interrogativas indirectas (destacado en negrita en (8)).

(7) *¿Cuánto mide un ángulo complementario?*
(8) *Quiero saber* ***si va a llover.***

Las interrogativas directas constituyen enunciados autónomos, mientras que las indirectas son una «variante de las oraciones subordinadas sustantivas» (Real Academia Española, 2009, 42.6a) y por lo tanto no son enunciados autónomos.

Los tres tipos de oraciones interrogativas definidos por la variable que presentábamos más arriba, parciales, totales y disyuntivas, se han considerado tradicionalmente tipos de interrogativas directas (Real Academia Española, 2009, 42.6a). Para la clasificación de las interrogativas indirectas se han considerado otros factores, como, por ejemplo, la función sintáctica de la indirecta en la oración en la que se inserta (Real Academia Española, 2009, 43.7a).

En este trabajo usaremos esta taxonomía:

- directas *vs.* indirectas para las oraciones interrogativas en general;
- totales, parciales y disyuntivas para las oraciones interrogativas directas.

Recordemos que, en nuestro trabajo, no describiremos las interrogativas indirectas, ya que consideramos las preguntas enunciados autónomos.

2.2.3 Oraciones interrogativas directas y foco

En este apartado trataremos algunas cuestiones relativas a la relación entre oraciones interrogativas y foco[51]. Nuestra aproximación en lo que se refiere a la relación foco-interrogativas sigue las ideas de Escandell (1999). Entre las posibles teorías que analizan la relación foco-interrogativas, seguimos la de Escandell (1999)

51 Pese a que el foco está más relacionado con aspectos pragmáticos y semánticos del lenguaje, lo abordamos aquí por su relación con fenómenos gramaticales de las interrogativas directas que trataremos en esta sección (como, por ejemplo, el orden de palabras).

por la sencillez de su planteamiento y por el hecho de que relaciona la noción de foco con otros fenómenos gramaticales que caracterizan a las interrogativas, tanto totales como parciales[52].

Por tanto, en esta sección no entraremos en la compleja cuestión de qué es el foco y cómo funciona. Nos limitaremos a la definición que maneja Escandell (1999) por considerarla adecuada y práctica para nuestra descripción gramatical de las interrogativas directas.

Para Escandell, la interrogación funciona como un operador, es decir, como un elemento que impone restricciones interpretativas sobre los constituyentes que caen bajo su dominio. Este dominio, denominado ámbito del operador, está determinado gramaticalmente y se corresponde con los constituyentes caracterizados como foco[53], que son aquellos que ocupan el primer plano informativo[54]. En esta aproximación, por tanto, el foco es aquella parte de la oración que posee mayor relevancia a nivel informativo. Además, en las interrogativas el foco se relaciona con el carácter de información nueva, frente a la información conocida o presupuesta por hablante y oyente. Como veremos a continuación, el foco se corresponde en las interrogativas con el segmento que define la incógnita o variable, por lo que es distinto en parciales y totales.

2.2.3.1 El foco en las interrogativas parciales

En las parciales, el foco (en mayúsculas en el ejemplo) lo constituyen las partículas o frases interrogativas:

(9) *¿QUIÉN escribió La bicicleta de Leonardo?*

El carácter de foco de las partículas interrogativas se marca a través de una serie de características gramaticales: especificidad léxica (cf. sección 2.4.3), prominencia prosódica (cf. sección 2.4.1) y posición inicial en la oración (cf. sección 2.4.2). La colocación en el inicio de la oración desencadena otro proceso sintáctico característico de las interrogativas parciales: la inversión del orden sujeto/verbo, como ocurre también con los constituyentes focalizados antepuestos (Zubizarreta, 1999, 64.3.4).

Como cualquier otro elemento caracterizado como foco, las partículas interrogativas pueden aparecer en su «posición canónica» en la oración si se marcan con una prominencia fonológica especial (sin que esto modifique la relación con el operador interrogativo). Esto ocurre en las denominadas «interrogativas de eco explicativas» (Escandell, 1999, p. 3935):

52 La mayoría de los estudios que se han ocupado de la relación foco-interrogativas han tratado solo las interrogativas parciales.

53 El carácter de operador focal de la interrogación y la necesidad de un dominio definido gramaticalmente para ese operador explican que interrogativas y construcciones remáticas compartan propiedades en su sintaxis interna. Estos factores también explican las coincidencias en la semántica de interrogativas y aseveraciones con foco (Rooth, 1992).

54 Hong (1995) también habla del «morfema-Q» como un operador sensible al foco.

(10a) —*No te lo vas a creer: ¡Acabo de ver a Schwarzenegger!*
(10b) —*¿Que has visto* A QUIÉN*?*

(Ejemplo tomado de Escandell, 1999, p. 3935)

Estas interrogativas se caracterizan por funcionar en el diálogo, ya que su estructura depende siempre de una interacción anterior sobre la que se apoyan. Por esta razón no consideramos las interrogativas de eco explicativas como preguntas y no nos ocuparemos de ellas en nuestro análisis.

Dejando el foco a un lado, los elementos restantes de la interrogativa parcial forman una presuposición, es decir, un contenido proposicional que hablante y oyente comparten. En (9), ese contenido proposicional conocido es *Alguien escribió La bicicleta de Leonardo*. La presuposición queda fuera del ámbito del operador interrogativo y, por lo tanto, no se cuestiona. Por eso las parciales admiten paráfrasis como (11):

(11) *Alguien escribió* La bicicleta de Leonardo*, ¿QUIÉN?*

(Ejemplo tomado de Escandell, 1999)

donde el contenido presupuesto aparece sintácticamente desgajado y solo el foco constituye la oración interrogativa.

2.2.3.2 El foco en las interrogativas totales

A diferencia de lo que ocurre en las parciales, en las totales el foco abarca en principio toda la estructura, de manera que la proposición entera cae bajo el ámbito del operador interrogativo:

(12) *¿Ha subido el gas?*
FOCO

Según Escandell (1999), como en el caso de las parciales, el orden de palabras característico de las totales (cf. sección 2.4.2.1), verbo/sujeto, es un medio gramatical para forzar la interpretación de toda la proposición como información nueva o en primer plano (interpretación «rem*á*tica» o de «juicio tético» de la proposición).

Como también ocurría en las parciales, puede focalizarse un determinado constituyente de la interrogativa total mediante medios gramaticales independientes (Kiefer, 1980) como la prominencia fonológica. En estos casos, el operador interrogativo solo actúa sobre el constituyente focalizado, de manera que solo este constituyente atrae hacia sí el sentido interrogativo, mientras que el resto de la proposición se interpreta como presuposición y «escapa» del dominio del operador:

(13) *¿Va a venir CAMINANDO?*
Presuposición: Va a venir.
Foco interrogado: ¿CAMINANDO?

(Ejemplo adaptado de Escandell, 1999)

La posibilidad de la paráfrasis:
(14) *¿Es caminando como va a venir?*

parece apoyar esta interpretación.

2.2.3.3 Cuantificadores y foco en las totales

Para Escandell (1999), los cuantificadores indefinidos en las interrogativas totales producen también un «efecto de focalización», de manera que el centro informativo se sitúa en ese constituyente, dejando el resto de la oración en segundo plano (presuposición):

(15a) *¿Gana MUCHO dinero?*
(15b) *¿Ha dejado ALGUIEN la puerta abierta?*

Este carácter de foco de los cuantificadores indefinidos en las interrogativas totales es el que lleva a interpretar los ejemplos de (15) como equivalentes a los de (16):

(16a) *¿CUÁNTO dinero gana?*
(16b) *¿QUIÉN ha dejado la puerta abierta?*

De hecho, según NGRALE (Real Academia Española, 2009, 42.7s y ss.), ciertas interrogativas totales se asimilan a las parciales y no se responden con *sí* o *no*, sino con respuestas que corresponden a alguna variable presente en la pregunta. Este sería el caso de las totales que contienen algún indefinido (16), que suelen responderse con la información que corresponde a la variable que el indefinido proporciona en la pregunta:

(17a) *¿Ha llamado ALGUIEN?* (¿QUIÉN ha llamado?) > tu madre
(17b) *¿Vas a ALGÚN SITIO?* (¿DÓNDE vas?) > al mercado

(Ejemplo tomado de Real Academia Española, 2009, 42.7s)

Para NGRALE, dentro de este grupo de totales que se «equiparan» a las parciales, aquellas que contienen expresiones cuantificativas:

(18) *¿Te costó mucho?*
(19) *¿Es muy tarde?*

constituyen un caso de interpretación más complejo, ya que puede aparecer el adverbio afirmativo o negativo como respuesta

(18a) *¿Te costó mucho?* > Sí. Cien euros.

pero en gran número de ocasiones, se omite.

(19a) *¿Te costó mucho?* > Cien euros.

(Ejemplos tomados de Real Academia Española, 2009, 42.7s)

Para la Academia, cuando el adverbio se omite

> no puede decirse que esas respuestas sean satisfactorias en términos sintácticos, en el sentido de que no se proporciona directamente en ellas la información que se solicita. Sin embargo, los juicios de valor que el hablante y el oyente comparten hacen que resulten casi siempre INFORMATIVAS, para el que las recibe; el que las contesta somete la información suministrada al juicio del que formula la pregunta, dando a entender con ello que debe ser él quien traduzca la respuesta a términos escalares.
> (Real Academia Española, 2009, 42.7t).

2.2.3.4 Negación y foco

En las interrogativas puede haber negación

(22) *¿Qué premiado por el Instituto Goethe **no** recogió el premio?*

Al igual que la interrogación, la negación es un operador y, por lo tanto, tiene un ámbito de acción. Este ámbito de acción es variable y puede afectar a un único constituyente o a toda la proposición. Por esta razón se distingue entre «negación interna» o «descriptiva», cuando esta afecta a un solo constituyente y «negación externa», «polémica» o «modal», cuando afecta a toda la proposición:

> En las oraciones enunciativas, la negación es interna cuando la proposición supone la aserción de una propiedad negativa; y es externa cuando indica el rechazo de una proposición afirmativa anterior.
> (Escandell, 1999, p. 3957)

Estos dos tipos de negación, interna (20) y externa (21), se dan también en las oraciones interrogativas:

(20) *¿En qué viviendas humanas no hay nunca moscas?*
(21) *¿No te parece que hace un buen día?*

Las estructuras negativas sobre las que opera la interrogación son distintas dependiendo del ámbito de la negación, de manera que el significado de la pregunta

con negación también es distinto dependiendo de si la negación es interna o externa.

Cuando la negación es interna (22), esta afecta tan solo al predicado, de manera que la interrogación actúa sobre este predicado negado dando lugar a una «predicación negativa simple que se interpreta como foco por defecto» (Escandell, 1999, p.. 3957):

(22a) ¿[Predicación negativa] ?
FOCO

(22b) ¿[*En qué viviendas humanas no hay nunca moscas*][55]?
FOCO

En el caso de la negación «externa», la estructura es más compleja: tenemos por un lado la negación y por otro una predicación afirmativa que está presupuesta. La negación es el foco, y la predicación afirmativa la presuposición:

(23a) NEGACIÓN [Predicación afirmativa]
FOCO Presuposición

(23b) ¿No [te parece que hace un buen día]?
FOCO Presuposición

La interrogación opera, por lo tanto, sobre una estructura compleja con un constituyente focalizado y, como ocurre en los casos de interrogativas con un elemento focalizado, el operador interrogativo actúa solo sobre el foco y deja fuera de su ámbito la predicación afirmativa:

(24a) ¿ NEGACIÓN ? [Predicación afirmativa]
FOCO

(24b) *¿No te parece que hace un buen día? > Te parece que hace un buen día, ¿no?*

Para Escandell (1999), el que la negación externa actúa sobre una proposición afirmativa lo demuestra el hecho de que esa proposición puede contener en su interior elementos de polaridad positiva (elementos que en las enunciativas correspondientes rechazan la presencia de la negación):

(25)
—¿No les habrá dicho ***algo****?*
—¿Qué les va a decir?

(Ejemplo adaptado de Escandell, 1999)

55 Las representaciones como la del ejemplo utilizadas en esta sección se han tomado de (Escandell, 1999).

Lógicamente, esto no es posible cuando la negación es interna

(26a) *¿No quiere llamarlo **también**? (vs. ¿No quiere llamarlo tampoco?)*
(26b) *¿No quiere intentarlo **siquiera**?*

(Ejemplo adaptado de Escandell, 1999)

La presencia de términos de polaridad positiva con negación es, por tanto, un indicador de que la negación es externa. Por otro lado, la presencia de términos de polaridad negativa es un indicador de que la negación es de tipo interno:

(27) *¿No ha avisado a **ninguno** de ellos?*

2.2.4 Características gramaticales que definen las oraciones interrogativas directas

El factor que presentábamos más arriba, la presencia de una incógnita que hace de la proposición contenida en una interrogativa una proposición abierta, es el que define como conjunto las oraciones interrogativas (tanto directas como indirectas). El carácter de enunciado autónomo define a su vez a las interrogativas directas frente a las indirectas. Por otra parte, hemos visto también que el carácter de la variable o incógnita define tres tipos de oraciones interrogativas directas: totales, disyuntivas y parciales. Pues bien, para cada uno de estos tipos de interrogativa directa, al tipo de variable se suman además una serie de características específicas que afectan a los siguientes ámbitos:

- curva entonativa;
- orden de constituyentes en la oración;
- presencia de determinados elementos característicos (partículas interrogativas): solo en las parciales.

Como veremos, la importancia de cada uno de estos ámbitos en la caracterización del tipo de interrogativa directa no es homogénea. La curva entonativa caracteriza fuertemente a las interrogativas totales (y, hasta cierto punto, a las disyuntivas), mientras que no es especialmente relevante para las parciales. El orden de constituyentes, sin embargo, distingue claramente a las parciales, mientras que no es especialmente distintivo en los otros dos tipos de interrogativas. Para terminar, la presencia de partículas afecta tan solo a las parciales.

Por otro lado, los tres ámbitos (curva entonativa, orden de constituyentes, presencia de elementos gramaticales característicos) no son, como veremos, igual de relevantes para la formalización de las preguntas. La curva entonativa queda totalmente fuera del ámbito de SpQA, ya que el *parser* está pensado para operar sobre texto escrito. Los otros dos ámbitos, sin embargo, ofrecen información interesante para la formalización en la gramática de SpQA.

A continuación describiremos las características propias de cada tipo de interrogativa en relación a los tres ámbitos mencionados: entonación, orden de palabras y presencia de elementos gramaticales característicos[56].

2.2.4.1 Curva entonativa

Tanto las interrogativas totales como las parciales poseen una curva entonativa característica[57]. En las siguientes secciones nos ocuparemos de su descripción detallada.

2.2.4.1.1 Curva entonativa en las totales

El esquema entonativo prototípico de las totales se caracteriza principalmente por su final descendente-ascendente o en «anticadencia».

Para la variante peninsular, Fernández Ramírez (Fernández 1951, I § 44 y ss.) describe este esquema de la siguiente manera: una rama inicial constituida por las sílabas anteriores al primer acento, con un movimiento ascendente que arranca de un nivel tonal un poco más elevado que el de la declarativa correspondiente; un cuerpo central descendente, hasta la última vocal acentuada; una rama final ascendente. Este patrón interrogativo se representa en el análisis de niveles como /2 3 1 3 ↑/ (Real Academia Española, 1973, p. 111), o como /(1 2 1) 1 2 ↑ / (Quilis, 1993, §14.5.), con un tono medio precedido por un nivel bajo y una juntura final ascendente.

Según los datos de Quilis (1985; 1993), Sosa (1991) o García (1996), en Canarias y el Caribe (Puerto Rico, Cuba, Venezuela...) la entonación no marcada para las totales es circunfleja, mientras que en Argentina, Colombia y México presenta otros contornos ascendentes distintos al peninsular.

Es interesante observar que, cuando no hay inversión sujeto/verbo (cf. sección 2.4.2.1), la entonación es el único elemento que distingue las interrogativas totales de las correspondientes enunciativas. Por eso, en la escritura, esta diferencia se marca mediante el uso de los signos de interrogación tanto al inicio como al final de la interrogativa.

2.2.4.1.2 Curva entonativa en las parciales

El esquema entonativo básico de las interrogativas parciales es bastante semejante al de una declarativa: un patrón de «cadencia», donde la partícula interrogativa se sitúa en la cima de la curva entonativa a la que sigue una melodía descendente hasta el final. Su representación, en el análisis de niveles, es la siguiente / (1 2) 1 1 ↓ /.

En el caso de las parciales, no parece haber variaciones dialectales en la curva entonativa.

56 En el caso de las disyuntivas no disponemos de información teórica en relación a los tres ámbitos, de manera que solo las trataremos en aquellos casos en los que esto sea posible.

57 Siempre que dispongamos de datos trataremos tanto la curva entonativa de la variante peninsular como la de otras variantes del español.

La similitud con las declarativas en lo que se refiere a la curva entonativa parece responder al hecho de que este tipo de interrogativas se diferencia claramente de las declarativas correspondientes por otros rasgos fuertemente característicos: el orden de palabras y la presencia de partículas interrogativas. Estos dos factores caracterizan suficientemente a las parciales como para distinguirlas de las declarativas. En el caso de las totales, como veremos, no existen partículas interrogativas específicas ni un orden de elementos tan marcado como en las parciales, de manera que la entonación es la principal característica que sirve para diferenciarlas de las declarativas correspondientes.

2.2.4.1.3 Curva entonativa en las disyuntivas

En el caso de las interrogativas disyuntivas, la entonación se fragmenta en dos grupos melódicos: el primero presenta básicamente el esquema de las totales; el segundo presenta una entonación descendente.

Esta entonación característica es la que permite, de hecho, distinguir claramente entre interrogativas disyuntivas e interrogativas totales con una disyunción. De hecho, en la escritura, esta distinción puede no ser posible:

(28) *¿Vendrás mañana o pasado?*

En (28) hay dos lecturas posibles:

(28a) *Vendrás mañana o pasado* [¿sí?/¿no?]
(28b) *Vendrás* [¿mañana? /¿pasado?]

2.2.4.2 Orden de constituyentes

Las variaciones en la colocación de los constituyentes de una oración son un medio gramatical de primer orden para denotar diferencias gramaticales, tanto sintácticas como discursivas (Escandell, 1999, p. 3951). Como veremos a continuación, el orden de palabras es un factor especialmente interesante en el caso de las interrogativas parciales (mucho menos para las totales y las disyuntivas).

2.2.4.2.1 Orden de constituyentes en las totales

Según Escandell (1999), el orden no marcado de las interrogativas totales es verbo/sujeto, con el resto de los argumentos tras el sujeto en orden relativamente libre[58]:

(29) *¿Es Fernando de Rojas el autor de La Celestina?*

Dicho orden queda muchas veces diluido por la elisión del sujeto en español.

Este orden parece responder a cuestiones discursivas relacionadas con la

58 A diferencia de las parciales, no disponemos de datos de corpus sobre el orden de los argumentos en las interrogativas totales.

estructuración de la información. Como hemos visto al tratar la cuestión del foco en las totales (cf. sección 2.3.3), para Escandell (1999) en las interrogativas totales la interrogación opera sobre toda la proposición, lo que hace necesario:

> [...] caracterizar dicho contenido como un 'juicio tético' (es decir, como una estructura remática, en la que todo el contenido proposicional se caracteriza como información nueva o de primer plano; 64.2), para que quede así bajo el dominio del operador interrogativo. El orden verbo/sujeto típicamente produce este efecto, y ello explica que ésta sea la disposición de constituyentes no marcada en las interrogativas totales.
> (Escandell, 1999, p. 3935)

Las interrogativas totales admiten también la colocación del sujeto antes del verbo. La colocación del sujeto antes del verbo parece tener también consecuencias en la estructura informativa de una interrogativa total y, por lo tanto, en su significado. La anteposición del sujeto provoca la interpretación de la interrogativa total como una estructura más compleja en la que la interrogación opera sobre una proposición ya cerrada, es decir: sobre una proposición completa preexistente (Escandell, 1999). Se trataría por lo tanto de «metaproposiciones»:

(30) *¿Pedro viene mañana?*
¿[Proposición]?

(Representación tomada de Escandell, 1999, p. 3953)

Según Escandell (1999), esta interpretación se debe a que la colocación del sujeto antes del verbo se relaciona con los «juicios categóricos», estructuras bimembres del enunciado donde el sujeto es temático o presupuesto y el predicado es remático o focal. En esta modalidad informativa, la relación entre el sujeto y el predicado se asevera, lo que, a priori, es incompatible con la interrogación. Esto provoca que en los casos en los que el sujeto se antepone al verbo la interrogación no actúe generando una proposición abierta, pues al actuar sobre una proposición ya aseverada (cerrada), actúa en un segundo nivel. Esta diferencia en la estructuración de la información en las totales es la que permite explicar el contraste entre:

(31a) *¿Ha hecho Juan el más mínimo esfuerzo por ayudarme?*
(31b) **¿Juan ha hecho el más mínimo esfuerzo por ayudarme?*

(Ejemplos tomados de Escandell, 1999, p. 3953)

En (31a) la interrogación puede legitimar la locución negativa ya que ambas forman una única estructura proposicional; en (31b), sin embargo, la interrogación opera sobre una proposición completa que contiene un término de polaridad negativa sin legitimar, de modo que el operador interrogativo no puede hacer nada para legitimar su presencia (Escandell, 1999, p. 3953).

La relación entre orden sujeto/verbo y estructuración de la información implica también que, para marcar en una total un elemento como tema, este debe

aparecer caracterizado sintácticamente como tal, desgajado, en una posición externa a la oración interrogativa (fuera por tanto del arranque de la entonación interrogativa):

(32) *Carlos, ¿te ha vuelto a llamar?*

(Ejemplo tomado de Escandell, 1999)

Ciertos tipos de interrogativas totales cuyo uso se restringe al diálogo pueden presentar un orden de constituyentes alternativo. Estas interrogativas totales pueden contener cualquier estructura oracional (con su orden característico, que puede ser distinto al de la interrogativa total que acabamos de ver) que corresponda a la forma de un enunciado atribuido a otro. De este modo, en el interior de estas interrogativas pueden aparecer oraciones enunciativas, imperativas, optativas o exclamativas:

(33)
— *Adivina qué me ha dicho.*
— ¿Ven a mi despacho inmediatamente?

(34)
— *Ojalá llueva.*
— ¿Ojalá llueva? *Entonces sí que pasaríamos calor...*

(35)
— *¡Qué bien me encuentro!*
— ¿Qué bien me encuentro? *Ya me conozco esa canción. Lo que tú quieres hacer es levantarte, y el médico ha dicho que necesitas reposo, así que ni lo sueñes.*

(Ejemplos tomados de Escandell, 1999, p. 3953)

Como ya hemos indicado, este tipo de interrogativas totales funcionan tan solo en el discurso, y, por lo tanto, no son preguntas. El orden esperable (al menos teóricamente) para las preguntas con forma de interrogativa total parece ser aquel en el que se produce inversión sujeto/verbo.

2.2.4.2.2 Orden de constituyentes en las parciales

A diferencia de lo que ocurre con las interrogativas totales (y con las oraciones del español en general, donde el orden de elementos suele ser bastante libre), las interrogativas parciales parecen presentar un orden prototípico de constituyentes. Este orden se caracteriza por dos rasgos:

- la colocación de una partícula interrogativa al inicio de la estructura;
- la inversión del orden sujeto/verbo.

El orden prototípico de constituyentes en las interrogativas parciales sería entonces el siguiente:

Partícula Interrogativa + Verbo + Argumentos

Los datos de corpus parecen apoyar esta afirmación. En un estudio sobre la colocación de los constituyentes en las interrogativas parciales, Gayo (2010) documenta este orden en un 96.7% de los casos (el resto de los casos pertenecen a otras ordenaciones que trataremos más abajo).

En cuanto a los argumentos postverbales, Gayo (2010) observa que estos muestran ciertas preferencias de ordenación, aunque ningún patrón de ordenación parece constituirse como prototípico (cf. Gayo, 2010, para más detalles).

Entre la partícula interrogativa y el verbo pueden colocarse adverbios de negación (36), adverbios de frecuencia (37) y también los adverbios aspectuales *ya* (38) y *todavía* (39) (Contreras, 1999, p. 1939).

(36) *¿Qué misión espacial Apollo* no *logró alcanzar la luna?*
(37) *¿Qué lugar* siempre *se visita en Granada?*
(38) *¿Por qué* ya *no se dedica al mundo del espectáculo Pepa Flores?*
(39) *¿Qué ciudad portuguesa* todavía *conserva los tranvías de los años 40 como medio de transporte?*

Junto a la presencia de partículas interrogativas específicas, el orden de elementos es el rasgo más característico de la sintaxis de las interrogativas parciales. Como ya hemos visto, la colocación de la partícula interrogativa y la inversión sujeto/verbo son fenómenos que derivan de caracterizar gramaticalmente el constituyente interrogado como foco.

NGRALE señala una serie de características relacionadas con la posición del sujeto en las interrogativas parciales. Se observa que el sujeto «suele ir tras el verbo» (Real Academia Española, 2009, 42.9c), y que puede ir en medio de la frase verbal cuando tenemos perífrasis (40) o tiempos compuestos (41).

(40) *¿Dónde podría* usted *comprar "pastéis de belém"?*
(41) *¿Cuándo habrá* el gobierno *solucionado la crisis?*

Por otro lado, según NGRALE, cuando la frase interrogativa está constituida por un adverbio de sentido causal o modal (*por qué*, *cómo*, *a santo de qué*, *a cuento de qué, hasta qué punto*, *de qué modo*, etc.) el sujeto puede aparecer antepuesto al verbo (Real Academia Española, 2009, 42.9d):

(42) *¿Hasta qué punto* el BCE *desea que España salga de la crisis?*

Para la Academia, esta ordenación se da porque esos adverbios interrogativos son más externos al predicado verbal que los de tiempo o lugar (Real Academia Española, 2009, 42.9g). Otros autores también recogen esta ordenación, pero ampliando el tipo de adverbios interrogativos que la permiten: para Goodall (2004),

los «interrogativos-adjuntos» permiten con más facilidad la colocación del sujeto antes del verbo; según Torrego (1984), la inversión de sujeto no es obligatoria con *cómo*, *cuándo*, *por qué*, *en qué medida.*

En Gayo (2010) se documentan dos órdenes distintos al orden «prototípico»: el que acabamos de ver, que se denomina allí anteposición tipo B[59], y otro distinto, denominado anteposición tipo A. En la anteposición tipo A el hablante coloca un argumento antes del interrogativo, generalmente el sujeto:

(43) ¿Ustedes *de dónde vienen?*
(44) ¿Tú *qué piensas de la Feria?*

(Ejemplos tomados de Gayo, 2010)

En cuanto a la anteposición tipo B, en la línea de lo que recogen los autores citado (cf. *supra*), en Gayo (2010) esta solo se documenta en casos con interrogativos funcionando como adjuntos: *por qué* (19 casos), *de qué* (2 casos), *desde cuándo* (1 caso), *por cuánto* (1 caso):

(45) ¿Por qué la revolución *no la harán unos tíos cachas?*
(46) ¿Por qué Sepúlveda *no quiso que lo operaran de nuevo?*

(Ejemplos tomados de Gayo, 2010)

La siguiente tabla recoge los datos generales sobre la ordenación prototípica y los dos tipos de anteposición en el estudio de Gayo (2010):

ORDENACIÓN	**Nº. CASOS**	**Valor en %**
Orden prototípico	2198	96,7
Anteposición tipo A	52	2,28
Anteposición tipo B	23	1

Tabla 1: Orden prototípico *vs.* órdenes con anteposición en Gayo (2010).

Como vemos, en el estudio de Gayo (2010), la anteposición tipo A es más común que la tipo B. En nuestra sección de corpus volveremos sobre estos órdenes «alternativos» (cf. capítulo 3).

Siguiendo con otras posibilidades de ordenación para las parciales, en el español del Caribe existe otro orden posible, con el sujeto precediendo al verbo, preferentemente si el sujeto es pronominal (Contreras, 1999, p. 1939; Real Academia Española, 2009, 42.9h):

59 En Gayo (2010) la anteposición tipo B no solo incluye la interposición del sujeto, sino también de otros constituyentes oracionales como el objeto directo.

(47) *¿Qué* tú *sabes?*

También puede darse un orden de constituyentes distinto al prototípico cuando tenemos un sintagma interrogativo complejo (Contreras, 1999, p. 1940).

(48) ¿A cuál de sus amigos de toda la vida *Marcos le robó la novia?*
(49) ¿Por qué razón entre las muchas posibles *Carlos habrá hecho eso?*

En conclusión: en las interrogativas parciales se da un orden de constituyentes marcado, con el interrogativo al inicio de la estructura e inversión sujeto/verbo. Los datos de corpus de Gayo (2010) parecen apoyar la hipótesis de que este orden es el prototípico en las interrogativas parciales.

La colocación del interrogativo tras el verbo, como en el caso de las interrogativas de eco explicativas (12b), es característico de interrogativas con fines distintos a la demanda de información, propias de contextos discursivos. Esta ordenación denota usos ligados a cuestiones pragmático discursivas.

Junto al orden prototípico parecen poder darse también otros, de uso más marginal. En Gayo (2010) se documentan dos: uno en el que el se coloca un constituyente (generalmente el sujeto) entre la frase interrogativa y el verbo cuando el interrogativo funciona como adjunto (45), (46), y otro en el que se coloca un constituyente (generalmente también el sujeto) antes del interrogativo (43), (44). En el español del Caribe se da otro orden en el que el sujeto se coloca entre el interrogativo y el verbo (47).

En la sección 1.2 del capítulo 3, aportaremos más datos sobre los distintos tipos de ordenación en las parciales.

2.2.4.3 Uso de partículas interrogativas

Hemos adelantado ya que el rasgo que más define las interrogativas parciales es la presencia de una partícula interrogativa, generalmente, al inicio de la estructura:

(50) ¿Quién *firmó la declaración de la independencia de Vermont?*

También hemos visto que si la partícula interrogativa se sitúa en otro lugar (cf. *supra:* «interrogativas de eco explicativas»), esto conlleva generalmente un significado marcado para la oración interrogativa, un uso específico distinto al de demanda de información que parece ligado a contextos discursivos.

Teniendo en cuenta la relevancia de las partículas interrogativas en el funcionamiento de las interrogativas parciales, en los siguientes apartados nos ocuparemos de la descripción de sus características gramaticales más importantes. En primer lugar abordaremos cuestiones generales, para luego analizarlas una a una, deteniéndonos en sus particularidades. Más adelante, cuando tratemos la semántica de las oraciones interrogativas, trataremos en profundidad sus aspectos semánticos (cf. sección 2.3 del presente capítulo).

2.2.4.3.1 Partículas interrogativas: características generales

Las partículas interrogativas que se usan para construir interrogativas parciales son las siguientes: *cuándo*, *dónde*, *cómo*, *por qué*, *qué*, *quién*, *cuál y cuánto*[60]. No tenemos en cuenta en nuestro análisis la forma *cúyo*, que funciona como adjetivo referido a un sustantivo para indicar «cosa poseída» (como su homónimo relativo), pues su uso es muy raro en la lengua actual (Real Academia Española, 2009, 22.14y) y parece haber quedado reducido a la lengua literaria (Porto Dapena, 1997, p. 51).

A nivel paradigmático, las partículas interrogativas (51) están directamente relacionadas con las partículas exclamativas (también tónicas) (52) y con los relativos (átonos) (53), ya que los tres tipos de palabras comparten un origen común (Alarcos, 1994).

(51) ¿Quién *descubrió América?*
(52) ¡Quién *te ha visto y* quién *te ve!*
(53) *Me da igual* quien *sea: eso no se puede hacer.*

Porto Dapena habla de hecho de «interrogativo-exclamativos» (Porto Dapena, 1997, p. 41), y Alarcos de «correlación entre las unidades relativas y las interrogativas o exclamativas» (Alarcos, 1994, p. 109).

En relación a su comportamiento sintáctico, las partículas interrogativas pueden funcionar aisladas (54) o formando constituyente sintáctico con otros elementos (55), (56):

(54) ¿Cuál *es el segundo pico de montaña más alto en el mundo?*
(55) ¿Qué envergadura *tiene el Airbus 380?*
(56) ¿Cuál de los Beatles *tocaba la batería?*

A nivel morfosintáctico, pueden desempeñar las siguientes funciones en la oración interrogativa: pronombre (57), determinante (58), adverbio (59) y adjetivo cuantificador (60):

(57) ¿Qué *significa mecate en nahuatl?*
(58) ¿Cuántos rusos *han estado en la Luna?*
(59) ¿Dónde *está el Mar de la Tranquilidad?*
(60) ¿Cuán largo *es el trayecto de Madrid a Palencia?*

De hecho, se caracterizan por su **especificidad funcional**: cada palabra interrogativa puede desempeñar unas funciones concretas:

- Pueden funcionar como **pronombres y determinantes**: *qué* (61), *cuánto* (62) y *cuál* (63).

(61a) ¿Qué *inventó Marconi?*
(61a) ¿Qué científico descubrió la penicilina?

60 Consideramos, siguiendo a la Academia, la forma *cuán* una variante de *cuánto* (cf. *infra*, p. 93).

(62a) ¿A cuántos *timó Mario Conde?*
(62b) ¿Cuántos minutos *tiene una hora?*
(63a) ¿Cuál *es la distancia entre Madrid y Sevilla?*
(63b) *¿A* cuáles pintores *se denomina expresionistas?*

Como señala la Real Academia Española (2009, 22.13j) y Porto Dapena (1997), *qué* y *cuál* alternan en su uso como determinantes en los grupos nominales de interpretación anafórica o catafórica, si bien *cuál* está siendo progresivamente sustituido por *qué* en este tipo de usos, especialmente en el área rioplatense y en Europa (Real Academia Española, 2009, 22.14a).

Cuánto puede funcionar también como adverbio:

(64) ¿Cuánto *duerme un koala al día?*

Y su variante *cuán* como adjetivo cuantificador:

(65) ¿Cuán pesado *puede llegar a ser un turismo?*

- Puede funcionar **solo como pronombre**: *quién.*

(66) ¿Quién *es el autor de la Gioconda?*

- Pueden funcionar **solo como adverbios** *dónde* (67), *cómo* (68), *cuándo* (69) y *por qué* (70).

(67) ¿Dónde *está el Empire State?*
(68) ¿Cómo *se hace el vino?*
(69) ¿Cuándo *se inició el Pleistoceno?*
(70) ¿Por qué *el cielo es azul?*

Este grado de especialización léxico funcional determina a su vez que podamos distinguir entre un conjunto de partículas interrogativas que funcionan prototípicamente como adjuntos (*cuándo*, *dónde*, *por qué, cómo*) y un conjunto que funciona prototípicamente como argumentos (*quién*, *qué*, *cuál*, *cuánto*)[61]. Esta especificidad de las partículas interrogativas constituye una ventaja para su formalización, ya que cada forma lingüística puede asociarse a una serie de funciones prototípicas. De hecho, como veremos (cf. capítulo 4), esta característica constituye una de las bases de nuestra formalización de las partículas interrogativas en la gramática de SpQA[62].

Ya hemos adelantado que otra de las características de las partículas interrogativas es el hecho de que pueden formar un constituyente sintáctico (frase interrogativa) con otras unidades:

61 *Cuán* funciona siempre como modificador.
62 Santalla (2002, p. 330) explota este mismo hecho en la formalización de las interrogativas indirectas parciales.

(71) ¿Qué *enseña Vital do Rego?*
(72) ¿Qué film *recibió el primer Premio de la Academia a la mejor película?*
(73) ¿De qué animal *vienen las alas de búfalo?*

Las partículas interrogativas que funcionan como determinantes (*qué, cuál y cuánto*), pueden ir seguidas de sustantivos (o sustantivos modificados) con los que forman una frase nominal:

(74) ¿Qué libro de aventuras de Verne *se desarrolla en un globo?*
(75) ¿Cuál animal *vive menos tiempo?*
(76) ¿Cuánto dinero *gana Nadal al año?*

Todas las partículas interrogativas que funcionan como pronombre pueden ir seguidas de una frase preposicional con la que forman un constituyente sintáctico. La frase preposicional tiene significado generalmente partitivo, modifica a los interrogativos que funcionan como núcleos pronominales de la construcción y puede ir situada tanto antes (77) como después del verbo (78):

(77a) ¿Quién de los escritores del 27 *escribió Poeta en Nueva York?*
(77b) ¿Cuál de los Bravos *era extranjero?*
(77c) ¿Cuántos de los países de la UE *están en crisis?*
(77d) ¿Qué de lo que dijo Rajoy en su último discurso *es más importante?*

(78a) ¿Quién *escribió Poeta en Nueva York* de los escritores del 27*?*
(78b) ¿Cuál *era extranjero* de los Bravos*?*
(78c) ¿Cuántos *están en crisis* de los países de la UE*?*
(78d) ¿Qué *es más importante* de lo que dijo Rajoy en su último discurso*?*

Las funciones que desempeñan este tipo de constituyentes interrogativos complejos son generalmente argumentales, en concreto: complemento directo, sujeto o predicativo.

Cuán puede ir seguido de un adjetivo (79) o de un adverbio (80):

(79) ¿Cuán preocupado *está Rajoy?*
(80) ¿Cuán lejos *queda Barcelona de París?*

Todas las partículas interrogativas (al igual que los constituyentes interrogativos complejos que acabamos de presentar) excepto *por qué*, pueden ir precedidos de una preposición con la que también forman un constituyente sintáctico:

(81a) ¿A qué altura *está la capa de ozono?*
(81b) ¿Con quién *está casado Tom Cruise?*
(81c) ¿En cuál de las regiones naturales *el suelo permanece helado durante casi todo el año?*

(81d) ¿Por cuánto *usted podía alquilar un escarabajo Volkswagen en 1966?*
(81e) ¿De dónde *provienen los cerdos de Guinea?*
(81f) ¿Desde cuándo *Portugal es una república?*

A *cómo* tampoco lo preceden preposiciones, exceptuando el uso coloquial con la preposición *a*, donde *cómo* tiene un valor equivalente a *cuánto*:

(82) ¿A cómo *vendió el Santander sus últimas acciones?*

2.2.4.3.2 Partículas interrogativas: particularidades

En los siguientes apartados analizaremos una a una las partículas interrogativas y sus características gramaticales. Empezaremos por aquellas que funcionan prototípicamente como adjuntos (*cuándo*, *dónde*, *cómo* y *por qué*) y a continuación trataremos aquellas que funcionan prototípicamente como argumentos (*quién*, *cuánto*, *cuál* y *qué*).

CUÁNDO

Adverbio.
(83) ¿Cuándo *se descubrió la pólvora?*

Funciona siempre como adjunto.
Según la Real Academia Española (2009, 22.15k), se puede combinar tan solo con las siguientes preposiciones: *de*, *desde*, *hasta* y *para*.

DÓNDE

Adverbio.
(84) ¿Dónde *se celebraron los Juegos Olímpicos de 2012?*

Funciona siempre como adjunto.
Puede ser término de las siguientes preposiciones: *a, de, desde, hacia, hasta, para, por,* y, en ciertos casos, *en* (Real Academia Española, 2009, 22.15c). Al respecto, Bosque (1984) señala que:

> Algunas preposiciones, entre las que figuran *en*, *con*, *ante* y *sobre* rechazan el adverbio *dónde* en todos los contextos [...]
> (Bosque, 1984, p. 266).

Sobre la combinación de la preposición *en* + *dónde* la Academia observa sin embargo que:

> No se percibe en los adverbios interrogativos y relativos de lugar la redundancia que se manifiesta en los demostrativos. Se considera incorrecta la secuencia en allí [...],

> pero se admite *en donde* (22.8o), y también *en dónde*, en pares como ¿Dónde trabaja? ~ ¿En dónde trabaja?
> (Real Academia Española, 2009, 22.15f)

Existe también el adverbio *adónde*, que se utiliza sobre todo con verbos de movimiento (85) (Real Academia Española, 2009, 22.15c).

(85) ¿Adónde *se dirigía Colón en origen?*

Adónde puede alternar con *a dónde* y también con *dónde* (86).

(86) *¿Adónde va la sociedad actual?* / *¿A dónde va la sociedad actual?* / *¿Dónde va la sociedad actual?*

Este tipo de alternancia se da también entre *en dónde* y *dónde*, al estar integrado en el adverbio el significado locativo de la preposición *en* (87).

(87) *¿En dónde está el Timanfaya?* / *¿Dónde está el Timanfaya?*

CÓMO

Adverbio.

(88) ¿Cómo *se contagia el virus del ébola?*

En cuanto a las funciones sintácticas que puede desempeñar, *cómo* puede funcionar como adjunto (89) o como predicativo (90):

(89) *¿Cómo se obtiene el área del hexágono?*
(90) *¿Cómo es el ambiente del cocodrilo tortuga laúd?*

Cómo se usa en la locución *cómo así (que)* en el español de ciertas zonas de Centroamérica, el Caribe continental y el área andina (Real Academia Española, 2009, 22.16k):

(91) *¿Y cómo así logró Bárcenas hacer un capital?*

Solo se combina con la preposición *a* con significado cuantitativo en usos coloquiales:

(92) *¿A cómo está el euro respecto al dólar?*

POR QUÉ

El caso de *por qué* (93) es más complejo en cuanto a su caracterización como unidad.

(93) ¿Por qué *la cerveza sabe amarga?*

NGRALE señala que en casos como (93) algunos autores analizan *por qué* como una locución adverbial (unidad léxica) similar a las existentes en otras lenguas (inglés *why*), mientras que otros lo analizan como un grupo preposicional (unidad sintáctica). Para la Academia «existen argumentos a favor y en contra de las dos opciones que se mencionan» (Real Academia Española, 2009, 22.16n), si bien en NGRALE se incluye su análisis entre el resto de las partículas interrogativas, porque así se ha hecho tradicionalmente y también porque la expresión adverbial cuenta con equivalentes en otras lenguas (Real Academia Española, 2009, 22.16n). En nuestro análisis también consideramos casos como los de (93) como locuciones adverbiales.

Un caso distinto lo constituye la combinación de la preposición *por* (en el sentido de 'a favor de' o 'en vez de') con el interrogativo *qué*. Este grupo preposicional se utiliza con verbos transitivos que también rigen complementos preposicionales que se construyen con la preposición *por*. En estos casos, *qué* también puede funcionar como pronombre (94) o como determinante (95).

(94) *¿Por qué se le cambió Manhattan a un indio?*
(95) *¿Por qué joya sacrificó el aventurero su vida?*

La locución adverbial *por qué* no se combina con otras preposiciones.
Su función prototípica es la de adjunto.

QUIÉN

Pronombre; variable en número e invariable en género: *quién* (96)*, quiénes* (97).

(96) ¿Quién *inventó el cinematógrafo?*
(97) ¿Quiénes *perpetraron el 11-S?*

Lo más común es que aparezca con verbos en tercera persona, aunque, como todos los interrogativos que admiten plural, también puede concordar con el verbo en primera o segunda persona del plural (Real Academia Española, 2009, 22.2k):

(98) *¿Quiénes conseguimos la victoria?*
(99) *¿Quiénes estabais allí?*

Puede combinarse con las siguientes preposiciones: *a*, *con*, *de*, *para*, *por.*

(100) *¿Para quién trabaja 007?*

Funciona prototípicamente como sujeto (101).

(101) *¿Quién asesinó a Kennedy?*

También puede funcionar como predicativo (102).

(102) *¿Quién es Barack Obama?*

Combinado con la preposición *a*, funciona como complemento directo (103) y como complemento indirecto (104); con otras preposiciones, como adjunto (105) y como suplemento (106).

(103) *¿A quién imita mejor Carlos Latre?*
(104) *¿A quién le salva la vida Lassie?*
(105) *¿Con quién está casado Imanol Arias?*
(106) *¿A quién se encomiendan las causas perdidas?*

Admite complementos partitivos con forma de frase preposicional, que, como hemos visto, puede ir situada antes del verbo (107) o después (108).

(107) ¿Quién de los diputados *es más joven?*
(108) ¿Quién *es más joven* de los diputados*?*

También podemos tener oraciones de relativo modificando al interrogativo o al grupo sintáctico que forma (Real Academia Española, 2009, 22.2l), que, al igual que la frase preposicional, pueden ir situadas antes (109) o después del verbo (110).

(109) *¿Quiénes* que conozcamos *asistieron a la última reunión con el presidente?*
(110) *¿Quiénes asistieron a la última reunión con el presidente* que conozcamos*?*

CUÁNTO

Pronombre (111), determinante cuantificativo (112) o adverbio (113). Como pronombre y determinante varía en genéro y número: *cuánto, cuánta, cuántos, cuántas*.

(111) *¿A* cuántos *afectó la bomba de Hiroshima?*
(112) ¿Cuántas veces *se debe regar una orquídea a la semana?*
(113) ¿Cuánto *costó el Titanic?*

Como todos los interrogativos que admiten el plural, también puede concordar con el verbo en primera o segunda persona del plural (Real Academia Española, 2009, 22.2k):

(114) *¿Cuántos tenemos la culpa de la situación económica actual?*
(115) *¿Cuántos cobraréis la jubilación?*

Puede combinarse con las siguientes preposiciones: *a*, *con*, *de*, *en*, *para*, *por.*
(116) *¿Por cuánto se vendió el último cuadro de Picasso subastado?*

Su función prototípica es la de complemento directo (117).
(117) *¿Cuánto ha invertido el Santander en Brasil?*

Sin preposición, puede funcionar también como sujeto (118) y como adjunto (119).

(118) *¿Cuántos trabajadores secundaron la última huelga?*
(119) *¿Cuánto debe dormir al día un niño?*

Con la preposición *a* puede funcionar como complemento indirecto (120) o directo de persona (120a); con la preposición *de*, como predicativo (121); con otras preposiciones, como adjunto (122) y suplemento (123).

(120) *¿A cuántos votantes les debe el PP su última victoria?*
(120a) *¿A cuántos lectores alimenta El País?*
(121) *¿De cuántos bits es el último procesador de Intel?*
(122) *¿Por cuánto se vendió el cuadro más caro de la historia?*
(123) *¿A cuánto asciende la deuda española?*

Admite complementos partitivos (124), (125) y complementos adjetivales (126), (127) (Real Academia Española, 2009, 22.14l). En ambos casos, se trata de frases preposicionales que, como hemos visto, pueden ir tanto antes (124), (126) como después del verbo (125), (127).

(124) *¿Cuántos* de los alumnos españoles *se presentaron al último examen de selectividad?*
(125) *¿Cuántos se presentaron al último examen de selectividad* de los alumnos españoles*?*
(126) *¿Cuánto* de aprovechable *tiene la carne de cocodrilo?*
(127) *¿Cuánto tiene la carne de cocodrilo* de aprovechable*?*

También podemos tener oraciones de relativo modificando al interrogativo o al grupo sintáctico que forma (Real Academia Española, 2009, 22.2l), situadas antes (128) o después (129) del verbo.

(128) *¿Cuántos políticos* que sean médicos *conoces?*
(129) *¿Cuántos políticos conoces* que sean médicos*?*

Según NGRALE, *cuánto* se acopoca en *cuán* ante adjetivos, adverbios y sus grupos sintácticos y ante las locuciones correspondientes (Real Academia Española,

2009, 22.14o), funcionando como adjetivo cuantificador (Real Academia Española, 2009, 22.1f y ss.):

(130) ¿Cuán *largo era ese pasaje?*

(Ejemplo tomado de Real Academia Española, 2009, 22.1f)

Este uso está más extendido en el español americano y es mucho menos frecuente en el español europeo. En el español europeo, así como en el área rioplatense, se utiliza la variante *cómo de* + grupo adjetival o adverbial para estructuras como las de (130) en las que se pregunta por el grado en el que se da una propiedad (Real Academia Española, 2009, 22.14q):

(131) ¿Cómo de *largo era ese pasaje?*

CUÁL

Pronombre (132) y determinante (133). Varía en número: *cuál, cuáles* (134).

(132) ¿Cuál *de los Beatles tocaba la batería?*
(133) ¿Cuál libro *es el más largo que se ha escrito?*
(134) ¿Cuáles *de estos ríos son afluentes: Miño, Sil, Guadiana, Ebro?*

Como todos los interrogativos que admiten el plural, también puede concordar con el verbo en primera o segunda persona del plural (Real Academia Española, 2009, 22.2k):

(135) *¿Cuáles fuimos acusados?*
(136) *¿Cuáles iréis allí?*

Ya hemos visto que *cuál* y *qué* alternan en el uso como determinante (137), y que, especialmente en el área rioplatense y en Europa, se está produciendo la sustitución progresiva de *cuál* por *qué* (Real Academia Española, 2009, 22.14a).

(137) *¿Cuál río es el más largo del mundo?* / *¿Qué río es el más largo del mundo?*

Puede combinarse con las siguientes preposiciones: *a*, *de*, *en*, *por.*

(138) *¿Con cuál de los Beatles se juntó Yoko Ono?*

Admite complementos partitivos desempeñados por frases preposicionales; la frase preposicional puede ir antes (139) o después del verbo (140).

(139) *¿Cuál* de los Beatles *era el más joven?*
(140) *¿Cuál era el más joven* de los Beatles*?*

También podemos tener oraciones de relativo modificando al interrogativo o al grupo sintáctico que forma (Real Academia Española, 2009, 22.2l), que pueden ir antes (141) o después (142) del verbo.

(141) *¿Cuáles de tus amigos* que conozcamos *van a la fiesta?*
(142) *¿Cuáles de tus amigos van a la fiesta* que conozcamos*?*

En cuanto a las funciones sintácticas que desempeña *cuál,* la prototípica es la de miembro de una estructura ecuativa (143).

(143) *¿Cuál de los afluentes del Miño es el más largo?*

También puede funcionar como sujeto (144) y complemento directo (145).

(144) *¿Cuál de los marinos del s. XV descubrió América?*
(145) *¿Cuál de los libros que escribió Cervantes tiene la Hispanic Society en Nueva York?*

Con la preposición *a*, puede funcionar como complemento directo de persona (146) y como complemento indirecto (147); con otras preposiciones puede funcionar como adjunto (148) o suplemento (149).

(146) *¿A cuál de los poetas del Barroco homenajeó la generación del 27?*
(147) *¿A cuál banco le atribuyen los expertos el origen de la crisis?*
(148) *¿Con cuál arma asesinaron a Kennedy?*
(149) *¿A cuál de los dioses se encomendaba Odiseo?*

QUÉ

Puede funcionar como pronombre (150) y como determinante (151), y es invariable en género y número.

(150) ¿Qué *inventó Edison?*
(151) ¿Qué empresa *fabrica el Escarabajo?*

Puede combinarse con las siguientes preposiciones: *a*, *bajo*, *con*, *contra*, *de*, *en*, *entre*, *para*, *por*, *sobre*.

(152) *¿De qué país era originario Kafka?*

Admite ciertos complementos partitivos desempeñados por frases preposicionales; la frase preposicional puede ir antes (153a) o después del verbo (153b).

(153a) *¿Qué* de todo lo expuesto por el presidente *es lo más improbable?*
(153b) *¿Qué es lo que más improbable* de todo lo expuesto por el presidente*?*

También podemos tener oraciones de relativo modificando al interrogativo o al grupo sintáctico que forma, situadas antes (154) o después (155) del verbo.

(154) *¿Qué* que se sepa *ha decidido Netflix sobre España?*
(155) *¿Qué ha decidido Netflix sobre España* que se sepa*?*

Puede funcionar como complemento directo (156), como sujeto (157) y como predicativo (158).

(156) *¿Qué descubrió Colón en 1492?* / *¿Qué continente descubrió Colón en 1492?*
(157) *¿Qué produjo el Tsunami de 2000?* / *¿Qué fenómeno natural produjo el Tsunami de 2000?*
(158) *¿Qué es Mercosur?* / *¿Qué tipo de organización es Mercosur?*

Con la preposición *a*, puede funcionar como complemento directo de persona (159) y como complemento indirecto (160); con otras preposiciones puede funcionar como adjunto (161) o suplemento (162).

(159) *¿A qué invitados trajo Pablo Motos en la última edición de El Hormiguero?*
(160) *¿A qué jueces les mintió Mario Conde?*
(161) *¿Con qué se fabrica el chicle?* / *¿Con qué material se fabrica el chicle?*
(162) *¿A qué se encomendó?* / *¿A qué dios se encomendó?*

2.2.4.3.3 Restricciones en la selección de las palabras interrogativas

Como apunta Ignacio Bosque (el resaltado es nuestro)

> Si queremos hacer una pregunta acerca de una persona, elegiremos el pronombre *quién*; si se trata de una cosa usaremos *qué*; utilizaremos *cómo* si preguntamos acerca del modo o manera en que se realiza un acto, *dónde* si nos interesa el lugar, etc. [...] Por un lado, los conceptos de "modo", "lugar" y, especialmente, "cosa", no son categorías gramaticales y, por tanto, no están definidos objetivamente por la propia gramática; por otro, el reflejo gramatical de tales conceptos muestra un difícil cruce de factores sintácticos, semánticos e incluso pragmáticos de gran complejidad, ya que **la selección de las palabras interrogativas no depende únicamente de la**

> **clase léxica en la que, en principio, pueda inscribirse la categoría que ocupa el núcleo del sintagma interrogado**.
> (Bosque, 1984, p. 246).

De acuerdo con la afirmación anterior, no solo son factores semánticos como los que presentaremos en la sección 3.2.1 los que determinan la selección de los interrogativos en las preguntas, sino que también entran en juego en esta selección factores sintácticos e incluso pragmáticos. En los apartados que siguen se presentarán algunos de estos factores[63].

Construcciones copulativas: ecuativas *vs.* identificativas en la selección de los interrogativos

En las estructuras copulativas, la selección de ciertos interrogativos depende del carácter atributivo o ecuativo de la oración.

En las construcciones atributivas (163a) se atribuye una propiedad a un sujeto, mientras que en las ecuativas o identificativas (163b) se identifican los referentes de los dos sintagmas que une la cópula (Bosque, 1984)[64].

(163a) *Juan es médico*[65].
(163b) *Juan es el médico.*

A continuación analizaremos los interrogativos cuya selección está determinada por el carácter atributivo o ecuativo de las copulativas.

Qué vs. Quién

Como señala Lyons (1977) *quién* ocupa el lugar del predicativo en las identificativas, pero no puede hacerlo en las atributivas (163).

(164) *¿Quién es Juan? > Juan es el médico / * médico*

Qué, por su parte, ocupa el lugar de un SN en las oraciones atributivas (Bosque, 1984, p. 249). Siguiendo con los ejemplos anteriores, observamos en (165) que *qué* puede utilizarse tanto con (156) como con (157):

(165) *¿Qué es Juan? > Juan es el médico / médico.*

Esto ocurre porque algunas ecuativas como (165) pueden interpretarse también como atributivas («Juan tiene la profesión de médico»). Lo mismo ocurre con (166), que tiene doble interpretación ya que admite los dos interrogativos:

63 Toda la información de esta sección procede de Bosque (1984).
64 Para más información sobre la distinción entre copulativas ecuativas y atributivas cf. (Lyons, 1977, p. 471 y ss.).
65 Todos los ejemplos de esta sección están tomados de (Bosque, 1984).

(166) *Pedro es su novio* > *¿Quién es Pedro?* / *¿Qué es Pedro?*
Sin embargo, (167) solo admite la lectura ecuativa y no acepta *qué:*

(167) *Su novio es Pedro* > *¿Quién es su novio?* / **¿Qué es su novio?*

***Qué vs. Cuál*: Funciones identificativa y anafórica de los pronombres interrogativos**

Cuál puede sustituir a los predicativos en las copulativas identificativas pero no en las atributivas (168a), mientras que *qué* actúa al revés: funciona en las atributivas pero no en las ecuativas (168b):

(168a) Identificativa: *La mejor novela española es El Quijote* > *¿Cuál es la mejor novela española?* / **¿Qué es la mejor novela española?*

(168b) Atributiva - *El Quijote es la mejor novela española* > **¿Cuál es El Quijote?* (en un contexto no anafórico) / *¿Qué es el Quijote?*

Se dan, no obstante, casos ambiguos (169) cuando la copulativa puede ser tanto atributiva (169a) como identificativa (169b):

(169) *Su vida es su trabajo.*
(169a) Atributiva > *¿Qué es su vida?*
(169b) Ecuativa > *¿Cuál es su vida?*

Por otra parte, frente a *qué*, *cuál* no puede usarse en sentido neutro: no puede sustituir a infinitivos, oraciones enteras ni a adjetivos neutros sustantivados (Bosque, 1984):

(170) *¿Qué es el/lo mejor?* / *¿Cuál es el mejor?* / **¿Cuál es lo mejor?*

Quién vs. Cuál

Los dos interrogativos se refieren a personas, pueden funcionar con valor anafórico y aparecer en oraciones copulativas identificativas:

(171) *¿Quién es Juan? / ¿Cuál es Juan?*

La diferencia entre ambos es que *quién* también puede apuntar a un elemento no presente en el co(n)texto (no es anafórico) mientras que *cuál* siempre es anafórico. En (172)

(172) *Estaban presentes Juan y María, pero no sé (quién/cuál) abrió la puerta.*

quién puede apuntar a Juan, María u otra persona, mientras que *cuál* solo a Juan o María. Por esta razón, *quién* puede funcionar en construcciones ecuativas identificativas sin que el referente se haya identificado con anterioridad, frente a *cuál:*

(173) *Pedro es el/un médico.*
(173a) *¿Quién es Pedro? Un/El médico.*
(173b) *¿Cuál es Pedro? *Un/El médico.*

Qué vs. Cómo

Al igual que *qué*, *cómo* puede ocupar el lugar del SN en las oraciones atributivas, siempre y cuando se cumpla una de las siguientes condiciones:

- En el SN correspondiente a *cómo*, el núcleo del SN debe pertenecer a un amplio paradigma de sustantivos valorativos, siempre usados con artículo, en el que figuran *desastre, éxito, horror, encanto, locura*, etc., o a un segundo paradigma de nombres, constituido por un grupo de adjetivos sustantivados (*un tonto, un inútil, un genio*) que suelen incorporar connotaciones valorativas extremas (174a).

- En el SN, el sustantivo debe ir calificado por un adjetivo valorativo (por lo tanto, quedan excluidos los de relación) que especifique alguna propiedad o cualidad del sujeto, o bien por un complemento restrictivo que cumpla una función similar (174b).

(174) *¿Cómo es Pedro?*
(174a) *Pedro es un encanto.*
(174b) *Pedro es despistado / *Pedro es un médico.*

Como se deduce del punto anterior, *cómo* se construye con adjetivos y participios que funcionan como atributos en oraciones copulativas (175), mientras que *qué* solo admite adjetivos de relación y sintagmas nominales en esas mismas construcciones (176).

(175) *¿Cómo es Jesusa? Fea y pesada.*
(176) *¿Qué es Jesusa? Vallisoletana / Profesora.*

Esto último también ocurre con los complementos predicativos: el uso del interrogativo depende de la naturaleza nominal (*qué*) o adjetival (*cómo*) del complemento. Por esta razón, los verbos que admiten sintagmas nominales y adjetivales en su complemento predicativo aceptan la alternancia *qué*/*cómo* (177a, 177b).

(177a) *¿Qué se considera? > Se considera el rey del mambo.*
(177b) *¿Cómo se considera? > Se considera guapísimo.*

A su vez, los infinitivos en función de complemento predicativo rechazan los pronombres o adverbios interrogativos:

(178) *Oyeron a Juan salir despacio > *¿Qué /Cómo oyeron a Juan?*

***Quién vs. qué* y *cuál:* objetos directos no preposicionales**

Quién no puede sustituir a objetos directos no preposicionales, frente a *qué* (179) y *cuál* (180).

(179) *Juan busca secretaria > ¿Qué busca Juan?/ *¿Quién busca Juan?*
(180) *Prefiero el alto > ¿Cuál prefieres? / *¿Quién prefieres?*

2.2.4.3.4 Interrogativas parciales múltiples

En las interrogativas parciales es posible el uso de más de una partícula interrogativa (181).

(181) ¿Cuándo *y* dónde *se produjo la batalla de Salamina?*

A estas interrogativas con varias incógnitas se las denomina interrogativas múltiples.

En SpQA no consideramos las interrogativas múltiples, por lo que no se abordará su descripción teórica. Para el estudio de las interrogativas múltiples cf., entre otros: Contreras (1999, p. 1942 y ss.), Dumitrescu (1992) o Real Academia Española (2009, 22.2g y ss.; 42.9ñ y ss.).

2.2.4.3.5 Interrogativas parciales con disyunción

Otro tipo especial de interrogativas parciales lo constituyen las interrogativas parciales con disyunción:

(182) *¿Quién escribió Rayuela: Borges o Cortázar?*

En las parciales con disyunción, el ámbito de la pregunta que se define a partir de la partícula interrogativa se restringe mediante la estructura disyuntiva. De este modo, en (182) el valor de *quién* se restringe a Borges o Cortázar.

Las parciales con disyunción constituyen por tanto un tipo particular de interrogativa parcial en el que el valor de la incógnita está más determinado y se restringe a uno o varios elementos presentes en la pregunta.

2.2.4.4 Otros rasgos gramaticales relevantes para la interrogación

Hasta ahora nos hemos ocupado de aquellos rasgos gramaticales que definen nuestras preguntas: curva entonativa, orden de palabras y todo lo relativo a las partículas interrogativas. En los siguientes apartados trataremos otros aspectos

gramaticales que, si bien no son característicos de las interrogativas, sí nos parecen relevantes para entender su funcionamiento y significado.

2.2.4.4.1 Interrogativas y subordinación: asociación a distancia

En el interior de una interrogativa podemos encontrarnos una (183), (184) o más (185) oraciones subordinadas.

(183) *¿Cuándo pensaba Nostradamus* que se acabará el mundo*?*
(184) *¿Qué edad tenía Leonardo da Vinci* cuando murió*?*
(185) *¿Cuándo se piensa* que se terminará el mundo que todos conocemos*?*[66]

Según NGRALE, lo más frecuente en estos casos es que el verbo de la cláusula principal sea un «verbo puente»[67], «que se caracteriza por formar construcciones parentéticas o incisos: *creer*, *parecer*, *suponer*, *pensar*, *decir*, etc.» (Real Academia Española, 2009, 22.17d).

En este tipo de casos, puede generarse ambigüedad cuando el interrogativo puede asociarse con dos o más predicados (Contreras, 1999):

(186) *¿Cuándo dijiste que llegaba Pepe? > cuándo-dijiste / cuándo-llegaba*
(187) *¿Quién cree que irá? > quién-cree / quién-irá*

Lógicamente, cuantas más subordinadas tengamos (más predicados), más posibilidades de asociación hay:

(188) *¿Cuándo dijiste que Pedro anunció que llegaba María? > cuándo-dijiste/ cuándo-anunció / cuándo-llegaba*

(Ejemplo tomado de Contreras, 1999)

Las posibilidades de que se genere ambigüedad en aquellos casos en los que el interrogativo es semánticamente compatible con dos o más predicados varían dependiendo del tipo y función del interrogativo, ya que no todas las funciones sintácticas pueden asociarse con igual facilidad a dos o más predicados. A continuación analizaremos los distintos casos, partiendo siempre de una situación en la que el interrogativo puede asociarse semánticamente a dos o más predicados de la pregunta.

66 Preguntas como (183), (184) y (185) constituyen un ejemplo de lo que en BR se denomina «preguntas complejas» (cf. p. 38), preguntas que son susceptibles de ser descompuestas en varias preguntas más simples.

67 Generalmente verbos de lengua o pensamiento, algunos de percepción. Para más información sobre el concepto de «verbo puente» cf. (Real Academia Española, 2009, 25.8f).

Interrogativos funcionando como adjuntos

En el caso de *dónde*, *cuándo* y *por qué*, la ambigüedad se genera siempre, ya que no hay ningún factor gramatical que nos indique a qué predicado se asocia el interrogativo:

(189) *¿Cuándo dijo que llegaba Pepe? > cuándo-dijo / cuándo-llegaba*
(190) *¿Dónde esperaba que llegara Pepe? > dónde-esperaba / dónde-llegara*
(191) *¿Por qué dijo que lo había hecho? > por qué-dijo / por qué- lo había hecho*

Con *cómo* ocurre lo mismo, pero parece más difícil encontrar ejemplos en los que el interrogativo sea semánticamente compatible con dos o más predicados (192).

(192) *¿Cómo piensas que vive? > cómo-piensas; cómo-vive*

En (192), si *cómo* se asocia a *piensas*, la interpretación natural parece ser 'cómo se te ocurre pensar que vive', es decir, que la oración tendría más bien valor exclamativo y no interrogativo.

Interrogativos funcionando como argumentos

En el caso del sujeto, siempre y cuando se dé concordancia entre el sujeto y los predicados, se produce ambigüedad:

(193) *¿Quién dijo que venía? > quién-dijo / quién-venía*
(194) *¿Cuál dijo que venía? > cuál-dijo / cuál-venía*
(195) *¿Cuántos dijeron que venían? > cuántos-dijeron / cuántos-venían*
(196) *¿Qué hizo que cayera? > qué-hizo / qué-cayera*

Con el complemento directo generalmente no se produce ambigüedad, ya que en estos casos los dos verbos (principal y subordinado) han de ser transitivos y cuando es así la subordinada tiene que interpretarse como complemento directo del verbo principal, y el interrogativo como complemento directo del verbo subordinado:

(197) *¿Qué libro me recomendó tu hermana que leyera?*

Con el interrogativo funcionando como complemento indirecto no es posible saber con qué verbo se asocia el argumento interrogado (198), ni siquiera cuando hay clíticos (199), (200):

(198) *¿A quién dijo que sonriera? > a quién-dijo / a quién-sonriera*
(199) *¿A quién le dijo que sonriera? > a quién-le dijo / a quién-sonriera*
(200) *¿A quién dijo que le sonriera? > a quién-dijo / a quién-le sonriera*

Con el atributo y el suplemento interrogados no se produce este tipo de ambigüedad.

Aspectos que restringen la ambigüedad

Existen ciertos parámetros que restringen la posible ambigüedad en estas estructuras. Ya hemos visto uno cuando el interrogativo funciona como sujeto: concordancia sujeto/verbo:

(201) *¿Quién cree que irás a la fiesta?*

Tampoco se produce ambigüedad cuando la subordinada es otra interrogativa parcial:

(202) *¿Cuándo sabrás* quién llegó*?*

(Ejemplo tomado de Contreras, 1999)

En estos casos, la interrogativa subordinada funciona como una «isla sintáctica» (Contreras, 1999, p. 1945), de manera que cada uno de los interrogativos se asocia a su verbo. Esto ocurre siempre excepto cuando se dan las siguientes condiciones simultáneamente (Contreras, 1999, p. 1946):

- la subordinada es de infinitivo;
- el elemento interrogativo inicial es referencial.

En este supuesto, los dos interrogativos se asocian al verbo subordinado:

(203) *¿Qué coche no sabes cómo reparar? > cómo + qué coche-reparar*

(Ejemplo tomado de Contreras, 1999)

Respecto a la pertinencia de la segunda condición, compárese (203) con (204):

(204) *¿Cómo no sabes qué coche reparar? > cómo-saber; qué coche-reparar*

(Ejemplo tomado de Contreras, 1999)

Al parecer, la diferencia entre expresiones referenciales como (203) (*qué coche*) y expresiones no referenciales como (204) (*cómo*) respecto a su interpretabilidad a distancia parece no ser exclusiva del español (Rizzi, 1990), y lo mismo sucede con las cláusulas de infinitivo *vs.* cláusulas finitas (Manzini, 1992).

2.2.4.4.2 Verbo no finito

Las interrogativas parciales pueden contener un verbo no finito:
(207) *¿Cómo encontrar a Kyogre en Pokemon Oro Corazón?*
(208) *¿Dónde encontrar a Espiritomb?*

2.2.5 Conclusiones

En los apartados anteriores hemos definido nuestro objeto de estudio. Para ello, en primer lugar hemos definido qué entendemos como pregunta, a saber: una estructura cuya forma gramatical se corresponde con una oración interrogativa directa y cuya finalidad es la demanda de información[68].

A continuación hemos caracterizado gramaticalmente las interrogativas directas, presentando aquellos rasgos que las caracterizan: curva entonativa, orden de palabras y presencia de partículas interrogativas, junto con otros rasgos gramaticales que son relevantes en su funcionamiento y significado (negación, asociación a distancia, etc.).

En las páginas que siguen, nos ocuparemos del análisis semántico pragmático de las preguntas. En primer lugar presentaremos las distintas aproximaciones teóricas más relevantes a la semántica de las preguntas; a continuación nos ocuparemos de aspectos semánticos y pragmáticos concretos que pueden ser interesantes para la formalización en la gramática de SpQA.

2.3 La semántica y la pragmática de las preguntas

La comprensión del significado subyacente a una pregunta es un aspecto prioritario de los sistemas de BR. Cuanto mayor sea la capacidad del sistema para entender, con la máxima concreción, cuál es la demanda de información que el usuario plantea con su pregunta, más posibilidades tendrá de encontrar una respuesta correcta a esa pregunta.

Como veremos en las siguientes secciones, en la interpretación del significado de una pregunta se entrecruzan factores semánticos y pragmáticos de un modo complejo, de manera que, en ocasiones, es imposible discernir la frontera entre los dos ámbitos.

En las páginas que siguen nos ocuparemos del estudio semántico y pragmático de las preguntas. En primer lugar presentaremos una revisión general de las distintas aproximaciones teóricas a la semántica de las preguntas, introduciendo nociones pragmáticas en aquellas teorías en las que semántica y pragmática son indisolubles. A continuación trataremos algunos aspectos pragmáticos relevantes para nuestro análisis. Finalmente, nos ocuparemos en mayor detalle de un aspecto concreto de la semántica de las preguntas que es clave para nuestro análisis: la semántica de las partículas interrogativas.

68 En nuestro caso, en un entorno de BR.

2.3.1 Teorías semánticas sobre las preguntas: la relación pregunta-respuesta

Los estudios centrados en la semántica de las preguntas iniciaron su andadura en el área de la lógica y la computación, a través de aproximaciones que trataban de computar mediante ciertos formalismos el significado de las preguntas. Desde el principio la relación pregunta-respuesta ocupó un lugar clave en estas teorías. Como veremos, en el área se produjo una paulatina evolución desde las primeras aproximaciones que partían de una semántica puramente formal, asentada en principios lógicos y que excluía aspectos pragmáticos, hacia las perspectivas más actuales en las que la semántica de las preguntas parece no poder entenderse sin la pragmática.

En las páginas siguientes llevaremos a cabo un repaso de las distintas teorías semánticas y semántico pragmáticas que se han ocupado del estudio de las preguntas. Dividiremos nuestro repaso en dos bloques: en primer lugar, presentaremos las teorías puramente semánticas (semántica formal) y a continuación aquellas de carácter semántico pragmático.

2.3.1.1 Semántica formal

2.3.1.1.1 Hamblin: la relación pregunta-respuesta

El estudio de la semántica de las preguntas se remonta (al menos) hasta Hamblin (1958). En esa época, el planteamiento teórico general respecto a las preguntas sostenía que la única diferencia entre interrogativas y declarativas era pragmática y no semántica. En su trabajo de 1958, Hamblin presenta una idea diferente: para él **la semántica de las preguntas debe ser definida a través de sus respuestas**. Desde entonces, esta interpretación del significado de las preguntas será la predominante en la mayoría de los estudios posteriores.

En su artículo de 1958, Hamblin enuncia tres postulados sobre la semántica de las preguntas a través de los cuales define la relación entre pregunta y respuesta. Estos postulados también estarán presentes, en mayor o menor grado, en todas las aproximaciones posteriores.

Postulados de Hamblin

> *Postulate 1. An answer to a question is a statement.*
> (Hamblin, 1958, pp. 162-163).

El primer postulado de Hamblin hace referencia a la **forma de las respuestas**.

Aquí *statement* significa «oración declarativa». Hamblin asume que las respuestas a preguntas son siempre oraciones completas y que las respuestas cortas son formas elípticas de esas oraciones. Veamos un ejemplo:

(209a) *¿En qué continente está Luxemburgo?*
(209b) *¿En qué continente está España?*

(210a) *Europa*.
(210b) *Luxemburgo está en Europa.*

(Ejemplo adaptado de Kaufmann, 2009)

En el ejemplo anterior, si utilizamos la respuesta-constituyente (210a) para definir (209a) y (209b), estas preguntas serían semánticamente equivalentes. Sin embargo, si tomamos la respuesta-oración (210b), esta solo responde a (209a) y, por lo tanto, (209a) y (209b) no son equivalentes. La consideración de respuesta-oración *vs.* respuesta-constituyente tiene por tanto una serie de implicaciones lingüísticas que deben ser tenidas en cuenta y que trataremos en detalle más adelante (cf. sección 3.1.1.3).

> *Postulate 2. Knowing what counts as an answer is equivalent to knowing the question.*
> (Hamblin, 1958, pp.162-163).

El segundo postulado se refiere al significado de las preguntas. Lo que Hamblin nos dice es que determinar el significado de las preguntas consiste en saber qué cuenta como respuesta. Debemos tener en cuenta que «saber qué cuenta como respuesta» no es lo mismo que «saber cuál es la respuesta»: **lo que el principio de Hamblin sugiere es que el significado de la pregunta restringe el ámbito de las respuestas**.

> *Postulate 3. The possible answers to a question are an exhaustive set of mutually exclusive possibilities.*
> (Hamblin, 1958, pp. 162-163).

Aquí *possibility* significa «proposición», o lo que es lo mismo: un conjunto de mundos posibles. Por lo tanto, una pregunta denota un conjunto de mundos posibles. *Exhaustive* significa que el conjunto de posibles respuestas agota el espacio lógico de posibilidades para la respuesta. No debe confundirse con la exhaustividad de las respuestas (determinar hasta qué punto la respuesta cubre la necesidad de información del que pregunta). A su vez, *mutually exclusive possibilities* significa que los posibles mundos de cada una de las posibles respuestas son mutuamente excluyentes.

En relación a la exhaustividad de las respuestas, Hamblin hace además una distinción entre *proper answers* y *no proper answers,* siendo las *proper answers* respuestas completas, es decir, mutuamente excluyentes:

(211) *¿Dónde está Luxemburgo?*
(211a) *O en Europa, o en Asia, o en África.*
(211b) *Luxemburgo está en Europa.*

(Ejemplo adaptado de Kaufmann, 2009)

En el ejemplo, (211a) sería una *no proper answer*, mientras que (211b) sería una *proper answer.*

Como ya apuntábamos, las aproximaciones teóricas posteriores a la semántica de las preguntas harán referencia o uso de uno o varios de estos postulados (solo el *Partition Approach,* cf. 3.1.2.1, hará uso de los tres).

2.3.1.1.2 Belnap y Stell: *Erotetic Logic*

La aproximación de Belnap y Steel (1976) se centra en el desarrollo de un formalismo adecuado para describir preguntas y respuestas. Para ello, se utiliza un lenguaje formal que es una extensión (llamada L) de la lógica de predicados de primer orden con igualdad[69] (FOPL=). La intención de la teoría no era tanto conseguir una representación formal del lenguaje natural como conseguir una representación formal para bases de datos y sistemas similares. Por lo tanto, Belnap no estaba interesado en una representación adecuada de fenómenos lingüísticos, aunque a menudo se valiese de ejemplos de lenguaje natural para la presentación de sus ideas (Belnap y Steel, 1976, pp. 139-148).

Belnap y Steel tratan solo las interrogativas directas parciales. En su representación, las preguntas constan de dos partes: *subject* (σ) y *request* (ρ), siendo ambas partes *tuplas*[70]. La parte de *subject* representa la pregunta en sí misma (proposición y variables). Esta parte contiene un rango de alternativas que pueden estar explícitas o indicadas usando una referencia a una condición o matriz[71]. Dada σ, la parte de *request* representa las especificaciones relacionadas con las respuestas solicitadas, tales como el número de respuestas solicitadas.

El sistema da cabida a preguntas con *who*, *when*, *where*, *why*, y *what*, siempre y cuando los interrogativos pregunten por «qué persona», «qué tiempo», «qué lugar», «qué razón» y «qué cosa», respectivamente.

Las preguntas pueden variar respecto al número de entidades que la denotación de la respuesta debe contener. Para los autores, hay modos en las lenguas naturales para indicar la *selection size* (Belnap), es decir, el tamaño de la selección de la pregunta, como por ejemplo el uso de cuantificadores:

(212) *Which* five *theories predict this fact correctly?*

69 http://www.philosophy-index.com/logic/systems/first-order.php
70 Una tupla, en matemáticas, es una secuencia ordenada de objetos, esto es, una lista con un número limitado de objetos. Cf., por ejemplo: http://es.wikipedia.org/wiki/Tupla
71 Una condición o matriz es una afirmación que contiene variables, y que se puede corresponder con un set finito o no finito.

Sin embargo, muchas preguntas no son específicas con respecto a este punto.

En relación al tamaño de la selección, normalmente se distinguen tres casos posibles:

- *Mention-one-interpretation*: la pregunta pide un solo ejemplo verdadero como respuesta.
- *Mention-all-interpretation*: la pregunta solicita una respuesta exhaustiva.
- *Mention-some-interpretation*: la pregunta demanda algunos (n) ejemplos de respuestas verdaderas.

A este respecto, posteriormente se ha discutido si las interpretaciones respecto al tamaño de la selección de la pregunta implican diferencias semánticas (deben recibir distintas representaciones semánticas) o si la pragmática debe dar cuenta de ellas; Groenendijk y Stokhof (1997), por ejemplo, se decantan por la distinción en el nivel semántico.

2.3.1.1.3 *Categorial Approach*

Los planteamientos categoriales se construyen en torno a la idea central de que un cierto número de fenómenos relacionados con las preguntas y las respuestas pueden ser explicados representando las preguntas como términos lambda (abstractos) cuyos tipos reflejan aquellos de la pregunta y la respuesta. De esta manera, la representación de la pregunta respondida puede derivarse por medio de una aplicación funcional desde la representación de la pregunta a la representación de la respuesta. Este planteamiento incorpora por tanto el segundo postulado de Hamblin.

Interrogativas parciales

El planteamiento categorial se basa en la idea de que una interrogativa parcial puede ser respondida por un **constituyente sintáctico «del tipo correcto»**. Ese «tipo correcto» se establece a partir de la frase interrogativa de la pregunta. Por lo tanto, los trabajos del enfoque categorial (Hausser, y Zaefferer, 1979; Keenan, y Hull, 1973; Krifka, 2001, entre otros) establecen una relación directa entre la categoría semántica de la pregunta (de la frase interrogativa) y una determinada categoría sintáctica (para la pregunta y la respuesta).

Las preguntas se representan como términos lambda donde el término lambda se corresponde directamente con el constituyente correspondiente de la posible respuesta-constituyente. Una pregunta es por lo tanto un *funtor*[72] que toma la respuesta como su argumento. La aplicación de la función produce directamente la proposición que representa la pregunta respondida. Hausser considera la respuesta-constituyente como primaria, y diseña su semántica de manera que la pregunta denota una función, la correspondiente respuesta-constituyente denota un posible argumento para esa función y si y solo si la respuesta es verdadera, la aplicación del

72 http://en.wikipedia.org/wiki/Functor

significado de la pregunta al significado de la respuesta resulta en una proposición verdadera. Esta proposición, a su vez, se corresponderá con el significado de la respuesta-oración.

(213)[73]

(a) *Who did Mary see?*
(b) <e,t>: λx[see' (x) (mary')]
(c) *John*.
(d) e:john'
(e) t: λx[see' (x) (mary')] (john') = see' (john') (mary')

(Ejemplo adaptado de Krifka, 2001, p. 288)

En el ejemplo (b) es la representación de (a). La respuesta (c) se representa mediante (d). Y la aplicación de la función produce la proposición que es transmitida a través de la respuesta (e). Así que en el análisis de Hausser, el interrogativo actúa como un operador lambda sobre la variable correspondiente.

Por otra parte, en este análisis, las preguntas parciales con un solo interrogativo como (213) son de tipo <e,t>. Este tipo es, a su vez, idéntico al de las cláusulas de relativo. Sin embargo, una pregunta con dos interrogativos es de distinto tipo, porque cada uno de los interrogativos es abstraído en términos lambda. De este modo, en:

(214) Who *bought* what*?*

el tipo será <e, <e,t>>. En general, una pregunta con n interrogativos denotará una relación n-lugar. Una pregunta total, con 0 interrogativos, denota una relación 0-lugar, lo cual es justamente una proposición.

Interrogativas totales

La idea general es que se manejan a través de las denotaciones *sí* y *no* (λ p.p y λ p.¬p, respectivamente), de manera que las aplicaciones funcionales de la representación de las preguntas totales a estas denotaciones (las posibles respuestas para las totales) resultan en la versión «afirmativa» o «negativa» de la representación de la pregunta.

Ventajas y desventajas del análisis

La principal ventaja de este análisis es que la relación pregunta-respuesta es modelada de una forma muy directa. El paso de la forma de la pregunta a su interpretación es muy sencillo. Por otra parte, también es muy interesante el modo en el que se capta la relación entre las interrogativas parciales y las oraciones relativas.

En cuanto a las desventajas, la mayor de todas es el hecho de que los distintos

73 El ejemplo está simplificado de manera que se usa una representación extensional y se asume la representación de nombres propios como constantes individuales (tipo e) (Fliedner, 2007).

tipos de preguntas se representan de distintas formas. Esto implica que, por ejemplo, sea muy difícil dar cuenta de la combinación de preguntas de distinto tipo.

2.3.1.1.3.1 Respuestas-oración *vs.* respuestas-constituyente

El planteamiento categorial defiende que las interrogativas parciales pueden ser respondidas por respuestas tipo constituyente, planteamiento que contrasta con el primer postulado de Hamblin (cf. *supra*, sección 3.1.1.1), al menos en su forma más pura. Como hemos visto, Hamblin sugiere que si se usa un constituyente, este debe recibir igualmente la interpretación de un enunciado completo en el contexto de una pregunta (Hamblin, 1958, p. 162).

Ingo Reich argumenta (Reich, 2003) que se debe asumir que una «respuesta canónica» a una parcial es siempre una oración completa y que una respuesta tipo constituyente se deriva de la elisión de material superfluo. A continuación presentaremos los argumentos de Reich para asumir que una respuesta tipo constituyente es de hecho una versión elidida de una respuesta-oración y luego presentaremos la regla de Reich que permite las elisiones correctas.

Modificación oracional

Reich observa (Reich, 2003, p. 25) que las respuestas tipo constituyente pueden ser modificadas por adverbios que son modificadores oracionales:

(215a) *Who will win the election?*
(215b) Probably *Bush.*

(Ejemplo tomado de Fliedner, 2007, p. 58)

Este hecho no puede ser explicado fácilmente cuando se asume una respuesta tipo constituyente, ya que este no sería de la categoría requerida (oración) que demanda un adverbio que es modificador oracional (*probably*).

Marca de caso

La respuesta-constituyente debe tener el mismo caso que el interrogativo de la pregunta. Por supuesto, esto es más evidente en lenguas con marca de caso abierto. Consideremos el siguiente ejemplo del alemán:

(216)
a. [*Welchen Studenten*]*dat hast du eine Eins gegeben?*
(To) which students have you one A given?
b. *To which students have you given an A?*
c. [*Allen Studenten*]dat .
(To) all students.

d. * [*Alle Studenten*]nom/acc .
All students.
e. *To all students*.
f. ? *All students*.

(Ejemplo tomado de Fliedner, 2007, p. 58)

En alemán, (d), donde el constituyente tiene el caso equivocado, es una respuesta claramente agramatical. En la traducción al inglés, juzgaríamos también que hay preferencia por (e) sobre (f) como respuesta a (b).

Reich apunta (Reich, 2003, p. 23) que la agramaticalidad de ejemplos como (216d) puede explicarse más fácilmente cuando uno asume material elidido en la respuesta tipo constituyente. Se asume generalmente que la asignación de caso se hace a través de reglas sintácticas determinadas por la palabra que rige la oración (Fliedner, 2007). Si se plantea una respuesta tipo constituyente, no está muy claro qué palabra debería regir este constituyente.

Pronombres reflexivos/recíprocos

Este punto está estrechamente ligado al precedente. Reich observa (Reich, 2003, p. 23) que una respuesta tipo constituyente puede consistir en un único pronombre recíproco o reflexivo.

En las teorías de corte generativista se asume generalmente que este tipo de pronombres se permiten solo cuando son *c-commanded* (exigidos) por su antecendente (Fliedner, 2007, p. 59). En este supuesto, es difícil ver cómo puede asumirse una relación *c-command* para un solo constituyente:

(217) *Who do John and Mary love?*
(217a) *Each other.*

(Ejemplo tomado de Fliedner, 2007)

Conclusión

Teniendo en cuenta estos tres argumentos, Reich llega a la conclusión de que las respuestas a las preguntas parciales son siempre respuestas oracionales. Para explicar las respuestas tipo constituyente, Reich sugiere que estas derivan de la forma de oración plena por elisión. Observa que en una respuesta adecuada para una parcial todos los constituyentes que corresponden a la frase interrogativa en la pregunta deben ser focalizados (y por lo tanto generalmente llevan el acento de foco). La respuesta oracional correspondiente a (217a) puede ser interpretada como sigue[74]:

(217b) *John and Mary love [each other]*F.

(Ejemplo tomado de Fliedner, 2007)

74 F marca el constituyente focalizado.

Reich propone una regla que plantea que en una respuesta oracional subyacente a la correspondiente parcial, todo el material no focalizado (y por lo tanto, especialmente todo el material que no corresponde a la frase interrogativa) puede ser elidido. En el ejemplo, la elisión del material no focalizado proveniente de (217b) lleva exactamente a la respuesta tipo constituyente esperada en (217a) (Reich, 2003, p. 26). A una conclusión similar llega Drubig (2003):

> Since there are languages in which term answers have a distinct syntactic form, I will assume that WH-questions induce focussed term answers and that sentential replies are redundant elaborations of term answers.
> (Drubig, 2003, p. 3)

2.3.1.1.4 *Structured Meaning approach:* foco y *background*

La aproximación de Reich (2003) integra la idea, tomada del *Propositional Approach* (cf. *infra*), de que las preguntas denotan conjuntos de proposiciones. Reich considera que las denotaciones son «conjuntos de proposiciones estructuradas», dividiendo las proposiciones en *focus* y *background*.

El *structured meaning approach* da cuenta de diversos fenómenos en los que el foco interactúa con la interpretación de preguntas y respuestas, fenómenos observables principalmente mediante patrones acentuales en las lenguas habladas. Como ya hemos visto, en el lenguaje natural el foco es principalmente expresado por métodos como el acento (los constituyentes focalizados reciben el acento) o por cambios de órdenes de palabras (cf. sección 2.3). En consecuencia, los hallazgos de esta teoría serían centrales para sistemas dialogados con lenguaje oral, donde podrían ayudar a lograr una entonación natural o a distinguir entre distintos tipos de preguntas[75].

En el *structured meaning approach* (Stechow, 1991), las representaciones semánticas de preguntas y respuestas consisten en dos partes: la parte de *background* y la parte de *focus*. El siguiente criterio define la relación de «respuesta»[76]:

> *Criterion for congruent question-answer pair Q – A, where [[Q]] = <B,R> and [[A]] = <B',F>:B' y F ∈ R.*
> (Tomado de Krifka, 2001, p. 296)

Es decir: en un par pregunta-respuesta congruente, el *background* de pregunta y respuesta debe ser el mismo, y el foco en la respuesta es un elemento restringido por la pregunta. Veamos un ejemplo[77]:

(218)
(a) *Who did Mary see?*
(b) <λx[see'(x)(m')],person'>

75 En un modelo de BR operando sobre texto como el que nosotros planteamos para SpQA estas cuestiones son menos relevantes.

76 B = *background*, F = foco y R = restricción de la pregunta.

77 Se sigue la práctica general y se marcan los límites de los constituyentes focalizados con F; el acento se marca con ´.

(c) *Mary saw [Jóhn]*F.
(d) <λx[see'(x)(m')],j'>, where j' ∈ person'
(e) **[Máry]*F *saw John*.
(f) <λx[see'(j')(x)],m'>, where m' ∈ person'

(Ejemplo tomado de Fliedner, 2007)

Dado el criterio de «respuesta» de Krifka, este ejemplo muestra cómo la propuesta del *structured meaning approach* puede explicar la observación de que (c) es una buena respuesta para (a), mientras que (e), donde Mary está acentuado, no. Las representaciones (b) y (d) forman un par congruente pregunta-respuesta bajo el criterio de respuesta (idéntico *background*, el foco es elemento de restricción), mientras que (b) y (f) no (el *background* difiere).

2.3.1.1.5 *Propositional Approach*

En este marco teórico, las preguntas se representan como el conjunto de proposiciones (mundos posibles) que constituyen sus respuestas.

Para Hamblin (1973), una pregunta es una función que va desde una situación a un conjunto de proposiciones que pueden ser posibles respuestas a esa pregunta. De hecho, Hamblin es el primero en tratar las preguntas como denotaciones de un set de proposiciones. En ese conjunto de proposiciones se incluyen todas las respuestas posibles, tanto las verdaderas como las falsas.

En su análisis, Hamblin solo se ocupa de las preguntas directas y no de las indirectas. Para él los interrogativos son denotaciones de conjuntos de individuos; por ejemplo, *who* denota un conjunto de humanos, *what* un conjunto de no-humanos. Hamblin introduce la noción de *denotation set* como el conjunto de denotaciones de cualquier expresión. De este modo cada expresión (pregunta, posibles respuestas) se representa (o se asocia) con un conjunto de denotaciones.

Por lo tanto, en este marco teórico, la relación de *answerhood* (el grado en el que una respuesta responde efectivamente una pregunta) se reduce a comprobar si la representación de la respuesta es un miembro del conjunto que denota la representación de la pregunta (ya que esta es exactamente un conjunto de proposiciones que son las posibles respuestas).

En Karttunnen (1977) las preguntas también son funciones desde situaciones a conjuntos de proposiciones. A diferencia de Hamblin, Karttunen trata tanto las interrogativas directas como las indirectas. Por otro lado, para Karttunen las proposiciones que denota una pregunta incluyen tan solo las respuestas verdaderas a dicha pregunta (Fliedner, 2007, p. 62).

Karttunen analiza una pregunta denotando, en cada mundo posible (o situación posible)

> the set of propositions which in that situation jointly constitute a complete and true answer to the question. The denotation of whether John walks in a given situation, is a set whose only member is either the proposition that John walks or the proposition that John doesn't walk, depending on which of these happens to be the

> true one. The denotation of who walks is the set of true propositions expressed by sentences of the form "x walks".
> (Karttunen, 1977)

Podríamos representar el análisis de Karttunen para cada tipo de interrogativo de la siguiente manera:

> **TOTALES** [[*whether John left*]] = {*John left, John didn't leave*}
> **PARCIALES** [[*who solved the problem*]] = {*John solved the problem, Mary solved the problem, ...}*
> **DISYUNTIVAS** [[*whether Ede wants coffee or tea*]] = {*Ede wants coffee, Ede wants tea*}

(Kaufmann, 2009)

En la aproximación de Karttunen se considera además que todas las preguntas derivan de un enunciado declarativo. Para ello se introducen las denominadas «proto-reglas», que hacen posible la transición desde la semántica de las declarativas a la semántica de las preguntas. Esta transición se realiza de la siguiente manera: Karttunen sugiere analizar las preguntas directas como si llevaran un verbo performativo encubierto, de forma que fueran equivalentes a una fórmula como «I ask you to tell me α» (una interrogativa indirecta), donde α es la correspondiente pregunta directa. De manera que una pregunta directa como:

(219) *What did John buy?*

es representada como si fuera

(220) *I ask you to tell me what John bought*[78].

En esta aproximación no se explica formalmente con detalle cómo se realiza el *matching* entre la representación de la pregunta y la representación de la respuesta (Fliedner, 2007). Hamblin lo describe como sigue:

> Semantically, an answer to a question on a given reading is any statement whose denotation-set on a suitable reading is contained in that of the question.
> (Hamblin, 1973, p. 52)

Hamblin también señala que puede haber condiciones sintácticas adicionales de correcta formulación en el par de pregunta-respuesta (*ibid*).

78 A este respecto, Potts ha destacado recientemente (Potts, 2006) que una buena teoría pragmática debe hacer innecesario presuponer tal estructura encubierta en la sintaxis o la semántica de las preguntas.

Críticas

La principal crítica al *Propositional Approach* es que este no explica las respuestas indirectas a preguntas (Fliedner, 2007, p. 63). Esta aproximación asigna representaciones uniformes para los diferentes tipos de preguntas, concretamente, conjuntos de proposiciones que cuentan como posibles respuestas. Incorpora por lo tanto tanto el primer como el segundo postulado de Hamblin. Sin embargo, no permite respuestas indirectas a preguntas, ya que solo las proposiciones que responden directamente la pregunta son predichas como respuestas posibles. Lo mismo se aplica para las diferentes aproximaciones categoriales (cf. sección 3.1.1.3): la *answerhood* se predice si la representación de la pregunta y la respuesta son iguales. Este vacío está cubierto en el *Partition Approach*, que describiremos en la siguiente sección.

2.3.1.2 Interrelación de la semántica y la pragmática en la interpretación de preguntas y respuestas

2.3.1.2.1 *Partition Approach*

Probablemente esta sea la aproximación reciente a la semántica de preguntas y respuestas más influyente. Se basa en la semántica de mundos posibles (Gochet, 2011). Como veremos a continuación, las particiones son sets de mundos posibles (índices) que se usan para representar el significado de las preguntas.

La definición semántica de preguntas y respuestas en esta aproximación se basa en la noción de **respuestas exhaustivas posibles**. Por ejemplo, una respuesta exhaustiva a la pregunta *Who sleeps?* apuntaría a cada individuo tanto si duerme como si no. Esta parece, sin embargo, una noción demasiado fuerte de *answerhood* como para representar las respuestas típicas del lenguaje natural.

Por esta razón, en el *Partition Approach* se utiliza una definición de respuesta parcial pragmáticamente definida. Esta definición permite capturar el hecho de que cualquier réplica que excluya al menos una posible respuesta contará como respuesta parcial. Esto a su vez permite explicar una serie de fenómenos conectados con preguntas y respuestas, entre ellos, aquellos relacionados con las respuestas indirectas que ni el *Categorial* ni el *Propositional Approach* pueden manejar.

La intuición es que las particiones se corresponden con el conjunto (completo) de todas las posibles respuestas exhaustivas. Veamos dos ejemplos, uno de una pregunta total (221) y otro de una pregunta parcial (222):

(221) *¿Está lloviendo?*
Particiones:
221.1 *Está lloviendo.*
221.2 *No está lloviendo.*

(222) *¿Quién está caminando?*
Particiones:
222.1 *Nadie está caminando.*

222.2 *Solo A está caminando.*
222.3 *Solo B está caminando.*
222.4 *Solo C está caminando.*
222.5 *Solo A y B están caminando.*
222.6. *Solo A y C están caminando.*
...
222.7 *Todos están caminando.*

(Adpatado de Groenendijk, y Stokhof, 1984)

En (221) tenemos las particiones correspondientes a la pregunta total *¿Está lloviendo?* Una partición (221.1) contiene los índices donde la extensión de la proposición de que está lloviendo es verdadera (es decir, está lloviendo en cada uno de estos mundos), el otro (221.2), aquellos donde la extensión es falsa (es decir, no está lloviendo). De esta manera, para las preguntas totales, hay exactamente dos particiones.

En (222) tenemos las posibles particiones correspondientes a la pregunta parcial *¿Quién está caminando?,* en un modelo con solo tres individuos: A, B y C. Aquí las particiones se distinguen determinando qué individuos contiene la extensión del predicado *caminar*: la primera partición (222.1), contiene todos los índices donde la extensión está vacía, de manera que *nadie está caminando*. A esta partición le siguen aquellas en las que exactamente un solo individuo camina (222.2, 222.3, 222.4), luego las posibles combinaciones de dos individuos (222.5 y 222.6) y finalmente los índices donde los tres individuos están caminando (222.7).

El significado de una pregunta es el conjunto de particiones, construidas como hemos esbozado en (221) y (222), para dividir todos los índices de acuerdo exactamente a todas las posibles respuestas exhaustivas. Todos los índices en los que la extensión de la proposición es equivalente constituyen exactamente una partición.

Una respuesta a una pregunta se evalúa de la siguiente manera: de la pregunta se deriva una representación semántica de la respuesta. La representación se usa después para seleccionar índices con los que esa extensión sea compatible. Por lo tanto, en general, un número de índices serán excluidos como no compatibles con esa respuesta concreta. Se dice que la respuesta constituye una respuesta exhaustiva cuando únicamente deja índices que forman parte de la misma partición, es decir, cuando únicamente se deja una partición. Como se mencionó más arriba, esta no parece ser una respuesta natural en la mayoría de los contextos. Típicamente, el que pregunta estará satisfecho con una respuesta parcial. De esta manera, las respuestas parciales a una pregunta pueden ser explicadas elegantemente por el *Partition Approach*. El requerimiento de que una respuesta deba excluir todas excepto una partición simplemente se cambia por el siguiente: **para ser una respuesta parcial, una respuesta debe excluir al menos una partición**. La intuición detrás de este planteamiento es la que sigue: excluyendo una partición, el que responde ha aportado información al que pregunta que es **pertinente** a la pregunta, de manera que el que pregunta puede ahora excluir una respuesta posible. Una respuesta como *A no está caminando* sería considerada una respuesta parcial a la pregunta, ya que excluye todas las particiones en las que A está andando.

La propuesta de Groenendijk y Stokhof asigna por tanto el siguiente significado a la pregunta en (222):

(223) λwλw'[λx[walk(w)(x)] = λx[walk(w')(x)]]

(Representación tomada de Fliedner, 2007)

La semántica de particiones tiene algo en común tanto con la semántica de las preguntas del *Proposition Approach* como con la del *Categorial Approach*. Puede verse una expresión lambda del *Categorial Approach* en la fórmula de (223): esta aporta el «tema» de la pregunta. Pero la fórmula completa de (223) es del tipo «relación entre mundos» (una relación de equivalencia, por tanto, una partición), y todas las preguntas pueden ser del mismo tipo. El *Partition Approach* también reconcilia la diferencia entre Karttunen y Hamblin, ya que hace uso en su semántica tanto del conjunto de todas las posibles respuestas completas como de las respuestas verdaderas.

Respuestas indirectas e inferencias

El *Partition Approach* puede explicar también las respuestas indirectas: como la *answerhood* se define a través de la incompatibilidad con conjuntos de mundos, las respuestas directas e indirectas pueden, de hecho, no distinguirse.

Consideremos una respuesta indirecta a (222) como *Llueve*. Si se sabe que A nunca camina cuando llueve, eso excluye una serie de particiones de (222) (concretamente, todas en las que A camina) y por lo tanto, da lugar a una respuesta parcial. Esta construcción sobre inferencia también permite explicar directamente que *B está conduciendo* contaría como una respuesta parcial indirecta (esto es, cuando tanto el que pregunta como el que responde están de acuerdo en el hecho de que alguien que está actualmente conduciendo posiblemente no puede estar al mismo tiempo caminando).

De este modo, en general se hace necesario modelar explícitamente el conocimiento y creencias del que pregunta y del que responde y especialmente su conocimiento compartido. En (Groenendijk y Stokhof, 1984), la descripción básica dada aquí es extendida para manejar explícitamente estas cuestiones modelando conjuntos doxásticos y epistémicos (para representar las creencias y el conocimiento, respectivamente).

Comparación de respuestas

Groenendijk y Stokhof proponen también una definición que permite comparar dos respuestas distintas para una misma pregunta (Groenendijk, y Stokhof, 1997, p. 1095). El núcleo de esta definición es el cálculo de lo informativo de una respuesta con respecto a la pregunta: si una respuesta excluye al menos una partición más que otra respuesta, la primera es considerada más informativa (ya que restringe más las posibilidades restantes).

Preguntas abiertas *vs.* preguntas informativas

Groenendijk y Stokhof también introducen una distinción importante entre lo que ellos llaman «preguntas informativas» y «preguntas abiertas» (Groenendijk, y Stokhof, 1997, p. 1108). Apuntan que el *Partition Approach* (como el resto de las aproximaciones discutidas aquí) no tiene cobertura para todas las preguntas en el lenguaje natural, sino que estas están limitadas únicamente al subtipo que ellos denominan «preguntas informativas».

Las preguntas informativas se caracterizan por el hecho de que todas las posibles respuestas directas pueden construirse directamente a partir de la pregunta. En el *Partition Approach* esto se hace, por ejemplo, partiendo exhaustivamente los índices dados basados en la extensión de la proposición contenida en la pregunta.

Por su parte, las preguntas abiertas se diferencian de las preguntas informativas en que sus posibles respuestas no pueden ser «obtenidas» de la preguntas. El ejemplo que utilizan Groenendijk y Stokhof es la pregunta *What are questions?* Los autores argumentan que no existe una respuesta simple y predifinida para esta pregunta. Una pregunta abierta parece necesitar un proceso creativo; por ejemplo, la respuesta puede darse a través de un ensayo completo más que a través de una simple proposición. Groenendijk y Stokhof también apuntan que la noción de verdad o falsedad de una respuesta a una pregunta abierta a menudo parece no tener sentido, y que otros conceptos (como «bueno» o «útil») parecen más intuitivos para describir este tipo de respuestas. Los autores llegan a la conclusión de que, por ahora, no existe una teoría útil para describir las preguntas abiertas.

Actualmente hay un creciente interés en BR por las preguntas abiertas (Burguer et al., 2001; Maybury, 2003). No existe acuerdo en torno a ningún esquema que resuelva cómo deben responderse este tipo de preguntas. Por otro lado, es difícil incluso identificar tales preguntas, ya que la distinción entre preguntas informativas y preguntas abiertas no es del todo clara. Ciertamente hay preguntas que parecen puramente informativas (*¿Cuándo fue John F. Kennedy asesinado?*), mientras que otras parecen puramente abiertas (*¿Qué son las preguntas?*). Pero hay otras preguntas que parecen situarse a medio camino, como, por ejemplo, *¿Por qué es famoso Cristóbal Colón?,* que puede responderse de modo informativo (*Porque descubrió América*) pero también escribiendo un ensayo completo como respuesta. Parece haber una cierta tendencia, sin embargo, a que las llamadas *definition-Q* (*¿Qué son las preguntas?*), *why-Q* (*¿Por qué es famoso Cristóbal Colón?*), *how-Q* (*¿Cómo se hace una quiche?*) y *what-if-Q* (*¿Qué ocurriría si Suíza le declarase la guerra a Austria?*) apuntan más bien a preguntas abiertas (Burguer et al., 2001; Maybury, 2003).

Conclusión

El *Partition Approach* puede verse como un camino para explicar los tres postulados de Hamblin, como destacan Groenendijk y Stokhof (cf. Groenendijk, y Stokhof, 1997, pp. 1076-8). Integra especialmente el segundo postulado, el cual

requiere que las posibles respuestas a una pregunta sean exhaustivas y constituyan un set de posibilidades mutuamente excluyentes. Esta es una diferencia importante con el *Proposition Approach*.

Hay una serie de fenómenos que están en el límite entre la semántica y la pragmática, tales como cuestiones ligadas a la presuposición en las preguntas o la dependencia del contexto de las preguntas y respuestas (cf. Groenendijk, y Stokhof, 1997, pp.1119-21). Esto lleva a Groenendijk y Stokhof a sugerir que únicamente una aproximación integrada y flexible puede manejar adecuadamente estas cuestiones: deben asumirse diferentes representaciones de las preguntas y las respuestas (dependiendo de los fenómenos a los que se desee dar cobertura en una específica pregunta o respuesta), donde reglas *type-coercion* establecerían correspondencias entre esas diferentes representaciones (Groenendijk, y Stokhof, 1997, pp. 1115-20). Por ejemplo, una pregunta recibiría tanto una representación *mention-some* como una *mention-all*, de las cuales la que mejor se adaptase al actual contexto sería la seleccionada. Manejar este tipo de fenómenos situados en el nivel pragmático podría hacerse a través de una aproximación integrada posiblemente en el marco de la semántica dinámica[79] (Groenendijk, y Stokhof, 1997, pp. 1120-2).

2.3.1.2.2 Jonathan Ginzburg: la importancia del contexto

En Ginzburg (1995; 1996), Jonathan Ginzburg describe una semántica de las preguntas donde los aspectos pragmáticos tienen una especial relevancia. Su propuesta se construye en el marco de la *situation theory* (Barwise, 1983), una teoría semántica fuertemente dependiente del contexto.

En la propuesta de Ginzburg el significado de la pregunta se construye a partir de un análisis semántico no dependiente de la situación (*situation-invariant*) junto con una serie de parámetros dependientes de la situación (*situation-dependent*): el contexto del discurso, el hablante, el oyente, el tiempo, etc. Ginzburg enfatiza el hecho de que encontrar una respuesta a una pregunta es *agent-relative,* es decir, depende en gran parte de las creencias, intenciones, etc., del usuario que la plantea. A continuación revisaremos algunos puntos relevantes de la teoría de Ginzburg.

El problema de enumerar todas las posibles respuestas

Ginzburg plantea que una representación de las preguntas que está basada en sus posibles respuestas (como hacen todas las propuestas presentadas hasta el momento, apoyándose en el segundo postulado de Hamblin) tiene dificultades cuando las respuestas no consisten en identificar uno o más individuos únicos. Así, propone el siguiente ejemplo:

(224) *What is the word for "relaxation" in Chukotian?*

(Tomado de Ginzburg, 1996, p. 400)

79 Cf.: http://plato.stanford.edu/entries/dynamic-semantics/

Ginzburg sostiene que, a partir de una pregunta como la de (224), no sería posible modelar una respuesta basada en alternativas. Si el hablante que plantea la pregunta no tiene o tiene muy poco conocimiento del idioma Chukotian, no puede concebir todo el posible rango de respuestas que la pregunta implica.

Sin embargo, contra esta crítica de Ginzburg podríamos argumentar que, intuitivamente, esperaríamos que alguien que pregunta (224) estará preparado para recibir como posible respuesta algo que considera una palabra, es decir, cualquier secuencia de sonidos o cualquier secuencia de caracteres escritos (Fliedner, 2007). Otra cuestión sería hasta qué punto aproximaciones semánticas como las presentadas tienen realidad psicológica en la mente de los hablantes (es difícil de asumir que tanto el que pregunta como el que responde necesiten representar mentalmente la pregunta como una enorme o casi infinita enumeración de todas las posibles respuestas).

Respuestas dependientes del contexto

Este es, como veremos, el punto en el que el planteamiento de Ginzburg se diferencia de los planteamientos presentados hasta ahora. Para Ginzburg, aquello que hace que una respuesta sea buena es altamente dependiente del contexto situacional. El conocimiento del que pregunta así como sus objetivos al realizar la pregunta deben ser especialmente tenidos en cuenta para evaluar si una determinada respuesta satisfará o no al que pregunta o, en términos de Ginzburg, si resolverá o no su pregunta. Así, por ejemplo, la siguiente pregunta:

(225) *¿Quién es Barack Obama?*

podría responderse con

(225a) *Es un hombre de 51 años.*
(225b) *Es el marido de Michelle Robinson.*
(225c) *Es el actual presidente de los EEUU.*

Las tres opciones son válidas, ya que en los tres casos se lleva a cabo una identificación del sujeto «Barack Obama». Sin embargo, las tres opciones no son igual de válidas en todas las situaciones. Dependiendo del contexto en el que se formule (225), se preferirá (225a), (225b) o (225c) como respuesta. Por ejemplo, en un entorno de BR factual (225c) se preferiría claramente, ya que cuando un usuario plantea una pregunta tipo *quién+ser*, lo que busca por lo general es información relacionada con la profesión o la actividad de la persona por la que se pregunta, y no una descripción física o una mera identificación. Volveremos sobre estos aspectos más adelante (cf. sección 3.1.4).

Granularidad de la respuesta

Este punto, muy relacionado con el anterior, se refiere al grado de especificación de una respuesta. La granularidad es altamente dependiente de los

intereses del que plantea la pregunta. Este factor es especialmente relevante cuando se necesita responder preguntas relativas a tiempo y lugar. Consideremos la siguiente pregunta:

(226) *Where is Deerfield, Illinois?*

(Ejemplo tomado de Fliedner, 2007, p. 82)

Las siguientes serían todas posibles respuestas correctas:

(226a) *87° 54' longitude, 42° 12' latitude.*
(226b) *Near Lake Michigan, about 20 miles north of Chicago.*
(226c) *Next to Highland Park.*
(226d) *On the planet Earth.*

(Tomado de Fliedner, 2007, p. 84)

También volveremos sobre este punto más abajo (cf. sección 3.1.4).

Representación de las preguntas de Ginzburg

En el planteamiento de Ginzburg se hace referencia explícita a las situaciones (elementos comparables a los mundos posibles en la lógica intensional[80], pero de los que se asume que tienen una estructura interna) y a un cierto marco de referencia (usado para modelar explícitamente al que pregunta). La definición de *resolvedness* de Ginzburg es la siguiente:

> A fact τ RESOLVES (s?μ) relative to a mental situation ms iff
> 1. Semantic condition: τ is a fact of s that potentially resolves μ
> 2. Agent relativization: τ ⇒ms Goal-content(ms)
> (Intuitively: τ entails the goal represented in the mental situation ms relative to the inferential capabilities encoded in ms.)
> (Ginzburg, 1996, p. 407)

Esto debe leerse de la siguiente manera:

- τ representa una solución a la pregunta, es decir, una respuesta satisfactoria.
- (s?μ) representa la pregunta. La situación *s* es la representación de la situación en la cual la pregunta se realiza y μ es su representación semántica (Ginzburg asume un término lambda como algo similar a la *subject part* de la representación de las preguntas de Belnap, cf. sección 3.1.1.2).
- τ ⇒ ms Goal-content(ms), significa que en ms, es decir, en el estado mental del que pregunta, algún hecho puede ser inferido que resuelva el actual objetivo del que pregunta (Ginzburg, 1995, pp. 499-504).

80 http://plato.stanford.edu/entries/logic-intensional/

Comparando este planteamiento con la solución sugerida por Groenendijk y Stokhof (cf. *supra*), se observa que la principal diferencia consiste en la representación explícita del objetivo del que pregunta y el requerimiento de que esto puede ser resuelto por la respuesta. La situación de Ginzburg puede (al menos) ser parcialmente descrita en términos de índices de la lógica intensional, el conocimiento del que pregunta puede ser modelado como sets epistémicos y doxásticos, pero el objetivo del que pregunta no es modelado explícitamente por Groenendijk y Stokhof. Este factor funciona además como un filtro adicional sobre respuestas posibles: en un ejemplo como (221) todas las respuestas en (221a), (221b) o (221c) se considerarían respuestas posibles en la aproximación de Groenendijk y Stokhof. La aproximación de Ginzburg aporta un filtro, al calcular las preferencias del que pregunta en lo concerniente a las distintas respuestas posibles.

2.3.1.2.3 Van Rooy

Van Rooy (van Rooy, 2003) supone un paso más allá en la interrelación entre semántica y pragmática. Al igual que Ginzburg, Van Rooy incide en la importancia del contexto para la interpretación de las preguntas.

En realidad, la influencia del contexto a la hora de determinar qué respuesta es apropiada no es algo tan controvertido, ya que se ha sugerido que el contexto puede ser el principal factor que determina si la interpretación de una pregunta es *mention some* o *mention all*, así como si es más apropiada una respuesta-constituyente o una respuesta-oración. Pero Van Rooy va más allá y argumenta que el contenido semántico de la pregunta en sí mismo es dependiente del contexto.

Por ejemplo, van Rooy (2003) da un par de escenarios relacionados que ilustran que *if Jill knows that she is in Helsinki, she may or may not be truly said to know where she is*:

(227) *Context: Jill about to step off a plane in Helsinki.*
Flight attendant: Do you know where you are?
Jill: Helsinki.
Flight attendant: Ah, OK. Jill knows where she is.

(228) *Context: Jill about to step out of a taxi in Helsinki.*
Driver: Do you know where you are?
Jill: Helsinki.
Driver: Oh, dear. Jill doesn't (really) know where she is.

(Tomado de van Rooy, 2003, p. 4)

La propuesta de van Rooy consiste en proporcionar un significado uniforme no especificado de las preguntas, aumentado por un mecanismo que después se encarga de especificar el significado en un contexto dado, teniendo en cuenta factores como los objetivos presumibles del intercambio de información. Por ejemplo: si una persona A pregunta dónde puede comprar un periódico italiano, puede ser relevante que esté tratando de decidir si caminar en dirección al museo o a

la estación; en ese caso, una «respuesta resolutiva» sería aquella que especificase si hay un lugar para comprar un periódico de camino al museo, de camino a la estación, o en ambos casos. Si hay uno, entonces mencionar otros lugares más lejanos que estos sería aportar información irrelevante extra (van Rooy, 2003).

En definitiva, en la línea de Ginzburg, van Rooy argumenta que la interpretación de los intercambios pregunta-respuesta depende del conocimiento de los objetivos de los participantes en una conversación. Para realizar esta interpretación, van Rooy utiliza consideraciones de la teoría de la información, así como medidas cuantitativas correspondientes a la «informatividad» o a la probabilidad de verdad de una proposición respecto al estado de información de una persona.

2.3.1.3 Conclusiones sobre la semántica de las preguntas

En esta sección hemos presentado distintas aproximaciones a la semántica de las preguntas: desde los planteamientos formales de los inicios, más próximos a la lógica, a los más modernos, que consideran indisoluble la relación entre semántica y pragmática.

A partir de los principios de Hamblin, los planteamientos formales trataron de representar la relación entre pregunta y respuesta partiendo del presupuesto de que el significado de una pregunta denota, de alguna manera, sus respuestas posibles. En la relación pregunta-respuesta, el análisis de la frase interrogativa se revela como fundamental para las preguntas parciales, pues es la frase interrogativa la que determina el tipo de variable que constituye la incógnita del que pregunta.

Pese a sus evidentes diferencias, Fliedner (2007, p. 49) señala que la mayoría de los análisis semánticos formales de la relación pregunta-respuesta se construyen, en mayor o menor medida, en torno a la siguiente idea central[81]:

- **Interrogativas totales**: representación: ?φ, donde φ es la proposición contenida en la pregunta. A través de la pregunta, el hablante quiere saber si la extensión de la proposición es verdadera o falsa (es decir, φ o ¬φ).

  (229) *Did Lee Harvey Oswald kill John F. Kennedy?*
  (230) *Yes, Lee Harvey Oswald killed John F. Kennedy.*

  (229a) ?kill' (lee_harvey_oswald' , john_f_kennedy')
  (230a) kill' (lee_harvey_oswald' , john_f_kennedy')

(Tomado de Fliedner, 2007)

- **Interrogativas parciales**: representación: ?x1 ...xnφ, donde las variables x1 ...xn corresponden a las frases interrogativas que aparecen en ?φ. Una respuesta (parcial) correspondería a φ cuando a1 ...an instancian x1 ...xn, respectivamente, de manera que no permanecen variables libres. Por lo tanto, una pregunta enlaza cada frase interrogativa con un individuo en el dominio:

81 El operador ? se usa para marcar «pregunta», *vs.* φ que indica «proposición».

(231) *Who killed John F. Kennedy?*
(232) *Lee Harvey Oswald killed John F. Kennedy.*

(231a) ?x1 kill' (x1, john_f_kennedy')
(232a) kill' (lee_harvey_oswald', john_f_kennedy')

(Tomado de Fliedner, 2007)

Por otra parte, las distintas aproximaciones formales han propuesto soluciones para cuestiones clave en la relación pregunta-respuesta como la relación entre preguntas y respuestas-constituyente/respuestas-oración (Hamblin, 1958; Reich, 2003), la interacción entre foco y *background* (Stechow, 1991; Reich, 2003; Krifka, 2001), el tamaño de la selección de la respuesta (Belnap y Steel, 1976, y, con nociones pragmáticas: Groenendijk y Stokhof, 1984; 1989; 1992; 1997, entre otros), etc.

Por otra parte, en las aproximaciones más recientes que combinan semántica y pragmática, se observa una importancia creciente en el papel que juega el contexto en la interpretación del significado de las preguntas. La evolución es clara desde aproximaciones como la de Groenendijk y Stokhof, que tiene mucho de semántica formal, a las teorías de van Rooy, donde se da un significado uniforme no especificado para las preguntas que es aumentado por un mecanismo que especifica el significado concreto en un contexto dado, manejando factores como los objetivos presumibles del intercambio de información. Aspectos como la granularidad de las preguntas (Ginzburg), o la modelización de los objetivos del que pregunta en la interpretación de la respuesta se muestran como especialmente interesantes para nuestra aproximación, como veremos a continuación.

2.3.1.4 Significado a través del contexto: algunos apuntes sobre la pragmática de las preguntas

Antes de pasar al análisis semántico de las partículas interrogativas, nos interesa esbozar en esta sección algunos aspectos relativos a la pragmática en un sistema de BR y la relación de estos aspectos con el significado de las preguntas.

Como los aspectos pragmáticos no son tratados en SpQA, no profundizaremos demasiado en estas cuestiones: solo nos interesa apuntar cuáles son los factores pragmáticos que pueden ser determinantes en un sistema de BR a la hora de establecer el significado de una pregunta.

2.3.1.4.1 Imprecisión del significado y factores pragmáticos en un sistema de BR

Como se ha apuntado y como veremos en detalle en la sección 3.2.1, muchas preguntas (no todas) tienen un significado impreciso. Por ejemplo, a la pregunta:

(233) *¿Cuándo se estrenará la película El Hobbit 2?*

podríamos responder con:

(233a) *En octubre.*
(233b) *En octubre de este año.*
(233c) *El 15 de octubre de 2013.*

Y todas las respuestas serían semánticamente válidas. La diferencia entre una u otra viene determinada por el grado de precisión en la respuesta (la granularidad) que le interese al usuario cuando plantea su pregunta.

Como apunta Jonathan Ginzburg (cf. *supra*), en estos casos lingüísticamente imprecisos los factores pragmáticos son clave para determinar el significado concreto de la pregunta y ofrecer una respuesta satisfactoria al usuario.

Escenario de interacción en un sistema de BR

En un escenario de BR en el que la interacción usuario-sistema se limita al planteamiento de la pregunta y la obtención de una respuesta automática, los factores pragmáticos a tener en cuenta son más limitados que los que entran en juego en una interacción lingüística entre humanos (no hay lenguaje gestual, por ejemplo, ni conocimiento compartido...). En ese escenario de BR, los factores pragmáticos que pueden determinar el significado de la pregunta son:

- **El usuario**: ya hemos visto a través de Ginzburg y otros planteamientos semántico pragmáticos que los objetivos del usuario, así como su conocimiento del mundo, juegan un papel determinante en el significado de la pregunta. Este factor siempre está presente, pero se hace especialmente relevante en casos como el de la pregunta (234), en los que el significado es impreciso y depende directamente de cuáles sean los intereses del usuario.

Acceder a los intereses del usuario en un entorno de BR solo es posible a través de un diálogo o algún tipo de interacción. Sin interacción, ante preguntas imprecisas como:

(234) *¿Dónde está Nueva York?*

siempre hay varias respuestas posibles igualmente válidas.

- **El perfil del sistema de BR**: hemos visto en el capítulo 1 que existen distintos tipos de sistemas de BR; sistemas de dominio general que manejan información factual, sistemas de dominio restringido que manejan información de cualquier tipo, etc. El perfil del sistema de BR puede influir en la determinación del significado de las preguntas. Por ejemplo, ante una pregunta como:

(235) *¿Quién es Mariano Rajoy?*

un sistema de BR general de tipo factual no respondería (o no debería responder, al menos) con

(236) *El marido de Elvira Fernández Balboa.*

Lo esperable sería que el sistema ofreciese una respuesta que recogiese otro tipo de información factual, como la profesión de Rajoy e incluso algún dato de tipo enciclopédico como el lugar de nacimiento, estudios, etc.

El perfil del sistema de BR está directamente relacionado con los intereses del usuario: un determinado perfil de sistema de BR está pensado para un determinado perfil de usuario (con unos determinados objetivos). Así, ante la pregunta:

(237) *¿Qué es el lupus?*

Un sistema de BR general de tipo factual ofrecería una respuesta corta y concisa, tipo definición, como:

(237a) *Es una enfermedad autoinmune crónica.*

Mientras que un sistema de BR especializado en medicina, ofrecería una respuesta más compleja y elaborada:

> (237b) *El lupus eritematoso sistémico (LES o lupus) es una enfermedad autoinmune crónica que afecta al tejido conjuntivo, caracterizada por inflamación y daño de tejidos mediado por el sistema inmunitario, específicamente debido a la unión de anticuerpos a las células del organismo y al depósito de complejos antígeno-anticuerpo* [...][82]

Un médico que buscase información en el primer sistema probablemente se sentiría insatisfecho con la respuesta (237a), del mismo modo que un usuario medio que plantease la misma pregunta y recibiese la respuesta (237b) probablemente recibiría un exceso de información.

En las secciones siguientes profundizaremos en la semántica de las palabras interrogativas, intentando discernir aquellas situaciones en las que el significado de la pregunta es inherentemente impreciso y, por lo tanto, depende de factores externos a la propia pregunta (como los anteriores) para ser concretado.

2.3.2 Aspectos semántico pragmático concretos

2.3.2.1 Semántica y pragmática de los interrogativos

En las secciones anteriores se ha explicado la importancia de las partículas interrogativas en el funcionamiento y significado de las interrogativas parciales.

82 Extraído de: http://es.wikipedia.org/wiki/Lupus

Como hemos visto, a partir del constituyente interrogativo se determina el valor semántico de la incógnita que expresa la pregunta y, al menos hasta cierto punto, el tipo de respuesta esperada. Es por esa razón por la que el correcto procesamiento del constituyente interrogativo es clave en un sistema de BR.

Por otra parte y como ya hemos adelantado, existen preguntas cuyo significado es inherentemente inespecífico. Esta falta de concreción del significado está directamente relacionada con el carácter del interrogativo y de las construcciones en las que este se inserta. De este modo, y como veremos a continuación, existen interrogativos con un valor semántico general concreto y otros con un valor semántico general impreciso. Además, algunos interrogativos de valor impreciso concretan su significado en ciertas construcciones sintácticas, mientras que otros no lo hacen, de manera que solo es posible establecer su valor concreto a través de la interacción con el usuario.

Como veremos (cf. capítulo 4, sección 4.4.3.4.2.2.2, en el análisis de SpQA se busca determinar el valor semántico que denota el interrogativo, es decir: el valor semántico de la variable o incógnita que representa el interrogativo. Para nosotros, este valor coincide con el del tipo de respuesta esperada, ya que, en la línea de Hamblin (cf. *supra*), consideramos que las frases interrogativas «generan» un espacio muy específico que solo puede ser llenado por estructuras semánticas concretas cuyas características están más o menos codificadas en la propia frase interrogativa. Para un sistema de BR es fundamental intentar establecer hasta qué punto la respuesta esperada está concretada en la pregunta. Por eso, en nuestro análisis prestaremos especial atención a este aspecto, intentando discernir el grado de concreción o precisión de la variable contenida en cada interrogativo.

En las páginas que siguen nos ocuparemos del estudio semántico de las partículas interrogativas. Las analizaremos una a una, intentando desentrañar los factores semánticos de su funcionamiento que son relevantes para establecer el significado general de las preguntas. Teniendo en cuenta que para determinar este significado no solo es importante el interrogativo como unidad léxica sino también las construcciones sintácticas en las que este se inserta, en las secciones que siguen nos ocuparemos de ambos aspectos: rasgos semánticos de la unidad léxica y rasgos semánticos de la unidad en determinadas construcciones sintácticas.

Se seguirá el mismo orden de exposición que en la sección 2.4.3: primero se presentarán aquellas partículas que funcionan prototípicamente como adjuntos (*cuándo*, *dónde*, *cómo* y *por qué*) y a continuación aquellas que funcionan prototípicamente como argumentos (*quién*, *cuánto*, *cuál* y *qué*).

2.3.2.1.1 El significado de las partículas interrogativas

CUÁNDO

Valores semánticos del interrogativo

(238) *¿Cuándo se estrenará la nueva colección de Mango?*

La variable que denota el interrogativo tiene un significado general de 'ubicación temporal'.

El valor temporal exacto de la variable no está especificado en el interrogativo, de manera que *cuándo* apunta siempre a una ubicación temporal no precisa. De este modo, cualquiera de las siguientes respuestas para (238) sería correcta desde el punto de vista semántico:

(238a) *En otoño.*
(238b) *En septiembre.*
(238c) *En 2013.*
(238d) *En septiembre de 2013.*
(238e) *El 12 de septiembre de 2013.*

Pues todas ubican en un espacio temporal el evento por el que se pregunta: el estreno de la nueva colección de Mango.

El valor temporal suele construirse con el verbo en pasado, ya que las preguntas con *cuándo* suelen apuntar a hechos ya ocurridos (239). No obstante, también son posibles preguntas con este valor temporal con el verbo en presente (240) o futuro (238).

(239) *¿Cuándo fue el descubrimiento de América?*
(240) *¿Cuándo comienza y cuando termina la cuaresma?*

Por otro lado, *cuándo* puede usarse también con un valor no temporal[83]. Es el caso de preguntas como:

(241) *¿Cuándo 2 o mas fracciones son equivalentes?*
(242) *¿Cuándo da positivo la alcoholemia en ciclomotores?*

En estos ejemplos, el valor semántico de *cuándo* sería equivalente a 'identificación de una situación en la que se da la condición expresada en la pregunta'. De este modo, el significado de (241) sería similar a 'quiero determinar la situación en la que dos fracciones son equivalentes', y el de (242) 'quiero saber la situación en la que se da positivo en alcoholemia en ciclomotores'. Probablemente este uso sea una extensión metafórica del uso temporal: de la ubicación temporal se pasa a la ubicación existencial en un escenario hipotético. Parece que este valor siempre se usa con el verbo en presente.

Con este valor, el significado de la variable también es impreciso, pero de un modo distinto al que veíamos para el valor temporal. En este caso, la identificación del escenario en el que se produce el evento de la pregunta puede llevarse a cabo con más o menos complejidad, dependiendo de los intereses del usuario. Así, la identificación de *cuál es el escenario en el que 2 o más fracciones son equivalentes* puede hacerse con distintos niveles de complejidad y extensión. Sin embargo, siempre se trata del mismo valor para la variable: la identificación del escenario en el que se produce ese evento, que solo es uno, pero que puede expresarse de distintas

83 Los trabajos teóricos consultados no recogen este valor semántico para *cuándo*.

maneras. En el caso de la localización temporal, el evento que se quiere localizar puede apuntar a varios escenarios temporales distintos y todos ellos válidos. Creemos que con el valor temporal las preguntas con *cuándo* son de tipo informativo, afectadas por el factor de la granularidad, mientras que con el valor hipotético, las preguntas con *cuándo* son preguntas abiertas.

Determinación del valor semántico de la variable

Distinguir casos como los de (241) y (242) de casos como (238), (239) y (240) solo es posible si se maneja conocimiento del mundo que permita diferenciar una ubicación temporal de un hecho ((238), (239) y (240)), de la identificación de un escenario hipotético ((241) y (242)). Aspectos lingüísticos concretos como el tiempo verbal pueden ayudar, pero no son determinantes.

Por otra parte, cuando el valor es temporal, hay dos elementos que pueden ayudar a concretar el tipo de localización temporal que se está demandando: el tipo de verbo y el tipo de evento que debe ser ubicado temporalmente. Veamos unos ejemplos:

(243) *¿Cuándo falleció Amy Winehouse?*
(244) *¿Cuándo fue domesticado el perro?*
(245) *¿Cuándo duerme el koala?*

Las tres preguntas anteriores pertenecen a un entorno de BR[84], y en cada una de ellas *cuándo* apunta a un valor temporal distinto: en (243) se pregunta por la ubicación temporal de un hecho histórico (*el fallecimiento de Amy Winehouse*), por lo tanto, lo más probable es que se pregunte por una fecha; en (244) se pregunta por otro hecho histórico, pero, en este caso, por uno cuya ubicación temporal exacta no es posible (*vs.* (243)), de manera que lo que se demanda es un período histórico; en (245) se pregunta por la ubicación temporal de una actividad de la vida de un animal (*dormir*), por lo que se pregunta por una parte del día.

Como vemos, tratar de determinar el valor temporal exacto de la variable a la que apunta *cuándo* implica manejar una gran cantidad de conocimiento lingüístico y también de conocimiento del mundo; eso nos permite saber, por ejemplo, que el fallecimiento de Amy Winehouse (*fallecer* + *Amy Winehouse*) en (243) es un evento considerado un hecho histórico que se ubica temporalmente a través de una fecha, mientras que *el momento en el que duerme un koala* (*dormir* + *el koala*) en (245) es una actividad de la vida diaria de un animal que se ubica temporalmente en una parte del día.

En un sistema de BR, la imprecisión en el valor de la variable temporal puede no representar un problema tan grave ya que puede resolverse en la fase de extracción de la respuesta. Esto es así porque los eventos en las preguntas con *cuándo* suelen tener una ubicación temporal más o menos definida, prototípica, y no varias posibles. Así, por ejemplo, un hecho histórico se ubica comúnmente mediante una fecha, de manera que si el sistema de BR realiza una búsqueda del hecho

84 (243) y (245) pertenecen a un portal de preguntas *online*, Wikirespuestas (cf. capítulo 3, sección 3.2.1.3) y 244 forma parte de uno de los sets de preguntas de las evaluaciones CLEF.

fallecimiento de Amy Winehouse, lo más probable es que se encuentre con una fecha y no, por ejemplo, con un período histórico. Sin embargo, esto no resuelve del todo el problema. En primer lugar, porque sería perfectamente posible encontrarse un evento como *el fallecimiento de Amy Winehouse* asociado en un texto a, por ejemplo, una parte del día:

(246) *Amy Winehouse murió a las 2:00 de la mañana en su domicilio.*

Esta sería una respuesta semánticamente compatible con la pregunta de (243). Y esto nos lleva a nuestro segundo punto: este problema podría resolverse si se asociase *cuándo* con un valor temporal concreto en el sistema de BR, por ejemplo, con el valor de 'fecha'. Sin embargo, haciendo esto, se limitaría bastante el tipo de preguntas que un usuario podría hacer ya que, por ejemplo, con una pregunta como (243), el usuario puede efectivamente estar interesado en la hora exacta del fallecimiento de Amy Winehouse y no en la fecha.

En definitiva, para el valor temporal, los múltiples valores temporales con los que sea semánticamente compatible la combinación del verbo y el evento contenido en la pregunta definen el rango de imprecisión de la variable contenida en *cuándo*.

Por lo tanto, en términos generales, asignar un valor a la variable definida por *cuándo* presenta dos problemas:

- En primer lugar, determinar si se trata de una variable de tipo temporal o no.
- En segundo lugar, dado que se trate de una variable de tipo temporal, establecer el valor exacto de esa variable entre los posibles (una fecha, un año, un período histórico, una parte del día, etc.).

En la sección de corpus (cf. capítulo 3, sección 3.3.3.1) profundizaremos en los posibles valores de *cuándo* en preguntas reales y retomaremos estas cuestiones.

DÓNDE

Valor semántico del interrogativo

(247) *¿Dónde está la ciudad de Los Ángeles?*

La variable que denota *dónde* tiene el significado general de 'localización espacial'. Con *dónde* ocurre algo similar a lo que acabamos de ver con *cuándo*: el valor locativo es impreciso, de manera que existe un amplio rango de posibilidades para contestar correctamente a una pregunta como (247):

(a) *Está en el condado de Los Ángeles.*
(b) *Está en el estado de California.*
(c) *Está en los Estados Unidos.*
(d) *Está en América.*

Con respecto a la imprecisión del valor de la variable, el caso de *dónde* es todavía más complejo que el de *cuándo*. Acabamos de ver que el tipo de evento con el que se asocia *cuándo* puede ayudar a determinar el valor exacto de la variable, ya que existen unos valores de ubicación temporal prototípicos para determinados eventos ('fecha' para un evento histórico, por ejemplo). Sin embargo, con el ejemplo (247) queda claro que existen distintas posibilidades de «ubicación en el espacio» igualmente válidas para una misma entidad. Todo depende del nivel de precisión deseado dentro del concepto 'localización espacial'. El rango de posibilidades que abre *dónde* parece ser, por lo tanto, mayor que el de *cuándo*.

Determinación del valor semántico de la variable

Al igual que con *cuándo*, el valor semántico de *dónde* es siempre impreciso. Para casi todos los tipos de localización espacial hay varias posibilidades igual de válidas que se diferencian por el grado de exactitud que requiere el que pregunta. Por lo tanto, la determinación del valor concreto de la variable solo parece posible si:

- Se accede a los intereses del usuario.
- Se establece de antemano ese valor en el sistema de BR (es decir: todas las preguntas con *dónde* se deben responder con determinado tipo de ubicación espacial, por ejemplo, con el país siempre que se trata de una ubicación geográfica).

La segunda opción tiene la contrapartida que ya hemos visto para *cuándo*: preestablecer un valor limita las posibilidades del sistema, ya que el usuario puede estar interesado en localización espacial con distintos niveles de precisión según la entidad por la que se pregunte.

Las preguntas con *dónde* parecen por tanto de tipo informativo, pero con un significado impreciso debido al factor de la granularidad.

CÓMO

Valor semántico del interrogativo

La variable que denota puede tener varios significados. El principal es el de 'modo' o 'manera' en que se realiza la acción expresada por el verbo:

(248) *¿Cómo se hace un risotto?*

NGRALE señala también un valor de tipo causal, donde *cómo* sería equivalente a *por qué* o a *cómo es posible que:*

(249) *¿Cómo sabía Gandalf que Frodo debía portar el anillo?*

Esta lectura causal está presente en muchas oraciones negativas (Real Academia Española, 2009, 22.16c):

(250) *Entonces, madre, si Céspedes mentía, ¿cómo no le estudiaste la pupila?*

(Ejemplo tomado de Real Academia Española, 2009, 22.16c)

El valor causal también aparece en oraciones no negativas, «cuando el que habla manifiesta no entender alguna situación o desconocer las causas que conducen a ella» (Real Academia Española, 2009, 22.16e):

(251) *Pero, hijas mías, ¿cómo pretendéis que yo la cure?*

(Ejemplo tomado de Real Academia Española, 2009, 22.16e)

Y en estructuras en las que la pregunta que *cómo* introduce se interpreta como apódosis de un período condicional, sea antepuesto o pospuesto (Real Academia Española, 2009, 22.16f):

(252) *¿Cómo dices que te interesa si no le prestas la menor atención?*

(Ejemplo tomado de Real Academia Española, 2009, 22.16f)

Creemos que el valor causal de *cómo* en todas estas construcciones parece más bien propio de contextos dialogados y no de preguntas sin contexto, por lo que es poco probable en un entorno de BR. En caso de documentarse, consideramos que lo más probable sería encontrarlo en la última construcción, como apódosis de un período condicional. Por lo tanto, creemos que en un contexto de BR el principal valor semántico de *cómo* es el modal. En la sección de corpus volveremos sobre estas cuestiones.

Determinación del valor semántico de la variable

El valor modal es en general bastante impreciso. La variable que denota el interrogativo tiene un significado general de 'modo, manera', pero el valor concreto de ese significado depende totalmente del verbo sobre el que incida *cómo*. De este forma, el valor concreto al que apunta *cómo* es distinto en (253), (254) y (255):

(253) *¿Cómo se llama al texto de una ópera?*
(254) *¿Cómo es el ambiente del pájaro cuclillo?*
(255) *¿Cómo se hace el mecate?*

En (253), *cómo* + *llamarse* denota un valor de 'denominación' para la variable; en (254), *cómo* + *ser* denota un valor de 'descripción' para la variable; en (255), *cómo*+*hacer* denota un valor de 'procedimiento'.

Por lo tanto, el significado concreto de *cómo* parece dependiente del verbo al que acompañe. Podemos establecer un significado impreciso de 'modo, manera', que es general y característico del interrogativo, frente a posibles valores semánticos

concretos que derivan de la suma del significado del interrogativo y el verbo al que acompaña. En la sección de corpus estudiaremos esta cuestión.

Las preguntas con *cómo* pueden ser, por tanto, tanto abiertas como informativas. Serán informativas cuando el valor de la variable sea un valor concreto (256), y abiertas cuando el valor de la variable sea impreciso (257):

(256) *¿Cómo se llama el actual presidente del gobierno?*
(257) *¿Cómo es un clima tropical?*

POR QUÉ

Valor semántico del interrogativo

El valor semántico general es el de 'causa' o 'razón' de aquello que es expresado por el resto de la pregunta:

(258) *¿Por qué flota un barco?*

Así, el significado de (258) sería algo similar a: 'quiero saber la razón por la cual un barco flota'.

La construcción *por qué* + *no* puede tener dos interpretaciones (Real Academia Española, 2009, 22.16o):

- la causal (259);

(259) *¿Por qué no llegó Colón a las Indias?*

- la de sugerencia (260).

(260) *¿Por qué no sales a que te dé un poco el aire?*

Como vemos, en la interpretación causal, la negación es interna (cf. capítulo 2, sección 4.1.1), mientras que en la interpretación de sugerencia la negación es externa.

NGRALE señala una serie de factores gramaticales que permiten distinguir las dos interpretaciones:

- Con tiempos verbales distintos al presente (261) y con perífrasis prospectivas (262) generalmente se da la lectura causal:

(261) *¿Por qué no han venido?*
(262) *¿Por qué no iba a estar bien?*

- Con el presente de indicativo se da la lectura de sugerencia (263), salvo cuando el sujeto es preverbal (264), en cuyo caso se da la lectural causal:

(263) *¿Por qué no sales un momento?*
(264) *¿Por qué el BCE no responde?*

Si el sujeto es postverbal (265), las dos interpretaciones son posibles:

(265) *¿Por qué no responde el BCE?*

- Con términos de polaridad negativa se bloquea la lectura de sugerencia:
  (266) *¿Por qué no ayuda nadie a Grecia?*

- Con las interrogativas de infinitivo la interpretación debe ser de sugerencia:
  (267) *¿Por qué no dejar eso a los políticos?*

Creemos que este valor de sugerencia no es propio de un sistema de BR de tipo factual, pero tal vez si lo sería de un sistema de BR que trabajase sobre opiniones o hipótesis (268):

(268) *¿Por qué no desocupar Irak?*

En la sección de corpus analizaremos esta cuestión.

Determinación del valor semántico de la variable

El valor semántico de *por qué* es de los más complejos entre los interrogativos (al menos desde el punto de vista de la BR). Eso es así porque la variable, 'causa', es altamente abstracta y, por lo tanto, imprecisa. Encontrar una respuesta que explique la causa de algo y determinar hasta qué punto esa respuesta es válida o suficiente es muy difícil y además depende, todavía más que en los casos anteriores, de los intereses del usuario.

Las preguntas con *por qué* son, por lo tanto, siempre preguntas abiertas.

Por esta razón, este tipo de preguntas apenas se han tratado en los sistemas de BR (para un ejemplo de este tipo de estudio cf. Verberne, 2010).

QUIÉN

Valor semántico del interrogativo

Quién denota el valor de 'persona', de la que se pide, en general, una identificación:

(269) *¿Quién dirigió la Aventura del Poseidón?*

En construcciones no copulativas, esa identificación suele consistir en aportar el nombre propio de la persona que cumple los requisitos establecidos en la pregunta (en (269), *haber dirigido la Aventura del Poseidón*). En estos casos, el valor de la variable es solo uno y concreto. En estas construcciones, también cabe como

respuesta el valor de 'descripción': en lugar de aportar el nombre de la persona por la que se pregunta, se aporta algún tipo de descripción que la identifica. Siguiendo con (269):

(270) *El mismo que dirigió El Millonario*.

Este tipo de respuestas, sin embargo, contarían como respuestas indirectas, ya que no identifican directamente a la persona por la que se pregunta.

En las construcciones copulativas, que como hemos visto son siempre de tipo ecuativo en el caso de *quién* (cf. capítulo 2, sección 2.4.3.3), podemos tener distintos tipos de identificación dependiendo de las características de la construcción sintáctica en la que se inserte *el interrogativo:*

- *Quién* + *ser* + frase nominal (cuyo núcleo es un nombre común) (271) o adjetiva (272): se manejan los mismos valores vistos hasta ahora, de manera que la identificación puede corresponderse con un nombre (271a, 272a) o con una descripción[85] (271b, 272b).

  (271) *¿Quién fue el descubridor de América?*
  (271a) *Cristóbal Colón.*
  (271b) *Un navegante genovés.*

  (272) *¿Quién de los hermanos Marx era mudo?*
  (272a) *Harpo Marx.*
  (272b) *El que llevaba peluca naranja y gabardina.*

- *Quién* + *ser* + nombre propio (273): en este caso, la identificación se corresponde con una **descripción** de la persona que se presenta en la pregunta.

  (273) *¿Quién es Barack Obama?*

Como el valor de 'descripción' es muy impreciso, ya hemos visto (cf. *supra*) que esta pregunta puede responderse de formas muy distintas:

(273a) *Es un hombre de 51 años.*
(273b) *Es el marido de Michelle Robinson.*
(273c) *Es el actual presidente de los EEUU.*

Todas estas respuestas son válidas para (273), puesto que implican una identificación de algún tipo para el sujeto *Barack Obama*. Por tanto, al igual que ocurre con *cuándo* o *dónde*, el valor 'descripción' para *quién* en este tipo de construcciones es impreciso.

85 Que, como en las no copulativas, da lugar a respuestas indirectas.

Determinación del valor semántico de la variable

Como acabamos de ver, en los casos no copulativos la variable apunta generalmente a la identificación mediante un nombre propio (269), aunque también se aceptaría una descripción como respuesta indirecta. En un entorno de BR de tipo factual, lo ideal sería aportar el nombre propio.

En las construcciones copulativas, el valor de la variable depende de la construcción sintáctica: si lo que hay que identificar no es un nombre propio, los valores son los mismos que para las construcciones no copulativas, nombre propio (271a, 272a) o descripción (271b, 272b). Como para las construcciones no copulativas, creemos que en un entorno de BR lo ideal sería aportar el nombre propio como respuesta. Si lo que hay que identificar es un nombre propio (273), el valor de la variable es siempre de 'descripción'. En estos casos, la variable es imprecisa y necesita de factores externos a la pregunta para ser concretada: intereses del usuario, perfil del sistema de BR, etc.

En la tradición de la BR general de tipo factual, este valor de 'descripción' de las construcciones copulativas tiene un significado más o menos establecido: consiste en determinar el rasgo o rasgos por los cuales es conocida la persona o personas por las que se pregunta. De hecho, en la Webclopedia uno de los Qtargets es *Why-Famous* (Hovy et al., 2000). En la mayoría de los casos, ese rasgo o rasgos apunta a una identificación de tipo profesional (la respuesta (273c) para (273)), pero no siempre (274a).

(274) *¿Quién es Frodo Bolsón?*
(274a) *El protagonista del Señor de los Anillos.*

En la sección de corpus volveremos sobre los posibles valores de *quién* en construcciones copulativas y no copulativas.

Las preguntas con *quién*, por tanto, pueden ser tanto informativas como abiertas. Cuando la variable es concreta y se pide un nombre propio, la pregunta es informativa. Cuando la variable es imprecisa y se pide una descripción, la pregunta es abierta.

CUÁNTO

Valor semántico del interrogativo

El valor de la variable consiste siempre en algún tipo de cuantificación de una entidad:

(275) *¿Cuánto ganas al mes?*
Variable = cantidad (de dinero) que ganas al mes.

(276) *¿Cuántos libros has comprado?*
Variable =cantidad de libros que has comprado.

La cuantificación siempre incide sobre alguna entidad. En los casos en los que *cuánto* funciona como determinante, la entidad cuantificada está presente en la pregunta (276). En los casos en los que *cuánto* funciona como pronombre, la entidad que es cuantificada se sobrentiende por el cotexto, fundamentalmente, se determina a través del contenido semántico del verbo más la noción de 'cuantificación'. Así, en (275) se pregunta por *cuánto* + *dinero* y en (277) por *cuánto* + *medida de longitud*:

(277) *¿Cuánto mide la Torre Eiffel?*

Determinación del valor semántico de la variable

En el caso de *cuánto*, no existe imprecisión en el valor de la variable: esta se corresponde siempre con una cuantificación concreta definida en la pregunta; en (275) la cuantificación de lo que alguien gana al mes; en (276) la cuantificación de los libros que alguien ha comprado; en (277) la cuantificación de lo que mide la Torre Eiffel.

Las preguntas en las que *cuánto* funciona como pronombre (275), (277), son más complejas semánticamente porque en ellas hay que deducir por el cotexto (por el verbo, fundamentalmente), qué tipo de entidad está siendo cuantificada.

Las preguntas con *cuánto* son, por tanto, siempre preguntas de tipo informativo.

CUÁL

Valor semántico del interrogativo

(278) *¿Cuál de los actores de El Barco salía en Verano Azul?*

El valor de la variable que denota *cuál* es distinto a los vistos hasta el momento. En este caso, la variable no tiene un contenido semántico propio «prototípico» como el de *cuándo* (tiempo), *dónde* (lugar), *cómo* (modo, manera), *por qué* (causa), *quién* (persona) y *cuánto* (cantidad).

La presencia de *cuál* implica la identificación de una o varias entidades respecto a un conjunto. En las preguntas con *cuál* (excepto en ciertas construcciones copulativas, como veremos), se nos presentan una serie de entidades entre las que debemos elegir la correcta de acuerdo a unos parámetros explícitos en la pregunta. El interrogativo simplemente apunta a esa entidad correcta como hace un relativo con su antecedente, aportando el significado de 'identificación'. En (278), por ejemplo, el conjunto de entidades es *los actores de El Barco* y la entidad (entre *los actores de El Barco*) a la que apunta *cuál* es aquella que cumple la condición de *salir en Verano Azul*. *Cuál*, por lo tanto, es un elemento vacío de significado que simplemente apunta a otra entidad, que puede ser de cualquier tipo: animada o no animada; concreta o abstracta; contable o no contable...

Para que la identificación sea posible, el conjunto de entidades a las que potencialmente apunta el interrogativo debe estar presente/activo en el contexto, bien

discursivo, bien situacional. Como recoge NGRALE: «El interrogativo *cuál* exige un antecedente [...], o bien un consecuente o un subsecuente, [...]» (Real Academia Española, 2009, 22.13h). Y como ya vimos en la sección 2.4.3.3.1 del capítulo 2, a diferencia de *qué*, *cuál* siempre apunta a otro elemento cuya presencia se exige en el co(n)texto inmediato. *Cuál* tiene además un valor tanto anafórico como catafórico, de manera que, en una pregunta, el conjunto de entidades al que apunta aparece siempre después del interrogativo (consecuente o subsecuente):

(280) *¿Cuál* continente[86] *se ubica totalmente en el hemisferio occidental?*
(281) *¿Cuál* de los hermanos Marx *no hablaba?*

Un caso especial lo constituyen ciertas construcciones copulativas con *cuál:*

(282) *¿Cuál era* el lema de la legión romana*?*

Estas copulativas son siempre ecuativas (cf. sección 2.4.3.3.1 del capítulo 2), de manera que lo que se produce es una identificación de *cuál* directamente con su referente (en (282) *el lema de la legión romana*). En este caso no hay un conjunto de entidades entre las que se debe seleccionar la correcta, sino que se nos proporciona directamente la entidad, sobre la que la pregunta nos exige algún tipo de información de carácter identificativo. La identificación exigida puede ser de distintos tipos. Puede demandarnos una entidad concreta: un lema en (282), un nombre en (283). O puede demandarnos conceptos más abstractos como el de (284), que son más imprecisos y difíciles de concretar.

(283) *¿Cuál es la zona térmica de Mongolia?*
(284) *¿Cuál fue el propósito del gobierno republicano y el segundo imperio?*

También es posible que se nos indique un conjunto en la estructura copulativa, dentro del cual debemos identificar la entidad correcta de acuerdo a una condición que es determinada generalmente por el constituyente funcionando como predicativo. En estos casos la pregunta suele contener una lista de elementos entre los que se debe elegir. Esa lista puede presentarse bajo formas sintácticas diferentes (285), (286).

(285) *¿Cuál de los siguientes elementos es más importante para la atmósfera: el nitrógeno, el oxígeno o el hidrógeno?*
(286) *¿Cuál de estos ríos, Darling, Zambeze, Ganges, Mackenzie, es el más importante que se localiza en Asia?*

Como ocurría con *quién*, en las construcciones con *cuál* también se puede responder con una descripción de la entidad por la que se pregunta:

86 Ya hemos mencionado en la sección 2.4.3.2 que el uso de *cuál* como determinante es extraño en España. Aún así, lo recogemos en nuestra descripción por ser usado ampliamente en otras zonas hispanohablantes.

(287) *¿Cuál de los Beatles compuso Julia?*
(288) *El que estaba casado con Yoko Ono.*

Como también ocurría con *quién*, este tipo de respuestas constituyen respuestas indirectas.

Teniendo en cuenta lo expuesto anteriormente, en un contexto de BR sin diálogo solo son esperables las preguntas en las que *cuál* apunta a un consecuente o subsecuente presente en la propia pregunta. Por tanto, en este contexto de BR solo son esperables preguntas con *cuál* con tres estructuras sintácticas:

1) *Cuál* como determinante acompañando a un sustantivo: el sustantivo constituye el conjunto, la clase o tipo del referente al que apunta el interrogativo (en (289): *ciudad*).

(289) *¿Cuál* ciudad *tiene más población en el mundo?*

2) *Cuál* como pronombre modificado por una frase preposicional con una frase nominal en su interior: la frase nominal determina el conjunto, la clase o tipo del referente al que apunta el interrogativo (en (290): *los Beatles*).

(290) *¿Cuál de* los Beatles *tocaba la batería?*

3) *Cuál* como pronombre en una estructura copulativa ecuativa en la que el referente al que apunta el interrogativo es directamente el constituyente de la ecuativa que funciona o bien como sujeto o bien como predicativo (en (291): *la causa del tsunami que asoló Malasia*).

(291) *¿Cuál fue* la causa del tsunami que asoló Malasia*?*

Determinación del valor semántico de la variable

Como ocurría con *quién*, *cuál* demanda la identificación de una entidad. Esta identificación puede ser precisa o imprecisa, dependiendo del tipo de entidad que deba ser identificada En la mayoría de los ejemplos que acabamos de ver, la entidad es una entidad concreta, de manera que la identificación demanda algo concreto como un nombre. En algunos casos, incluso, el nombre debe ser seleccionado de un conjunto o lista proporcionado en la propia pregunta (285), (286). En otros casos, sin embargo, la entidad que se debe identificar es más abstracta, un 'propósito' en (284) o una 'causa' en (291). Esto implica que la identificación sea también más imprecisa. En la sección de corpus analizaremos distintos valores posibles para la variable representada por *cuál*.

Como ocurre con *cómo*, las preguntas con *cuál* pueden ser informativas o abiertas dependiendo del valor concreto de la variable a la que apunta el interrogativo. Si la variable es concreta, la pregunta es informativa. Si la variable es abstracta e imprecisa, la pregunta es abierta.

QUÉ

El interrogativo *qué* apunta de modo general a una o más entidades no animadas:

(292) *¿Qué países colonizó Hong Kong?*
(293) *¿Qué envolvió el artista Christo?*
(294) *¿Qué país es el mayor productor de diamantes?*

El valor concreto de la entidad no animada puede estar más o menos determinado en la propia pregunta, dependiendo de si el interrogativo funciona como pronombre o como determinante.

Pronombre

La entidad está más indeterminada. Lo único que sabemos de ella es que es no animada y que cumple los requisitos recogidos en la pregunta:

(295) *¿Qué causa la gripe?* > entidad = «algo» que causa la gripe

Puede tratarse, por lo tanto, de una o más entidades, concretas (296) o abstractas (297):

(296) *¿Qué fabrica un luthier?*
(297) *¿Qué piensa Rouco Varela del matrimonio entre personas del mismo sexo?*

A mayor abstracción de la entidad (297), mayor imprecisión de la variable y, por lo tanto, mayor complejidad a la hora de buscar una respuesta.

El pronombre también puede tener un valor específico de tipo cuantitativo (de manera que se equipara a *cuánto*) cuando *qué* se construye con un verbo que implique medida (*costar, valer, pesar*, etc.), «con más frecuencia en el habla coloquial que en registros formales» (Real Academia Española, 2009, 22.13ñ).

(298) *¿Qué costó la construcción de la Ciudad de la Cultura?*

En estos casos, el valor de la variable es preciso y concreto (como ocurre con *cuánto*). La entidad a la que cuantifica *qué* se deduce del tipo de verbo presente en la pregunta (también de forma paralela a lo que ocurre con *cuánto*). Para Bosque (1984) la alternancia *qué*/*cuánto* es más frecuente si la pregunta hace referencia a conceptos como 'dinero' (*¿Qué/Cuánto invirtió, gastó, ...*), 'valor, peso o medida' (*¿Qué/Cuánto mide, pesa, ...?*), 'tiempo' (*¿Qué/Cuánto ha tardado, durado, ...?*), 'distancia' (*¿Qué/Cuánto recorrió, hay entre...?*), etc.

En determinadas construcciones copulativas, *qué* funcionando como pronombre presenta otro valor:

(299) *¿Qué es la ONU?*

El patrón sintáctico de estas construcciones es el siguiente:
frase interrogativa con *qué* + *ser* + frase nominal (= entidad)

Este tipo de construcciones son paralelas a las de *quién* + *ser* + nombre propio (*¿Quién es Pep Guardiola?*). Como en las construcciones con *quién*, en estas construcciones con *qué* se nos pide una descripción de la entidad. Ese valor de 'descripción' es también impreciso. En la tradición de los sistemas de BR generalmente se ha identificado este valor con el de 'definición'.

Determinante

Funcionando como determinante, *qué* tiene valores similares (excepto el de definición) a los que tiene como pronombre, con ciertos matices. La principal diferencia con los usos como pronombre es que la entidad a la que apunta el interrogativo está más especificada en la pregunta; retomando el ejemplo (300):

(300) *¿Qué causa la gripe?*
(301) *¿Qué virus causa la gripe?*

La diferencia entre (300) y (301) radica en la información que tenemos sobre la entidad a la que apunta *qué*: en (300) está indefinida, y, por los rasgos del interrogativo, solo sabemos que es una entidad no animada; en (301) se nos proporciona la clase a la que pertenece la entidad, *virus*, de manera que sabemos que aquello por lo que se nos está preguntando es un tipo de virus (en concreto, el que *causa la gripe*).

Al igual que como pronombre, como determinante *qué* apunta a una (302), (303) o varias entidades (304) no animadas, que pueden ser concretas (302) o abstractas (303), cuya clase está especificada en la pregunta:

(302) *¿Qué* empresa *creó Ruiz Mateos en primer lugar?* > entidad = empresa
(303) *¿Qué* explicación *ha dado Cospedal a las actuales políticas del gobierno?* > entidad = explicación
(304) *¿Qué* países *forman el Consejo de Cooperación del Golfo?* > entidad = países

Como ocurre siempre, a más abstracción de la entidad (303), más probabilidades de imprecisión en la determinación de cuál es la respuesta correcta.

Como determinante, *qué* también puede presentar el valor semántico de 'tipo o clase'. En las preguntas con este valor, aquello por lo que se pregunta no es una entidad concreta, sino un 'tipo o clase' de entidad:

(305) *¿Qué cerveza se consume más en Bélgica?* > entidad = tipo de cerveza (*la Pilsen*)
(306) *¿Qué libros prefería Don Quijote?* > entidad = tipo de libro (*los libros de caballerías*)

La interpretación de 'tipo o clase' se da con los sustantivos contables solo en plural. Comparemos (307) y (308):

(307) *¿Qué libros prefería Don Quijote?*
(307a) *Los de caballerías.*

(308) *¿Qué libro prefería Don Quijote?*
(308a) *El Amadís.*

Con sustantivos que pueden ser contables o no contables en singular, se puede dar la lectura de 'tipo o clase' en singular (309) y en plural (310):

(309) *¿Qué cerveza se consume más en Bélgica?*
(309a) *La de trigo.*

(310) *¿Qué cervezas se consumen más en Bélgica?*
(310a) *Las de trigo.*

El problema de este valor es que no es posible distinguirlo del de 'entidad':
(311) *¿Qué cerveza se consume más en Bélgica?*
(311a) *La de trigo.* (Tipo de cerveza)
(311b) *La Hoegaarden.* (Marca de cerveza de trigo)

(312) *¿Qué libros prefería Don Quijote?*
(312a) *Los de caballerías.* (Tipo de libro)
(312b) *El Amadís y Tirante el Blanco.* (Libros concretos)

De esta manera, estas construcciones son siempre ambiguas y solo es posible aclarar su significado acudiendo al usuario que pregunta o bien preestableciendo para ellas un valor de antemano.

En construcciones copulativas, caben también los dos valores que hemos visto para el determinante, 'entidad' (313), (314) y 'tipo/clase' (315):

(313) *¿Qué país es el mayor productor de diamantes?* > entidad = país
(314) *¿Qué galaxia es la más cercana a la vía láctea?* > entidad = galaxia
(315) *¿Qué libros eran los favoritos de Don Quijote?* > entidad = tipo de libro / libros

El determinante también acepta el valor de cuantificación cuando determina a ciertos nombres no contables en singular (Real Academia Española, 2009, 22.13m) como *altura*, *peso*, *profundidad,* etc.:

(316) *¿Qué altura tiene el Empire State?*

Esta interpretación cuantitiva también se da en el habla coloquial (Real Academia Española, 2009, 22.13s) con algunos sustantivos en plural, sobre todo con *años* (317), pero no solo (318).

(317) *¿Qué años tenía Colón cuando murió?*
(318) *¿Qué metros cuadrados tiene el Santiago Bernabéu?*

A modo de resumen, los valores posibles para *qué* son los siguientes:

1) **Una o más entidades no animadas, concretas o abstractas**: suele demandarse el nombre en entidades concretas; si es una entidad más abstracta, lo que se demanda es más impreciso.

1.1) No se da en la pregunta el tipo o clase de la entidad (*qué* pronombre).
(319) *¿Qué causa los tifones?*

1.2) Dando en la pregunta el tipo o clase de la entidad (*qué* determinante).
(321) *¿Qué fenómeno natural causa los tifones?*
(322) *¿Qué países componen la UE?*

2) **Tipo o clase del sustantivo que se nos da en la pregunta** (*qué* determinante): difícil de discernir del valor 1:
(323) *¿Qué cerveza gusta más en Praga? > la Pilsen* (tipo) *vs. la Grimbergen* (cerveza concreta).

(324) *¿Qué libros odia Umberto Eco? > la novela rosa* (tipo) *vs. 100 años de soledad y La Regenta* (libros concretos).

3) **Cuantificación de una variable:**
3.1) Sin especificar la variable cuantificada (*qué* pronombre):
(325) *¿Qué mide el estadio Bernabéu?*

3.2) Especificando la variable cuantificada (*qué* determinante):
(326) *¿Qué altura tiene el Empire State?*
(327) *¿Qué años tiene Mafalda?* (coloquial)

4) **Definición de una entidad**: solo en determinadas construcciones copulativas.

(328) *¿Qué es la Commonwealth?*

Como ocurría con otros interrogativos, las preguntas con *qué* pueden responderse con una descripción en lugar de con la entidad por la que se pregunta:

(329) *¿Qué país es el mayor productor de diamantes?*
(329a) *Un país que limita al oeste con Namibia.*

Respuestas como (329a) serían también respuestas indirectas.

Determinación del valor semántico de la variable

Como hemos visto, el interrogativo *qué* presenta diversos valores.

El principal problema a la hora de determinar la variable presente en las preguntas con *qué* lo encontramos con las entidades de tipo abstracto:

(330) *¿Qué causa los terremotos?*
(331) *¿Qué provoca la lluvia?*
(332) *¿Qué dijo Cospedal en su última comparecencia?*
(333) *¿Qué opinión tiene Zapatero de la crisis?*

Estas variables pueden corresponderse con procesos, opiniones, juicios, descripciones, etc., de manera que son inherentemente inespecíficas y pueden generar problemas a la hora de recuperar una respuesta válida. En los casos en los que el interrogativo funciona como determinante.

El significado del verbo y sus argumentos presentes en la pregunta también nos pueden dar pistas para establecer si se está preguntando por una entidad concreta o abstracta:

(334) *¿Qué alimentos consumen los Tarahumaras?*
Entidad = alimento (+concreta)
Otra información: «alimento» que consumen los Tarahumaras (información semántica contenida en el significado del verbo *consumir* y su sujeto *los Tarahumaras*).

Cuando el interrogativo funciona como pronombre, solo tenemos la información contenida en el verbo más sus argumentos para determinar por qué tipo de entidad se está preguntando.

(335) *¿Qué comen los zorros?*
Entidad = «algo»
Otra información: «algo» que comen los zorros (información semántica contenida en el significado del verbo *comer* + sujeto *los zorros*).

Las preguntas con *qué* también pueden ser informativas o abiertas. Cuando la entidad es concreta (cuantificación, etc.), la pregunta es de tipo informativo. Cuando la entidad es abstracta e imprecisa (descripción, opinión, etc.) la pregunta es abierta.

2.3.2.2 El significado de las partículas interrogativas: combinaciones con preposiciones

Hemos visto ya que todos los interrogativos excepto *por qué* pueden combinarse con preposiciones (sección 2.4.3.1). Exceptuando aquellos casos en los que el valor de la preposición es de tipo sintáctico (objeto indirecto), las

preposiciones añaden valor semántico a la frase interrogativa. Por esta razón, su análisis nos parece interesante para BR.

En SpQA tenemos en cuenta las siguientes preposiciones: *a*, *ante*, *bajo*, *con*, *contra*, *de*, *desde*, *durante*, *en*, *entre*, *hacia*, *hasta*, *para*, *por*, *según*, *sin*, *sobre*, *tras*.

A continuación detallamos sus significados básicos, y recuperamos la información sobre sus posibles combinaciones con cada uno de los interrogativos[87].

A: presenta distintos usos:

1) Elemento vacío de significado semántico que se usa en las siguientes construcciones sintácticas:

- Precediendo al complemento directo e indirecto.
  Se combina con: *qué*, *quién*, *cuál* y *cuánto*.
  (336) *¿A qué/cuál presidente eligió el pueblo americano en 1960?*
  (337) *¿A quién quiere vengar Hamlet en la obra de Shakespeare?*
  (338) *¿A cuántos asesinos atrapa Sherlock Holmes en la obra de Conan Doyle?*

- Precediendo a un complemento preposicional que se construye con infinitivo regido por un verbo que indica el comienzo, aprendizaje, intento, logro, mantenimiento o finalidad de la acción. (Real Academia Española, 2011).

  Se combina con *qué* y *cuál.*

  (339) *¿A qué/cuál cosa se aprende primero: a escribir o a leer?*

- Precediendo a un complemento preposicional que se construye con nombres y verbos de percepción y sensación, para precisar la sensación correspondiente. *Sabor a miel.* (Real Academia Española, 2011).

  Se combina con *qué* y *cuál.*

  (340) *¿A qué/cuál cosa huele el azufre?*

- Precediendo al complemento nominal o verbal que es régimen de ciertos verbos. *Condenar a muerte. Jugar a las cartas.* (Real Academia Española, 2011).

87 Somos conscientes de que, al menos a primera vista, algunas de las construcciones que se presentan a continuación pueden parecer poco esperables en un entorno de BR. No obstante, como se ha hecho en todo el capítulo, nos inclinamos hacia una descripción de las posibilidades lingüísticas de las preguntas lo más amplia y exhaustiva posible. En el capítulo 4, detallaremos cómo esta información se orienta al área de BR en la descripción de la gramática sobre la que se construye SpQA.

La combinación con interrogativos depende del tipo de entidad semántica que es régimen del verbo. *Qué* y *cuál* funcionan en general para casi cualquier tipo de entidad:

(341) *¿A qué/cuál deporte juega Federer?*

Cuando la entidad es de tipo cuantitativo, también se combina con *cómo*[88] y *cuánto*:

(342) *¿A qué/cuál precio se vende el Porsche más barato?*

(343) *¿A cuánto/cómo se venden hoy las acciones de Endesa?*

- Precediendo al complemento de algunos adjetivos. *Suave al tacto. Propenso a las enfermedades.*

  Se combina con *qué, cuál* y *cuánto*.

  (344) *¿A qué/cuáles/cuántas enfermedades es propenso un alcohólico?*

2) Indica 'destino' con verbos de movimiento.
Se combina con: *dónde*, *cuál*, *quién*, *qué*, *cuánto*:

(345) *¿A dónde se dirigía Colón?*
(346) *¿A qué/cuál lugar se dirigían los exploradores Speke y Grant?*
(347) *¿A quién se debe dirigir una reclamación sobre las autopistas gallegas ?*
(348) *¿A qué lugar se dirigía?*
(349) *¿A cuántas ciudades ha viajado?*

ANTE: su significado básico es 'delante de'.
Se combina con los siguientes interrogativos: *cuál*, *qué*, *quién*, *cuánto.*

(350) *¿Ante qué/cuáles reyes se postró Colón?*
(351) *¿Ante quién perdió las elecciones Rajoy?*
(352) *¿Ante cuántas personas tocó en su último concierto Bob Dylan?*

BAJO: su significado básico es 'debajo de' o 'al amparo de'.
Se combina con *cuál*, *qué*, *cuánto*.

(353) *¿Bajo qué está Eloísa en la obra de Jardiel Ponciela?*
(354) *¿Bajo cuál régimen ocurrieron las masacres en Serbia?*
(355) *¿Bajo cuántos dictaduras ha vivido Nicaragua?*

CON: medio, modo o instrumento que sirve para hacer algo. (Real Academia Española, 2011).

88 Uso coloquial.

Se combina con *cuál*, *qué*, *quién*, *cuánto*.
(356) ¿*Con qué se hacían las pinturas rupestres?*
(357) ¿*Con cuáles/cuántos apoyos cuenta Merkel en la UE?*
(358) ¿*Con quién se juntó Aznar en las Azores?*

CONTRA: su significado general es el de 'oposición'. (Real Academia Española, 2011).
Se combina con *cuál*, *qué*, *quién*, *cuánto*.

(359) ¿*Contra qué/cuál cosa chocó el Titanic?*
(360) ¿*Contra quién perdió la última Copa del Rey el Real Madrid?*
(361) ¿*Contra cuántos pueblos luchó Alejandro Magno?*

DE: puede aportar los siguientes significados:

- 'Posesión' o 'pertenencia' con el verbo *ser*.
  Se combina con *qué*, *quién*, *cuál*, *cuánto*.
  (362) ¿*De qué/cuál compositor es el Vals del Danubio Azul?*
  (363) ¿*De quién es La Macarena?*
  (364) ¿*De cuántos herederos es el castillo de Soutomaior?*

- 'Origen de algo'.
  Se combina con *dónde*, *cuál*, *qué*, *cuánto*, *quién*.
  (365) ¿*De dónde viene la ensaimada?*
  (366) ¿*De qué/cuál destino partió Marco Polo en su primer viaje?*
  (367) ¿*De cuántos puertos distintos partió Magallanes en su vida?*
  (368) ¿*De quién surgió la idea del cinematógrafo?*

- 'Materia de la que está hecho algo'.
  Se combina con *qué*, *cuál*, *cuánto*.
  (369) ¿*De qué está hecho el chicle?*
  (370) ¿*De cuál material son los muros de Alcatraz?*
  (371) ¿*De cuántos ingredientes se compone la sangría?*

DESDE: su principal significado es el de 'origen', 'punto de partida'.
Se combina con *cuándo*, *dónde*, *cuál*, *qué*, *cuánto*.

(372) ¿*Desde cuándo se usa la rueda?*
(373) ¿*Desde dónde se empiezan a contar las páginas para hacer el índice?*
(374) ¿*Desde qué/cuál lugar partió Vasco de Gama?*
(375) ¿*Desde cuánto comienza un coeficiente intelectual a ser considerado alto?*

DURANTE: denota simultaneidad de un acontecimiento con otro. (Real Academia Española, 2011).
Se combina con *cuál*, *qué*, *cuánto*.

(376) *¿Durante cuál semana de embarazo la hormona gonadotropina está más elevada?*
(377) *¿Durante qué período hay más horas de sol en Santiago?*
(378) *¿Durante cuánto tiempo están obligados los servicios de asistencia técnica a proporcionar piezas de repuesto?*

EN: su significado básico es el de 'localización'.
Se combina con *dónde*, *cuál*, *qué*, *quién*, *cuánto*.
(379) *¿En dónde se fabrican los Ferrari?*
(380) *¿En qué/cuál país el chocolate está más rico?*
(381) *¿En quién confió el electorado en las últimas elecciones generales de España?*
(382) *¿En cuántos mares estuvo Francis Drake?*

ENTRE: tiene los siguientes significados:

- 'Localización en medio de dos lugares'.
  Se combina con: *cuál*, *quién*, *qué*, *cuánto*.
  (383) *¿Entre qué/cuáles países se sitúa España en la UE en cuanto a resultados académicos en secundaria?*
  (384) *¿Entre quiénes se sientan los novios en la mesa de una boda?*

  (385) *¿Entre cuántos aspirantes fueron elegidos los seleccionados para participar en la última edición de Operación Triunfo?*

- 'Cooperación'.
  Se combina con: *cuál*, *quién*, *qué*, *cuánto*.
  (386) *¿Entre qué/cuáles mentes se gestó la Revolución de Octubre?*
  (387) *¿Entre quiénes se urdió el asesinato de John Lenon?*
  (388) *¿Entre cuántos planificaron el secuestro de Delorean?*

HACIA: su significado básico es el de 'dirección a la que apunta un movimiento'.
Se combina con *dónde*, *cuál*, *quién*, *qué*, *cuánto*.

(389) *¿Hacia dónde se dirigía Alejandro Magno en sus conquistas?*
(390) *¿Hacia qué/cuál lugar viajaban los portugueses bordeando África?*
(391) *¿Hacia quién se dirigió Colón como última opción para financiar sus viajes?*
(392) *¿Hacia cuántos países diferentes se encaminó Marco Polo?*

HASTA: denota el término de tiempo, lugares, acciones o cantidades. (Real Academia Española, 2011).
Se combina con *cuándo*, *dónde*, *cuál*, *quién*, *qué*, *cuánto*.

(393) *¿Hasta cuándo esperó Penélope?*
(394) *¿Hasta dónde se extiende Rusia?*

(395) *¿Hasta qué/cuál altura puede volar un globo?*
(396) *¿Hasta cuánto puede costar un coche deportivo?*
(397) *¿Hasta quién ha acudido por desesperación?*

PARA: denota 'fin' o 'término'.
Se combina con *cuándo*, *dónde*, *cuál*, *qué*, *quién*, *cuánto*.

(398) *¿Para cuándo saldrá El Hobbit II?*
(399) *¿Para dónde trasladaron al preso Antonio García Vidriel?*
(400) *¿Para qué/cuál asunto viajó Rajoy a Polonia tras el primer rescate?*
(401) *¿Para quién se construyó el Taj Mahal?*
(402) *¿Para cuántos países ha aprobado rescates el FMI?*

POR: tiene los siguientes significados:

- 'Localización física'.
  Se combina con *dónde*, *cuál*, *qué*, *cuánto.*
  (403) *¿Por dónde pasa el Ebro?*
  (404) *¿Por qué/cuál país discurre el Nilo?*
  (405) *¿Por cuántos países pasa el Danubio?*

- 'Localización temporal'.
  Se combina con *qué* y *cuál*.
  (406) *¿Por qué/cuál época anidan las golondrinas?*

- 'En calidad de'.
  Se combina con *qué* y *cuál.*
  (407) *¿Por qué/cuál cosa tomaron a los españoles los nativos americanos?*

- 'Causa'.
  Se combina con *cuál* y *qué*.
  (408) *¿Por qué/cuál causa la Tierra es redonda?*

- 'Medio o modo de ejecutar algo'. *Por fuerza. Por señas* (Real Academia Española, 2011).
  Se combina con *qué* y *cuál*.
  (409) *¿Por qué/cuál método consiguió hacerse rico Amancio Ortega?*

- 'Precio'.
  Se combina con *cuál*, *qué*, *cuánto*.
  (410) *¿Por qué/cuál precio se vendió el cuadro más barato de Van Gogh?*
  (411) *¿Por cuánto vendió Picasso su primera pintura?*

- 'A favor o en defensa de alguien o de algo'. *Por él daré la vida.* (Real Academia Española, 2011).
  Se combina con *quién*, *qué*, *cuál*, *cuánto*.
  (412) *¿Por quién se sacrifó Jesús?*

(413) *¿Por qué/cuál país luchó Lord Byron?*
(414) *¿Por cuántos países peleó el Che?*

SEGÚN: expresa 'conforme a', 'con arreglo a'.
Se combina con *cuál*, *qué*, *quién*, *cuánto*.

(415) *¿Según qué/cuál pensador griego la Tierra giraba alrededor del Sol?*
(416) *¿Según quién el sol tenía manchas?*
(417) *¿Según cuántos observadores un OVNI chocó en Roswell en 1947?*

SIN: denota 'carencia' o 'falta de algo'. (Real Academia Española, 2011).
Se combina con *cuál*, *qué*, *quién*, *cuánto.*
(418) *¿Sin qué/cuál miembro se quedó Take That en 2010?*
(419) *¿Sin quién no aguantaron juntos los Beatles?*
(420) *¿Sin cuántos caballeros de la Mesa Redonda se quedó el Rey Arturo tras todas sus batallas?*

SOBRE: tiene los siguientes significados:

- 'Encima'.
  Se combina con *cuál*, *qué*, *quién*, *cuánto*.
  (421) *¿Sobre qué/cuál placa se sitúa el continente europeo?*
  (422) *¿Sobre quién recayó la responsabilidad de sacara adelante a Grecia?*
  (423) *¿Sobre cuántos países ha caído alguna vez lluvia ácida?*

- 'Acerca de'.
  Se combina con *cuál*, *qué*, *quién*, *cuánto*.
  (424) *¿Sobre qué/cuáles temas trata la República de Platón?*
  (425) *¿Sobre quién habla la obra de teatro Citizen?*
  (426) *¿Sobre cuántos temas distintos escribió Erasmo?*

TRAS: tiene el significado general de 'detrás de, después de' con dos sentidos:

- Físico.
  Se combina con *cuál*, *qué*, *quién*, *cuánto*.
  (427) *¿Tras qué/cuál actor se sitúa Tom Cruise en la lista de los actores más guapos de Fotogramas?*
  (428) *¿Tras quién se escondía Alonso Quijano?*
  (429) *¿Tras cuántos se escondió de su culpa Hitler?*

- Temporal.
  Se combina con *cuál*, *qué*, *quién*, *cuánto*.
  (430) *¿Tras qué/cuál mes viene marzo?*
  (431) *¿Tras quién salió al escenario Muse en su última actuación?*
  (432) *¿Tras cuántos minutos sale a la superficie una gaviota que se sumerge a pescar?*

En la siguiente tabla sintetizamos todos los valores de cada una de las preposiciones, así como su uso con los distintos interrogativos:

	Cuándo	**Dónde**	**Cómo**	**Cuál**	**Quién**	**Qué**	**Cuánto**
A (sintaxis) -objeto directo/indirecto				X	X	X	X
A (sintaxis) -infinitivo regido				X		X	
A (sintaxis) -complementos de nombres y verbos de percepción				X		X	
A (sintaxis) -complemento nominal o verbal				X		X	X
A (sintaxis) -complemento de adjetivos				X		X	X
A +destino		X		X	X	X	X
ANTE +delante de				X	X	X	X
BAJO +debajo de				X		X	X
CON +medio, modo, instrumento				X	X	X	X
CONTRA +oposición				X	X	X	X
DE +posesión				X	X	X	X
DE +origen		X		X	X	X	X
DE +materia				X		X	X
DESDE +origen	X	X		X		X	X
DURANTE +duración				X		X	X

	Cuándo	**Dónde**	**Cómo**	**Cuál**	**Quién**	**Qué**	**Cuánto**
EN +localización		X		X	X	X	X
ENTRE +localización				X	X	X	X
ENTRE +cooperación				X	X	X	X
HACIA +dirección		X		X	X	X	X
HASTA +término	X	X		X	X	X	X
PARA +fin	X	X		X	X	X	X
POR +localización física		X		X	X	X	X
POR +localización temporal				X		X	
POR +en calidad de alguien/algo				X		X	
POR +causa				X		X	
POR +medio o modo de ejecutar algo				X		X	
POR +precio				X		X	X
POR +a favor de alguien/algo				X	X	X	X
SEGÚN +conforme				X	X	X	X
SIN +carencia				X	X	X	X
SOBRE +encima				X	X	X	X
SOBRE +acerca de				X	X	X	X

	Cuándo	**Dónde**	**Cómo**	**Cuál**	**Quién**	**Qué**	**Cuánto**
TRAS +detrás físico				X	X	X	X
TRAS +detrás temporal				X	X	X	X

Tabla 2: Combinaciones de preposiciones e interrogativo.

A partir de los párrafos anteriores y de la tabla, podemos establecer una clasificación de las preposiciones en *clusters* semánticos. Diferenciamos cuatro grandes grupos:

1) Preposiciones ligadas tanto al concepto de **tiempo** como al de **espacio**: *de, desde, a, hacia, hasta, para, por, en, entre, tras*.
2) Preposiciones ligadas solo al concepto de **tiempo**: *durante*.
3) Preposiciones ligadas solo al concepto de **espacio**: *ante, bajo, sobre*.
4) Preposiciones ligadas a otros conceptos semánticos (no relacionadas con tiempo y espacio):
   - 'Medio': *con*.
   - 'Oposición': *contra*.
   - 'Posesión': *de*.
   - 'Materia': *de*.
   - Cooperación': *entre*.
   - 'Causa': *por.*
   - 'En calidad de': *por.*
   - Valor cuantitativo: *por*, *a*.
   - 'Conforme a': *según*.
   - 'Carencia': *sin*.

De estos campos semánticos, los más interesantes para BR (al menos para BR de tipo factual) son 'espacio', 'tiempo' y 'cantidad', pues son los más ligados al ámbito fáctico. De hecho, vimos en el capítulo 1 que el reconocimiento y marcado de relaciones espacio-temporales es una de las técnicas que se utiliza en BR. Por otro lado, 'tiempo', 'espacio', 'cantidad' y 'causa' son también los únicos valores de los citados que pueden ser expresados directamente por los interrogativos (*dónde*, *cuándo, cuánto y por qué*, respectivamente). Como veremos en el capítulo 4, este hecho nos será útil para tratar la paráfrasis de la frase interrogativa en SpQA.

2.3.2.3 Conclusiones sobre la semántica de los interrogativos

En la siguiente tabla resumimos los valores semánticos señalados para los interrogativos:

	VALOR 1	VALOR 2	VALOR 3	VALOR 4	TOTAL
CUÁNDO	temporal	hipotético			2
DÓNDE	locativo				1
CÓMO	modal	causal			2
POR QUÉ	causal	sugerencia (+no)			2
QUIÉN	identificación	descripción			2
CUÁL	x	x	x	x	x
CUÁNTO	cantidad				1
QUÉ	entidad	tipo entidad	cantidad	definición	4

Tabla 3: Valores semánticos de los interrogativos.

Como vemos, algunos interrogativos tienen un solo valor (*cuánto*, *dónde*), mientras que otros presentan varios valores posibles (*qué*, *cómo*, etc.). El caso de *cuál* es el más particular: al ser un elemento deíctico, *cuál* puede tener cualquier valor (*n* valores). Observemos los siguientes ejemplos, en los que *cuál* reproduce el valor de cualquier interrogativo:

1) *Cuándo*
¿Cuándo nació Ghandi?
¿Cuál es la fecha del nacimiento de Ghandi?

2) *Dónde*
¿Dónde nació Ghandi?
¿Cuál es el lugar en el que nació Ghandi?

3) *Cómo*
¿Cómo se hace un risotto?
¿Cuál es el modo de hacer un risotto?

4) *Por qué*
¿Por qué llueve?
¿Cuál es la causa de que llueva?
5) *Quién*
¿Quién pintó La Gioconda?
¿Cuál es el autor de la Gioconda?

6) *Cuánto*
¿Cuánto mide la Torre Agbar?
¿Cuál es la altura de la Torre Agbar?

7) *Qué*
¿Qué provoca un tsunami?
¿Cuál es la causa que provoca un tsunami?

La paráfrasis con *cuál*, de hecho, implica la especificación de la variable que corresponde a la incógnita en un sustantivo concreto.

Hay que tener en cuenta que, a los valores básicos recogidos en la tabla 3 se suman los valores que pueden aportar las preposiciones con las que se combina cada interrogativo (cf. sección anterior).

Si atendemos a la mayor o menor concreción de las variables, obtenemos la siguiente tabla:

	Nº de valores precisos	**Nº de valores imprecisos (granularidad)**	**Nº de valores imprecisos (abstractos)**
CUÁNDO	0	1 (temporal)	1 (hipotético)
DÓNDE	0	1 (locativo)	0
CÓMO		0	2 (modal, causal)
POR QUÉ	0	0	2 (causal, sugerencia)
QUIÉN	1 (identificación)	0	1 (descripción)
CUÁNTO	1 (cantidad)	0	0
CUÁL	1 (entidad)	0	1 (entidad)
QUÉ	3 (entidad, cantidad, tipo de entidad)	0	2 (entidad, definición)
TOTAL	6	2	9

Tabla 4: Tipo de valor semántico (preciso *vs.* impreciso).

La tabla muestra que hay un alto índice de imprecisión en los posibles valores semánticos de los interrogativos (9 posibles valores imprecisos *vs.* 6 concretos).

Si estableciéramos una escala de complejidad en la determinación de la variable para cada interrogativo, tendríamos *cuánto* en un extremo con un solo valor

concreto, y *qué* o *cuál* en el otro, con muchos valores semánticos posibles y muchos de ellos de tipo abstracto.

Por otro lado, el análisis de los interrogativos muestra que hay una serie de partículas interrogativas con un significado inherente bastante definido y constante, y otra serie de partículas interrogativas con un significado menos concreto que permiten apuntar a más tipos de entidades distintas. Al primer grupo pertenecerían los interrogativos adverbiales: *cuándo*, *dónde*, *por qué*, *cómo* y *cuánto* y, por otra parte, *quién*. Al segundo, *cuál* y *qué*. Este hecho se refleja en la tabla 4, donde observamos que *cuál* y *qué* acumulan el mayor número de valores semánticos posibles. Por lo tanto, podemos establecer una clasificación de los interrogativos en dos grupos:

- Interrogativos «llenos» o «transparentes»: son aquellos que tienen un significado más definido.
  A este grupo pertenecen: *cuándo*, *dónde*, *por qué*, *cómo*, *cuánto* y *quién*.

- Interrogativos «vacíos» u «opacos»: son aquellos que tienen un significado menos definido y que, por tanto, pueden potencialmente apuntar a cualquier entidad.
  Pertenecen a este grupo: *cuál* y *qué*.

Los interrogativos del primer grupo suelen funcionar como adverbios o pronombres. El carácter adverbial limita las posibilidades combinatorias del interrogativo: no puede modificar sustantivos, ni sustituir a otros elementos y solo puede ser modificado por preposiciones ligadas al concepto espacial, temporal o de otro tipo que el adverbio expresa (cf. *supra*). El caso de *quién* y *cuánto* es algo distinto. *Quién* solo funciona como pronombre con el rasgo +*humano*, por lo que tampoco puede modificar a sustantivos y las preposiciones que lo pueden modificar tienen que ser compatibles con ese rasgo semántico de +*humano*. *Cuánto* tiene el rasgo semántico básico de +*cantidad*. Por eso, aunque sí puede funcionar como determinante, su significado siempre está restringido y limitado al concepto de cantidad.

Al contrario, los interrogativos del segundo grupo pueden funcionar como pronombres o como determinantes. Como pronombre, *qué* está más restringido semánticamente, ya que solo puede apuntar a entidades no animadas. *Cuál*, sin embargo, puede apuntar como hemos visto a cualquier tipo de entidad. Como determinantes, las posibilidades semánticas de estas partículas se multiplican: las dos pueden apuntar a cualquier tipo de entidad:

+animada: *qué-cuál persona/animal*
+inanimada: *qué-cuál objeto*
+concreta: *qué-cuál mesa*
+abstracta: *qué-cuál cualidad*

De esta manera, el segundo grupo tiene la potencialidad de desempeñar las funciones del primero (es decir, de adquirir valores adverbiales, por ejemplo), funcionando como determinantes y combinados con preposiciones, generalmente:

+*entidad principal del discurso:* qué-cuál cineasta/edificio/comida
+*entidad «marco» del discurso:* en qué-cuál año; desde qué-cuál ciudad

Es por eso que el segundo grupo permite una mayor cantidad de paráfrasis, mientras que el primero es mucho más limitado en este aspecto. Volveremos sobre estos aspectos en el capítulo 4.

2.4 Aspectos lingüísticos relevantes para la formalización de las preguntas en la gramática de SpQA

En las secciones anteriores hemos descrito una serie de fenómenos lingüísticos relacionados con las preguntas y la relación pregunta-respuesta. Nuestra descripción teórica ha intentado ser exhaustiva en lo que se refiere al funcionamiento de las preguntas y la relación pregunta-respuesta desde una perspectiva general.

No obstante, como el objetivo principal de este trabajo es el análisis de preguntas en un contexto de BR, de los fenómenos descritos nos interesan especialmente aquellos que son relevantes para la formalización de preguntas en una gramática orientada a tareas de BR. Por esta razón, presentaremos a continuación una selección de aquellos fenómenos que son más importantes desde esta perspectiva. Como veremos en el capítulo 4, algunos de estos fenómenos determinan ciertas características de SpQA.

Al igual que hasta ahora, estructuraremos los fenómenos en dos bloques: fenómenos gramaticales y fenómenos semántico pragmáticos.

2.4.1 Rasgos gramaticales

Negación y foco

Las interrogativas pueden contener negación. Esa negación puede ser de tipo interno o externo. Que la negación sea de un tipo u otro tiene consecuencias en la interpretación semántica de la preguntas. De este modo, (433) y (434):

(433) *¿Bush no le dijo la verdad a los americanos?*
(434) *¿Quién no ha ido a la ópera alguna vez?*

pueden interpretarse de dos maneras:

(433a) *¿Bush no le dijo la verdad a los americanos?* > negación interna
(433b) *Bush le dijo la verdad a los americanos, ¿no?* > negación externa

(434a) *Alguien no ha ido a la ópera alguna vez, ¿quién?*
(434b) *Todo el mundo ha ido a la ópera, ¿verdad?*

(Ejemplos adaptados de Escandell, 1999)

Atendiendo solo a la información gramatical, no existe modo alguno de distinguir entre ambos tipos de negación a menos que aparezcan términos de polaridad negativa (negación interna) o positiva (negación externa).

Debido a su significado, la mayoría de los casos de interrogativas con negación externa son «peticiones de confirmación de una presuposición positiva» (Escandell, 1999, p. 3960). Por esta razón, lo esperable en un entorno de BR es que la mayoría de los casos de negación sean de tipo interno y no externo. Como veremos en la sección de análisis de corpus, nuestros datos apoyan esta intuición.

No obstante, aunque extrañas, creemos que sí serían posibles preguntas con negación externa en un entorno de BR:

(435) *¿No descubrió Colón América?*

En estos casos, el vacío informativo del que pregunta se limita a la confirmación de una presuposición que él considera, ya de antemano, cierta.

Orden de palabras en totales y parciales

Parece existir un orden de constituyentes prototípico tanto en las interrogativas totales como en las parciales.

En las parciales el orden prototípico es: interrogativo + verbo + complementos.

(436) *¿Cuándo llegó el hombre a la Luna por primera vez?*

En el estudio de corpus de Gayo (2010), este orden con el interrogativo encabezando la estructura es claramente el preferido para las interrogativas parciales.

Además de este orden, existen otros que parecen más marginales:

- En el caso de interrogativos funcionando como adjuntos: es posible otra ordenación con un constituyente entre el interrogativo y el verbo:

(437) *¿Desde cuándo Portugal es una república?*

Esta ordenación se documenta en Gayo (2010), donde se denomina «anteposición tipo B».

- En Gayo (2010) se documenta también la «anteposición tipo A»: en esta ordenación se coloca un constituyente antes del interrogativo:

(438) *¿Y Rajoy qué opina de Bárcenas?*

(Ejemplo adaptado de Gayo, 2010)

- En el español del Caribe se da además otro orden en el que se coloca el sujeto entre el interrogativo y el verbo:

  (439) *¿Qué tú quieres?*

  En las totales el orden prototípico es: verbo + sujeto+ complementos.

  (440) *¿Llegará el hombre a Marte algún día?*

Según Escandell (1999), la anteposición del sujeto al verbo en las totales provoca la interpretación de la interrogativa total como una estructura más compleja en la que la interrogación opera sobre una proposición ya cerrada. Esto implica que la proposición ya esté presente en el contexto anterior de alguna forma, por lo que este orden no sería esperable, al menos en principio, en un contexto de BR.

(441) *¿Pedro viene mañana?*
¿[Proposición]?

(Representación tomada de Escandell, 1999)

En el capítulo 3 investigaremos los distintos órdenes posibles en corpus de preguntas, con el fin de arrojar más luz sobre esta y otras cuestiones..

Uso de partículas interrogativas

Como hemos visto, el papel de las partículas interrogativas es crucial en el funcionamiento de las interrogativas parciales, tanto a nivel gramatical como semántico. Es por ello por lo que consideramos que todos los aspectos tratados en la sección relativa al uso de las partículas interrogativas (características morfosintácticas, factores que influyen en la selección de las partículas interrogativas, etc.), son especialmente relevantes para el análisis de las preguntas.

Interrogativas y subordinación: asociación a distancia

Este fenómeno también es relevante en la interpretación de las preguntas. Como vimos en la sección 2.4.5.2, cuando la pregunta contiene una subordinada (generalmente, con un verbo puente, cf. 2.4.5.2), puede no estar claro con qué verbo se asocia el interrogativo (si con el verbo de la principal o con el verbo de la subordinada):

(442) *¿Quién cree que ganará? > quién – cree / quién – ganará*

Siempre y cuando tengamos dos o más verbos con los que el interrogativo se puede asociar, que haya ambigüedad o no depende, entre otras cosas, del tipo de interrogativo.

Verbo no finito

Las interrogativas parciales pueden contener un verbo no finito:

(445) *¿Dónde* comprar *una katana?*

2.4.2 Aspectos semánticos

Segundo postulado de Hamblin

> *Postulate 2. Knowing what counts as an answer is equivalent to knowing the question.*
> (Hamblin, 1958, pp. 162-163).

De los tres postulados de Hamblin, este nos parece el más relevante para el análisis y representación de las preguntas[89]. Como mencionábamos más arriba:

> El segundo postulado se refiere al significado de las preguntas. Lo que Hamblin nos dice es que determinar el significado de las preguntas consiste en saber qué cuenta como respuesta. Debemos tener en cuenta que «saber qué cuenta como respuesta» no es lo mismo que «saber cuál es la respuesta»: **lo que el principio de Hamblin sugiere es que el significado de la pregunta restringe el ámbito de las respuestas**.

Como hemos visto en la sección de análisis semántico de los interrogativos[90] (cf. sección 3.2.2.1), la forma de la pregunta determina (restringe) las posibles respuestas, en unos casos de forma precisa y clara, en otros de forma más imprecisa. Formalizar este tipo de cuestiones es crucial para el análisis de preguntas en un sistema de BR.

Foco y significado de las interrogativas: diferencia entre totales y parciales

Como hemos visto tanto en las interrogativas totales como en las parciales el significado de la oración depende en gran medida de cuál es el dominio del operador interrogativo, lo que a su vez está directamente relacionado con cuál es el constituyente o los constituyentes marcado(s) como foco.

Lo general parece ser que el foco sea la frase interrogativa en las parciales (con el resto de la información formando la presuposición) y toda la oración en las totales. Esto será así siempre y cuando en la interrogativa no haya otros elementos

89 Nuestra posición respecto al primer postulado de Hamblin la recogeremos más adelante, cuando tratemos la cuestión de «respuesta tipo constituyente *vs.* respuesta tipo oración».

90 En nuestro análisis semántico de las partículas interrogativas ya hemos aplicado este principio de Hamblin, al establecer cierta equivalencia entre el significado de una pregunta y sus posibles respuestas correctas.

sensibles al foco, tales como partículas de polaridad negativa y/o cuantificadores. Por tanto, en nuestro análisis, el significado esperable para (446) y (447) sería el siguiente:

(446) *¿Quién descubrió la penicilina? > Alguien descubrió la penicilina, ¿quién?*
(447) *¿Es Messi un jugador del FC Barcelona? > Messi es un jugador del FC Barcelona, ¿sí? / ¿no?*

Este es el análisis que proponemos para parciales y totales en un entorno de BR, donde no se presupone interacción ni un contexto textual previo a las preguntas. Esta interpretación encaja además con la interpretación semántica que la mayoría de las aproximaciones que hemos visto dan para las parciales y las totales, que reproducimos aquí de nuevo:

TOTALES
(448) *Did Lee Harvey Oswald kill John F. Kennedy?*
(449) *Yes, Lee Harvey Oswald killed John F. Kennedy.*

(448a) ?kill' (lee_harvey_oswald' , john_f_kennedy')
(449a) kill' (lee_harvey_oswald' , john_f_kennedy')

PARCIALES
(450) *Who killed John F. Kennedy?*
(451) *Lee Harvey Oswald killed John F. Kennedy.*

(450a) ?x1 kill' (x1, john_f_kennedy')
(451a) kill' (lee_harvey_oswald' , john_f_kennedy')

(Tomado de Fliedner, 2007)

Respuesta oración *vs.* respuesta constituyente en las preguntas parciales

En relación a esta cuestión, seguimos la propuesta de Ingo Reich (Reich, 2003): desde una respuesta oracional subyacente a la parcial todo el material no focalizado (y por lo tanto especialmente todo el material que no corresponde a la frase interrogativa) puede ser elidido.

Si bien este aspecto está más relacionado con la extracción y selección de la respuesta, consideramos que también es un aspecto importante para el análisis de la pregunta, pues supone que se distinga claramente entre aquel material de la pregunta que corresponde a la frase interrogativa (material focalizado) y aquel que corresponde a la proposición (material no focalizado).

Preguntas abiertas *vs.* preguntas informativas

Esta diferenciación de Groenendijk y Stokhof (1997) nos parece

especialmente interesante para el análisis de preguntas en BR. En nuestro análisis de los interrogativos (cf. *supra*), hemos aplicado esta clasificación intentando discernir, en cada caso, qué interrogativos permitían cada uno de los tipos de preguntas.

Valores semánticos de los interrogativos

En referencia a este punto, toda la información semántica de la sección 3.2.2 nos parece relevante para el análisis de preguntas en SpQA.

Relación de factores lingüísticos relevantes para SpQA

Rasgos gramaticales

- Negación y foco.
- Orden de palabras en totales y parciales.
- Características de las partículas interrogativas.
- Interrogativas y subordinación: asociación a distancia
- Verbo no finito.

Rasgos semánticos

- 2º postulado de Hamblin.
- Foco y significado de las interrogativas.
- Respuesta-oración *vs.* respuesta-constituyente en las preguntas parciales.
- Preguntas abiertas *vs.* preguntas informativas.
- Valores semánticos generales de los interrogativos.

En el capítulo siguiente nos centraremos en el análisis en corpus de algunos de estos rasgos. Ya en el capítulo 4, retomaremos todas estas cuestiones junto a los datos obtenidos en el estudio de corpus en la descripción de SpQA.

2.5 Conclusiones generales del capítulo

Este capítulo se ha centrado en el estudio lingüístico exhaustivo de las preguntas y de la relación pregunta-respuesta desde una perspectiva general.

En la primera sección se ha definido qué es una pregunta, diferenciando este concepto de otros como petición de información u oración interrogativa. Para nosotros, una pregunta es una demanda de información con la forma de una oración interrogativa directa.

En la segunda sección se han caracterizado las oraciones interrogativas directas. Para ello, se han definido de modo general las oraciones interrogativas; se han diferenciado oraciones interrogativas directas de indirectas; se han presentado las características gramaticales que definen las oraciones interrogativas directas (curva entonativa, orden de constituyentes y partículas interrogativas); y se han presentado otros aspectos gramaticales relevantes para el funcionamiento de la oraciones interrogativas directas (foco, negación, etc.).

En la tercera sección nos hemos ocupado de los aspectos semánticos de las preguntas y la relación pregunta-respuesta. Con este fin, en primer lugar se han revisado las distintas teorías sobre la semántica de las preguntas, desde las perspectivas puramente formales hasta las perspectivas pragmático-semánticas. En segundo lugar, hemos profundizado en la semántica y pragmática de los sistemas de BR, y nos hemos ocupado del análisis semántico exhaustivo de las partículas interrogativas.

Finalmente, hemos retomado toda la información lingüística anterior desde el punto de vista de la formalización en la gramática de SpQA, seleccionando aquellos aspectos lingüísticos que consideramos más relevantes para el análisis de preguntas en un entorno de BR.

En el capítulo siguiente nos ocuparemos del estudio en corpus de algunos de estos aspectos lingüísticos.

Capítulo 3

Análisis de corpus

En la sección final del capítulo anterior señalamos una serie de aspectos lingüísticos como especialmente relevantes (al menos teóricamente) para el análisis de preguntas en BR.

En este capítulo presentaremos el análisis de algunos de estos aspectos lingüísticos documentados en corpus reales de preguntas. El objetivo de dicho análisis es determinar hasta qué punto las afirmaciones teóricas respecto a tales aspectos son completas y adecuadas. Los datos extraídos del análisis de corpus unidos a las descripciones teóricas del capítulo anterior nos aportarán una visión de conjunto más completa sobre el funcionamiento de las preguntas. Esto, a su vez, nos permitirá llevar a cabo una descripción de las preguntas en la gramática formal de SpQA (cf. capítulo 4) más precisa y realista.

3.1 Rasgos analizados

No todos los rasgos seleccionados al final del capítulo anterior se han analizado en corpus, bien porque su análisis no nos parecía relevante, bien porque este no era posible (ciertos aspectos semánticos). El análisis se ha centrado en aquellos aspectos (gramaticales y semánticos) sobre los cuales se ha considerado que el estudio en corpus podía aportar datos útiles a la formalización.

De los aspectos gramaticales señalados al final del capítulo anterior, se han estudiado en corpus los siguientes:

- **Orden de constituyentes**: el estudio en corpus aporta información sobre si los órdenes considerados «prototípicos» lo son en realidad, además de sobre la incidencia de los otros órdenes descritos.
  - Orden en las totales: incidencia del «orden prototípico»; incidencia del orden con el sujeto antepuesto.
  - Orden en las parciales: incidencia del «orden prototípico»; incidencia de otras ordenaciones.
- **Negación**: documentación del adverbio *no* en los corpus. Incidencia de negación externa e interna.

- **Partículas interrogativas**: el estudio da información sobre la incidencia de cada una de las partículas en los corpus.

- **Preposiciones**: el estudio aporta datos sobre las combinaciones de las preposiciones con cada partícula interrogativa.

Se han dejado fuera del estudio dos aspectos gramaticales: asociación a distancia y uso de verbo no finito en las preguntas. Esta decisión se debe a que se ha considerado que su estudio en corpus no aportaría datos relevantes para la formalización de preguntas en SpQA (ambos rasgos se han formalizado directamente en la gramática, cf. capítulo 4).

En cuanto a los aspectos semánticos, nos hemos ocupado de aquellos cuyo estudio en corpus era posible y que a la vez eran más relevantes para la formalización. Por esta razón, nos hemos centrado en la semántica de los interrogativos, concretamente, en el análisis de los posibles significados para cada una de las partículas interrogativas.

3.2 Metodología

3.2.1 Corpus utilizados

Para nuestro estudio hemos compilado tres corpus distintos. Dichos corpus contienen preguntas extraídas de diferentes fuentes. Los tres corpus son:

3.2.1.1 Trivial

Trivial es un corpus formado por preguntas tipo *quiz,* es decir, preguntas de conocimiento general sobre diversas áreas, como las que nos podemos encontrar en el juego *Trivial Pursuit*.

El corpus está constituido por 249 preguntas extraídas de la web[91], 22 totales (1) y 227 parciales (2), (3):

(1) *¿Fue Felipe V de España hijo de Felipe IV?*[92]
(2) *¿Con cuántos países limita Mónaco?*
(3) *¿Cuál era el oficio de los Siete Enanitos?*

3.2.1.2 Clef

Clef es un corpus formado por preguntas utilizadas en distintas ediciones de las conferencias CLEF: QA@CLEF 2004, QA@CLEF 2006 y QA@CLEF 2007.

91 249 preguntas de 251 casos extraídos el 08/05/12 de: http://platea.pntic.mec.es/jescuder/pregunta.htm

92 Todos los ejemplos de este capítulo pertenecen a alguno de los tres corpus.

En total, el corpus se compone de 586 preguntas parciales (no hay totales).

(4) *¿De qué está recubierto el continente antártico?*
(5) *¿Qué país se reincorporó a la UNESCO tras 38 años de ausencia?*

3.2.1.3 Wiki

Wiki es un corpus formado por preguntas reales de usuarios extraídas de un portal de preguntas y respuestas llamado Wikirespuestas[93]. El corpus contiene todas las preguntas disponibles en el portal en el momento de la extracción[94].

Sobre este conjunto de preguntas se realizó un procesado simple que consistió en la eliminación de duplicados y la corrección de la acentuación de las partículas interrogativas. El resultado final es un corpus de 169.319 estructuras, entre preguntas y peticiones de información.

Al ser un corpus no supervisado elaborado a partir de datos reales de usuarios, Wiki presenta una serie de problemas. En primer lugar, el corpus tiene graves problemas ortográficos. En segundo lugar, una gran parte de las estructuras utilizadas en el corpus no son preguntas (no son oraciones interrogativas), sino peticiones de información. Veamos algunos ejemplos:

(6) *100 centimetros*[95] *en metros?*
(7) *Mariposa en nahualtla?*
(8) *Buscar los productos alimenticios originarios de mexico que sean aportaciones para la gastronomia mundial?*

Los ejemplos (6) y (7) no tienen verbo (son frases nominales). En (8) tenemos una oración de infinitivo «descolgada».

Los errores ortográficos del corpus hacen imposible determinar automáticamente cuántas de las estructuras de Wiki son preguntas y cuántas peticiones de información, por lo que sería necesaria una revisión manual para obtener este dato. Teniendo en cuenta el altísimo número de casos del corpus, esto no ha sido posible. Para las preguntas parciales podemos manejar datos estimativos del número de casos, a partir de la presencia de las partículas interrogativas. Diferenciar de modo automático las preguntas totales de las peticiones de información, en cambio, no es posible con garantías. Por esta razón, el corpus Wiki se ha utilizado con especial cuidado. En todos los análisis realizados se ha intentado utilizar solo preguntas (mediante selecciones y filtrados previos) y no otro tipo de estructuras como las de los ejemplos anteriores.

Las características de los tres corpus compilados responden al interés por manejar preguntas próximas a las que se harían en distintos tipos de sistemas de BR. Por una parte, utilizamos preguntas que se emplearon en el pasado para evaluar sistemas reales de BR (corpus CLEF): preguntas parciales y, en la mayor parte de los

93 http://respuestas.wikia.com/wiki/WikiRespuestas
94 10/05/2012.
95 Reproducimos los ejemplos de Wiki tal cual están en el corpus, sin corregir los errores ortográficos.

casos, de tipo factual. Por otra, utilizamos preguntas reales sobre conocimiento general extraídas de la web con un perfil similar a las de TREC y CLEF (Trivial). Finalmente, utilizamos preguntas reales planteadas por usuarios en un portal de preguntas y respuestas (Wiki). Los dos primeros corpus provienen de entornos «controlados», no presentan prácticamente errores lingüísticos y contienen solo preguntas. Wiki, sin embargo, proviene de un entorno no supervisado, lo que implica ciertos problemas (el lenguaje utilizado por los usuarios está lleno de errores que dificultan el análisis automático; no todas las estructuras son preguntas), pero, al mismo tiempo, implica un especial interés, ya que constituye una muestra de estructuras reales que los hablantes utilizan para preguntar en un entorno no controlado.

DATOS GENERALES – CORPUS

	Trivial	Clef	Wiki	Total
P. Totales	22	0	69.148[96]	69.170
P. Parciales	227	586	100.171	100.984
Total	249	586	169.319[97]	170.154

Tabla 5: N.º de preguntas por corpus.

Como decíamos más arriba, en el caso de Wiki no es posible saber con exactitud cuántas de las estructuras del corpus son preguntas parciales o totales y cuántas otro tipo de estructuras. El único cálculo aproximativo posible es para la parciales; para realizarlo, se extrajeron todas las preguntas que iban encabezadas por interrogativo/preposición+interrogativo.

Lo anterior implica que solo en el caso de Trivial y Clef podamos hacer cálculos absolutos de la incidencia de determinados aspectos sobre un conjunto de totales o parciales.

Además de estos tres corpus, contamos también con los datos de Gayo (2010), donde se presentan datos de corpus[98] relativos a distintos aspectos gramaticales de preguntas parciales. Los datos de Gayo (2010) se utilizarán con una finalidad contrastiva.

96 Como no es posible determinar automáticamente el número de totales, ofrecemos un cálculo aproximado resultado de restar las parciales al número total de casos del corpus (incluidas las peticiones de información).

97 Como no es posible calcular cuántas de las estructuras del corpus son preguntas, utilizamos el dato correspondiente al total de estructuras del corpus (incluidas las peticiones de información).

98 El corpus utilizado en Gayo (2010) corresponde a la sección contemporánea del Archivo de textos hispánicos de la Universidad de Santiago (ARTHUS). Los textos de ARTHUS (http://www.bds.usc.es/) son en su mayor parte de tipo literario (con mayor incidencia de la narrativa), aunque también hay una buena parte de textos orales y periodísticos.

3.2.2 Análisis de corpus

La aproximación al trabajo en corpus no ha sido homogénea ya que los rasgos a analizar tampoco lo son: no es lo mismo (en términos de análisis) analizar la presencia de términos de polaridad negativa en las preguntas que el uso de un valor semántico de cierta partícula interrogativa. El primer aspecto conlleva una búsqueda,

extracción y cuantificación de casos; el segundo, un análisis lingüístico detallado inviable si hablamos de miles de ejemplos.

En general, siempre que ha sido posible, se ha utilizado un análisis de tipo automático que abarcase el mayor número de casos posibles en los tres corpus. Cuando se ha tratado de aspectos que requerían un análisis manual, el número de casos se ha reducido. En las secciones siguientes presentaremos los aspectos analizados uno a uno, detallando las particularidades del trabajo realizado en cada caso.

3.3 Estudio de corpus

3.3.1 Orden de constituyentes

En esta sección analizaremos los aspectos relativos al orden de constituyentes en las preguntas. En primer lugar nos ocuparemos de las preguntas totales y, a continuación, de las parciales.

3.3.1.1 Orden de elementos en las interrogativas totales

En el capítulo anterior veíamos que las preguntas totales presentan, según (Escandell, 1999), un orden no marcado (con el verbo al inicio de la estructura seguido del sujeto y a continuación los otros complementos), junto con un orden marcado (con el sujeto colocado antes del verbo). Para (Escandell, 1999), ambos órdenes están relacionados con la estructuración de la información y conllevan significados y usos distintos: en el orden no marcado la proposición es abierta y toda la pregunta es información nueva; en el orden marcado la interrogación opera sobre una proposición cerrada, de manera que la pregunta contiene una proposición ya presente de alguna manera en el discurso. Este hecho nos lleva a pensar que, si (Escandell, 1999) está en lo cierto, el orden no marcado sería el esperado en las preguntas sin contexto, mientras que el orden marcado sería más bien propio de contextos dialogados.

Para determinar si la afirmación anterior es correcta hemos hecho un estudio del orden de constituyentes en las preguntas totales de nuestros corpus. Trivial se ha utilizado completo ya que las preguntas totales son escasas; en el caso de Wiki, debido al tamaño del corpus y a sus características, nos resultaba imposible analizar todas las preguntas totales. Por esta razón se realizó una selección de 145 preguntas totales para el estudio[99]. Como ya hemos dicho, en Clef no hay preguntas totales.

99 El número de totales es reducido, pero se debe a que la extracción de totales de Wiki requiere un análisis manual complejo. Recordemos que en Wiki solo es posible distinguir automáticamente las

	Trivial	Wiki	Total
P. Totales	22	145	167

Tabla 6: N.º de preguntas totales analizadas por corpus.

DATOS OBTENIDOS

Trivial

- Totales con orden no marcado = 17 (9).
- Totales con orden marcado = 5 (10).

(9) *¿Beben las gaviotas agua de mar?*
(10) *¿La araña es un insecto?*

Wiki

- Totales con orden no marcado = 40.
- Totales con otro orden = 105.

En estas 105 totales con orden alternativo, el constituyente antepuesto al verbo puede ser de dos tipos:

- Sujeto: 91 casos
  (11) *¿El leon es carnivoro?*
- Adjunto: 14 casos
  (12) *¿En colombia abra pokemon oro corazon y plata alma en español?*

Conclusiones sobre el orden en las totales

	Trivial	Wiki	Total
O. no marcado	17	40	57
O. marcado	5	91	96
Otros órdenes	0	14	14
Total	22	145	167

Tabla 7: Orden de constituyentes en las preguntas totales.

preguntas parciales. De hecho, estas 145 preguntas resultaron ser las únicas totales (extraídas mediante un análisis manual) de 1.000 estructuras (que no eran preguntas parciales) de Wiki.

A la luz de los datos, el orden no marcado no es tan común como se esperaría. En Trivial sí es el más abundante, pero no en Wiki.

Los datos demuestran que en las totales el orden verbo+sujeto alterna con otros órdenes, especialmente con el orden sujeto+verbo. La mayor incidencia de este último orden en Wiki nos hace pensar que el orden sujeto+verbo en las totales parece ser tan común como el orden no marcado. Por tanto, aunque esta ordenación implique una estructura informativa compleja en la que la interrogación opera sobre una proposición cerrada, tal como defiende (Escandell, 1999) (cf. sección 2.4.2.1 del capítulo 2), este hecho parece no interferir para que este tipo de interrogativas se utilicen como preguntas con fines informativos en contextos no dialogados.

3.3.1.2 Orden de constituyentes en las parciales

En el capítulo anterior vimos que las parciales presentan un orden que parece ser prototípico, con el interrogativo encabezando la estructura, seguido del verbo. Según (Escandell, 1999), este orden responde, como en el caso de las totales, a cuestiones relativas a la estructuración de la información (cf. capítulo 2, secciones 2.2.3.1 y 2.2.4.2.2). Vimos también que este orden puede alternar con otros dos posibles: la anteposición tipo A y la anteposición tipo B, ambos documentados (con muchos menos casos que el orden «prototípico») en el estudio de corpus de Gayo (2010).

En las secciones que siguen nos ocuparemos del estudio del orden de constituyentes en las parciales de nuestros tres corpus. Analizaremos dos aspectos: orden «prototípico» *vs.* «órdenes alternativos»; «órdenes alternativos»: incidencia de la anteposición tipo A, incidencia de la anteposición tipo B y otras ordenaciones posibles.

	Trivial	Clef	Wiki	Total
P. Parciales	227	586	100171	100984

Tabla 8: N.º total de parciales en los tres corpus.

3.3.1.2.1 Orden prototípico *vs.* órdenes alternativos

En esta sección se analizará qué órdenes se documentan en una muestra de interrogativas parciales de nuestros tres corpus.

Como el número total de parciales es muy elevado y el análisis requerido no se podía ejecutar automáticamente[100], se han seleccionado al azar 100 preguntas parciales de cada corpus y se ha analizado el orden de constituyentes.

100 La anteposición tipo A sí se puede detectar automáticamente (porque implica algún elemento entre el interrogativo inicial y la partícula interrogativa), pero no así la anteposición tipo B, donde se coloca un elemento entre el interrogativo y el verbo (recordemos que Wiki está lleno de errores, con lo cual, es imposible reconocer en muchos casos el verbo de la pregunta).

DATOS OBTENIDOS

Trivial

Las 100 preguntas seleccionadas presentaban el orden prototípico:
(13) *¿Por qué condimento pagan los chefs el precio más alto?*

Clef

Todas las preguntas de CLEF presentaban orden prototípico.
(14) *¿Cómo se llama la mujer de Bill Gates?*

Wiki

Las 100 preguntas seleccionadas al azar presentaban orden prototípico (26).
(15) *¿Cuántos cromosomas tiene un caracol?*

Conclusiones

La siguiente tabla recoge los datos generales obtenidos sobre ordenación en las parciales en las 300 preguntas seleccionadas al azar en los corpus:

	Trivial	Clef	Wiki	Total
O. Prototípico	100	100	100	300
Ant. tipo A	0	0	0	0
Ant. tipo B	0	0	0	0
Otros	0	0	0	0
Total	100	100	100	300

Tabla 9: Órdenes documentados para la selección de parciales en los tres corpus.

Como queda claro en la tabla, el orden denominado «prototípico» parece ser el más común en nuestros tres corpus. Esta conclusión preliminar se verá confirmada con los datos de las secciones siguientes.

3.3.1.2.2 Órdenes alternativos

3.3.1.2.2.1 Anteposición tipo A

Como hemos visto, en la anteposición tipo A se coloca un constituyente antes del interrogativo, generalmente, el sujeto[101]:

101 Un análisis más pormenorizado de los casos de anteposición tipo A de Gayo (2010) ha revelado

(16) *¿Y* yo *dónde meto a mi vaca?*

(Ejemplo tomado de Gayo, 2010)

A continuación analizaremos la incidencia de la anteposición tipo A y sus características en nuestros tres corpus. Como la anteposición tipo A se puede detectar automáticamente, utilizaremos los datos de nuestros tres corpus completos.

DATOS OBTENIDOS

Trivial

No hay casos de anteposición tipo A.

Clef

No hay casos de anteposición tipo A.

Wiki

De las 100.984 parciales en Wiki, 31 presentaban anteposición tipo A:

- 5 con sujeto antepuesto:
  (17) *Ecuador a qué organismo internacional pertenece?*

- 1 con complemento indirecto antepuesto:
  (18) *A aggron porque no le afecta derribo ni doble filo y porqué es el único?*

- 22 con adjunto antepuesto:
  (19) *¿En enfermería para qué sirve la bioquímica?*

  20 de los 22 casos corresponden a la tematización del adverbio *actualmente*:
  (20) *¿Actualmente cuantos habitantes tiene colombia?*

- 3 casos con *Y* + sujeto:
  (21) *¿Y saturno cuantas lunas tiene?*

Conclusiones

En la siguiente tabla recogemos los datos totales para la anteposición tipo A en nuestros tres corpus:

que el constituyente que se antepone con más frecuencia es el sujeto, con un 83,7% de los casos (datos no publicados).

	Trivial	Clef	Wiki	Total
P. Parciales	227	586	100.984	101.797
Ant. tipo A	0	0	31	31

Tabla 10: Anteposición tipo A en los tres corpus.

Como muestra la tabla, la incidencia de la anteposición tipo A es muy baja en nuestros corpus. Solo se documenta en uno de los tres, Wiki, con una incidencia casi imperceptible (0.03% del total).

En relación al argumento antepuesto al interrogativo, estos son los datos:

	Trivial	Clef	Wiki	Total
Sujeto	0	0	6	6
Adjunto	0	0	24	24
O. Indirecto	0	0	1	1
Total	0	0	31	31

Tabla 11: Función del constituyente antepuesto al interrogativo.

Como recoge la tabla, en nuestros corpus la función que más se antepone al interrogativo es la de adjunto, seguida de lejos por el sujeto. Este dato difiere de los de Gayo (2010).

Esta y otras posibles diferencias tal vez se deban a que el contexto de uso de los casos con anteposición tipo A en los textos de Gayo (2010) es diferente de los de nuestros corpus: allí la mayoría de los ejemplos corresponden a un contexto discursivo en el que estas estructuras parecen usarse para introducir un nuevo tema en el discurso (el tema se corresponde con el elemento antepuesto al interrogativo). El siguiente ejemplo sería un buen ejemplo de este uso (el destacado es nuestro):

(22)
- Está bien. Se lo digo claramente: usted se está quedando calvo.
Me miró sin pestañar.
-**¿Y por casa cómo andamos?** --preguntó--. De acuerdo, le concedo, soy un pelado y además un ingenuo.
¿O el ingenuo es usted? Porque si no me equivoco está pidiendo que me malquiste con las autoridades

(Ejemplo tomado de Gayo, 2010)

Este uso parece encajar con las características de los casos de anteposición tipo A documentados en Gayo (2010)[102]:

- El sujeto es el elemento que más se antepone (83,7% de los casos).
- El tipo de unidad que más se antepone es el pronombre personal (69,7% de los casos).
- En un 39,5% de los casos el elemento tematizado va precedido de la partícula *y*, que se utiliza para marcar al constituyente que le sigue como tema (Escandell, 1999, p. 3955).
- El 76,5% de los casos pertenece al género oral (teatro y discurso) del corpus ARTHUS; los casos procedentes de la narrativa y el ensayo se corresponden con contextos de estilo directo.

3.3.1.2.2.2 Anteposición tipo B

Como hemos visto, en la anteposición tipo B se coloca un constituyente (generalmente el sujeto) entre el interrogativo y el verbo:

(23) *¿Por qué* los submarinos *no tienen ventanas para ver el fondo del mar?*

(Ejemplo tomado de Gayo, 2010)

Un estudio de los casos con anteposición tipo B en (Gayo, 2010)[103] muestra que la estructura de esta es más homogénea que la de la anteposición tipo A. En la anteposición tipo B:

- El interrogativo siempre es no argumental (*por qué* en un 81% de los casos).
- El constituyente antepuesto funciona en el 90.5% de los casos como sujeto.

A continuación analizaremos la incidencia de la anteposición tipo B en nuestros tres corpus. Utilizaremos Trivial y Clef completos; dado el tamaño de Wiki y que la identificación de la anteposición tipo B requiere un análisis manual (cf. *supra*, nota 89), en el caso de este corpus hemos trabajado sobre una selección aleatoria de 200 preguntas parciales.

DATOS OBTENIDOS

Trivial

No hay casos de anteposición tipo B.

Clef

En Clef se documentan 3 casos de anteposición tipo B. En todos los casos el constituyente entre el interrogativo y el verbo es el sujeto.

102 Datos no publicados.
103 Datos no publicados.

A continuación detallaremos la estructura sintáctica (funciones y estructura interna de la frase interrogativa) de los casos de anteposición tipo B documentados en Clef.

N.º Total de casos = 3
Interrogativo funcionando como adjunto = 3

Qué = 2:

(24) *¿En qué fecha Estados Unidos invadió Haití?*
(25) *¿En qué fecha El Corte Inglés compró Galerías Preciados?*

Cuándo = 1:

(26) *¿Desde cuándo Portugal es una república?*

Wiki

No se documenta ningún caso de anteposición tipo B en la muestra de 200 preguntas parciales.

Conclusiones

La tabla recoge los datos generales para la anteposición tipo B en los tres corpus:

	Trivial	Clef	Wiki	Total
P. Parciales	227	586	200	1.013
Ant. tipo B	0	3	0	3

Tabla 12: N.º Total de casos con anteposición tipo B en nuestros tres corpus.

Como muestra la tabla, la incidencia de la anteposición B también es muy baja; además, solo se documenta en uno de los tres corpus.

En los tres casos documentados, el interrogativo funciona como adjunto, siendo en dos casos *qué* (con la preposición *en*) y en uno *cuándo* (con la preposición *desde*).

Con el objetivo de realizar una comparación entre la incidencia de anteposición tipo A y B en nuestros corpus, hemos analizado también la presencia de casos de anteposición tipo A en la selección de 200 preguntas de Wiki. Los resultados son los siguientes:

	Trivial	Clef	Wiki	Total
P. Parciales	227	586	200	1013
Ant. tipo A	0	0	0	0
Ant. tipo B	0	3	0	3

Tabla 13: Datos globales sobre la anteposición A y B en los tres corpus.

Los datos de la tabla muestran que, dentro de la escasez de casos de ambos tipos de construcción, parece que la anteposición tipo B puede ser más común que la anteposición tipo A en nuestros corpus. Recordemos que en (Gayo, 2010) la anteposición tipo A era más abundante que la tipo B, aunque, como hemos visto, esta construcción parece tener allí una función diferente a la documentada en Wiki (cf. *supra*).

3.3.1.2.2.3 Otros elementos antepuestos al interrogativo

En nuestro análisis relativo al orden de elementos en las parciales, hemos podido observar que, además de las preposiciones, también se documentan otros elementos antepuestos al interrogativo (solo en Wiki). Estos elementos son los siguientes:

1) Adverbios o adverbios modificados:

(27) ¿Aproximadamente *cuántos soldados murieron en la segunda Guerra Mundial?*
(28) Dentro *de que Bioma se encuentra la Republica Dominicana?*

Podríamos considerar (27) como un caso de anteposición tipo A. No obstante, no nos parece que se trate de la misma estructura. En la anteposición tipo A el elemento que se coloca antes del interrogativo es un constituyente con entidad semántica propia que, creemos, el hablante coloca ahí para darle relevancia informativa. En casos como el de *aproximadamente* consideramos que el adverbio se coloca ahí por su fuerte relación con el interrogativo y la noción de cuantificación, no para darle relevancia desde un punto de vista informativo o semántico.

Hemos documentado en Wiki los siguientes adverbios funcionando en el mismo tipo de estructuras:

- *Aproximadamente*: 11 casos.
  (29) *¿Aproximadamente cuantos habitantes hay en el mundo?*

- *Exactamente*: 5 casos.
  (30) *¿Exactamente a cuanto equivale una onza liquida?*

- *Cada*: 36 casos;
  (31) *¿Cada cuántos años se celebran los juegos olímpicos?*

2) Al igual que en Gayo (2010), hemos documentado en Wiki casos con la partícula «y» iniciando la pregunta. En total, hay 56 preguntas con esta estructura:

(32) *Y como ordeno cronológicamente el cuento Quilla pan?*

3.3.1.2.2.4 Conclusiones sobre el orden en las parciales

En las secciones anteriores nos hemos ocupado del análisis del orden de constituyentes en las interrogativas parciales.

En primer lugar, hemos realizado un estudio con el fin de comprobar la incidencia del orden denominado prototípico *vs.* otros órdenes en nuestros corpus. Como hemos visto, solo el orden prototípico se documenta.

A continuación nos hemos ocupado del estudio de las ordenaciones alternativas, la anteposición tipo A y la anteposición tipo B. En los tres corpus completos, solo se documentan 31 casos de anteposición tipo A (todos en Wiki). En dos de los tres corpus completos y una selección de 200 preguntas de Wiki, la anteposición tipo B solo se documenta en 3 casos (y siempre con el interrogativo funcionando como adjunto). En esta misma selección de preguntas, la anteposición tipo A ni siquiera se documenta. Esto nos hace pensar que:

- Ambos tipos de ordenación tienen una incidencia prácticamente imperceptible en nuestros tres corpus;

- Probablemente la anteposición tipo B sea más común que la tipo A.

3.3.2 Negación

En el capítulo 2 hemos visto la influencia de la negación en el significado de las preguntas: el hecho de que esta sea interna o externa genera interpretaciones distintas para una pregunta (cf. capítulo 2, sección 2.3.5).

En esta sección exploraremos el uso del adverbio *no* en nuestros corpus (datos generales de uso y combinado con cada interrogativo). Analizaremos, además, en qué casos la negación es de tipo interno o externo.

DATOS OBTENIDOS

3.3.2.1 *No*

Trivial

Se documentan 4 casos.

(33) ¿Qué función vital no se puede hacer al mismo tiempo que deglutir?

(34)

En los cuatro casos la negación es interna.

Clef

Se documenta 1 caso:
(34) *¿Qué premiado por el Instituto Goethe no recogió el premio?*

La negación es interna.

Wiki

En wiki se documentan 771 casos.
(35) *Qué paises no existian hace dos siglos en europa?*

Hemos hecho un análisis detallado de 100 de estos 771 casos para determinar en cuántos la negación era interna y en cuántos la negación era externa. De esos 100, solo dos casos pueden interpretarse como negación externa:

(36) *Porque los que preguntan sobre pokemon mejor no se meten en la wikidex?*
(37) *7 no es divisor de 100?*

(36) es, como veremos más abajo, uno de los dos casos documentados de *por qué* + *no* con valor de sugerencia, y la interpretación solo puede ser de negación externa. En (37), sin embargo, la interpretación puede ser de negación interna (38) o externa (39):

(38) *7 no es divisor de 100, ¿verdad?*
(39) *7 es divisor de 100, ¿no?*

Conclusiones

La siguiente tabla recoge los datos generales sobre la incidencia de la negación en nuestros corpus:

	Trivial	Clef	Wiki	Total
No	4	1	771	0

Tabla 14: Incidencia de la negación en los tres corpus.

Como vemos en la tabla, la negación se documenta en nuestros tres corpus, aunque con una incidencia muy baja.

Respecto a los tipos de negación, externa o interna, obtenemos la siguiente tabla:

	Trivial	Clef	Wiki	Total
N. Interna	4	1	98	103
N. Externa	0	0	2	2
TOTAL	4	1	100	105

Tabla 15: Incidencia de la negación interna y externa en los tres corpus.

Por otro lado y como era de esperar, prácticamente todos los casos de negación constituyen casos de negación interna y no externa. Hay que tener en cuenta además que los dos casos con negación externa corresponden a Wiki, el corpus con estructuras que más se acercan a la oralidad.

Si atendemos a la presencia del adverbio *no* por tipo de pregunta, el reparto queda como sigue:

	Trivial	Clef	Wiki	Total
P. Totales	0	0	90	90
P. Parciales	4	1	681	686
TOTAL	4	1	771	776

Tabla 16: Presencia de *no* por tipo de interrogativa.

Vemos en la tabla que la negación es más común en las preguntas parciales que en las totales en nuestros tres corpus (esto era lo esperable, por otra parte, ya que las preguntas parciales son más comunes que las totales en nuestros corpus; cf. *supra*).

Respecto a la incidencia de la negación en las parciales por tipo de interrogativo, obtenemos los siguientes resultados:

	Trivial	Clef	Wiki	Total
Cómo	0	0	4	4
Cuál	0	0	30	30
Cuándo	0	0	3	3
Cuánto	0	0	5	5
Dónde	0	0	2	2
Qué	4	1	129	134
Quién	0	0	2	2
Por qué	0	0	506	506
Total	4	1	681	686

Tabla 17: Presencia de *no* por tipo de interrogativo.

En cuanto al tipo de interrogativo, *por qué* es con diferencia el interrogativo que más se combina con predicados con negación, seguido de *qué; cuál* se sitúa a una cierta distancia y el resto de los interrogativos presentan muy pocos casos.

3.3.3 Partículas interrogativas

3.3.3.1 Incidencia de los interrogativos por corpus

En esta primera sección ofrecemos datos de la incidencia de los interrogativos en cada uno de los tres corpus (tenemos en cuenta los casos con y sin preposición).

	Trivial	Clef	Wiki	Total
Qué	114	278	28.825	29.217
Quién	4	102	3.325	3.431
Cuál	51	45	20.764	20.860
Cuánto/cuán	34	68	19.808	19.910
Cuándo	1	29	990	1.020
Dónde	8	23	4.632	4.663
Cómo	15	41	21.573	21.629
Por qué	0	0	2.54	19.756
Total	227	586	100.171	100.984

Tabla 18: Incidencia de los interrogativos por corpus.

Como muestra la tabla, los tres corpus coinciden en los extremos: el interrogativo más documentado es *qué* y el menos documentado *por qué*.

Qué, *cuál* y *cuánto* están entre los cinco interrogativos más documentados en los tres corpus, mientras que *cuándo* y *por qué* están entre los tres menos documentados también en los tres corpus.

En términos generales, la jerarquía en la incidencia en corpus quedaría así: *qué*, *cómo*, *cuál*, *cuánto*, *dónde*, *quién*, *cuándo*, *por qué*.

En esta jerarquía hay una clara diferencia en cuanto a la incidencia: los cuatro primeros interrogativos superan o se sitúan en torno a los 20.000 casos, mientras que los cuatro últimos están por debajo de los 5.000. Como veremos más abajo cuando tratemos los aspectos semánticos de los interrogativos, de los cuatro más documentados el que presentaría menos problemas de procesamiento sería *cuánto*; los otros tres (*qué*, *cuál* y *cómo*) son de los más complejos.

3.3.3.2 Preposiciones

En el capítulo 2 hemos visto que todos los interrogativos excepto *por qué* se pueden combinar con preposición para formar un constituyente complejo. En relación a estas combinaciones, NGRALE y Bosque (1984) señalan ciertas restricciones para *cuándo* y *dónde* (cf. sección 2.4.3.2).

Con el fin de determinar con qué preposiciones se combina cada una de las partículas interrogativas hemos analizado todas las combinaciones documentadas en nuestros tres corpus completos. Recogemos los datos a continuación, tratando uno a uno los interrogativos.

DATOS OBTENIDOS

Cuándo

En Clef y en Wiki se combina con una sola preposición (y con un solo caso en ambos corpus): *desde*.

En Trivial con ninguna.

Dónde

En Trivial se combina con una preposición: *de* (1)[104].

En Clef se combina con una: *de* (1).

En Wiki se combina con cuatro: *a* (44), *de* (14), *en* (583), *por* (475).

Cuál

En Trivial y Clef no se documenta con preposición.

En Wiki se combina con cuatro: *a* (6), *de* (13), *en* (75), *por* (14).

104 La cifra entre paréntesis corresponde al número de casos documentados.

Cuánto

En Trivial se documenta con dos preposiciones: *con* (1), *en* (1).
En Clef se documenta con dos: *a* (5) y *de* (1).
En Wiki se documenta también con cuatro: *a* (283), *de* (105), *en* (329), *por* (50).

Quién

En Trivial no se documenta con preposiciones.
En Clef se documenta con dos preposiciones: *a* (1) y *para* (2).
En Wiki se documenta con dos: *a* (6), *de* (1).

Qué

En Trivial se combina con seis: *a* (2), *con* (9), *contra* (1), *de* (12), *en* (11), *por* (1).
En Clef se combina con ocho: *a* (14), *con* (2), *contra* (2), *de* (24), *en* (58), *entre* (1), *para* (1), *sobre* (1).
En Wiki se combina con seis: *a* (54), *bajo* (4), *de* (32), *en* (92), *para* (34), *por* (2).

Conclusiones

La siguiente tabla recoge todas las posibles combinaciones de preposición + interrogativo documentadas por corpus:

	Trivial	Clef	Wiki
Cuándo	0	desde – 1	desde – 1
Dónde	de – 1	de – 1	a – 44; de – 14; en – 583; por – 475
Cuál	0	0	a – 6; de – 13; en – 75 por – 14
Cuánto	con – 1; en – 1	a – 5; de – 1	a – 283; de – 105; en – 329 por – 50
Quién	0	a – 1; para – 2	a – 6; de – 1

	Trivial	Clef	Wiki
Qué	a – 2 ; con – 9; contra – 1; de – 12 en – 11; por – 1	a - 29 ; con – 4 contra – 2; de – 52 en – 126; entre - 1 para – 8; sobre - 4	a - 54; bajo – 4; de - 32 en – 92; para – 34; por - 2

Tabla 19: Combinación de los interrogativos con preposición en los tres corpus.

Por número total de combinaciones con preposiciones distintas, estos son los resultados de nuestros corpus para cada interrogativo:

- *Cuándo*: se combina solo con una preposición (*desde*).
- *Quién*: se combina con tres preposiciones distintas.
- *Dónde, cuál y cuánto*: se combinan con cuatro preposiciones distintas.
- *Qué*: se combina con nueve preposiciones distintas.

Recordemos que, según Bosque (1984), la combinación de *dónde* con *en* no es correcta (Bosque, 1984, p. 266), mientras que para NGRALE esta combinación sí se acepta (Real Academia Española, 2009, 22.15f). En nuestros corpus, esta construcción se documenta solo en Wiki, pero con un alto número de casos: 583. Recordemos, por otra parte, que Wiki es el corpus con un lenguaje menos cuidado.

3.3.3.3 Análisis semántico de los interrogativos

En el capítulo anterior hemos mostrado que las partículas interrogativas pueden presentar diferentes valores semánticos con distintos niveles de exactitud o precisión (cf. capítulo 2, sección 3.2.2.1). También hemos visto que, para algunas partículas interrogativas, existen distintos valores semánticos dependiendo de la estructura sintáctica en la que se integren. Por otro lado, hemos apuntado que los valores concretos en los casos de significado impreciso pueden variar dependiendo de factores pragmáticos como los intereses del usuario o el perfil del sistema de BR en el que se integre una pregunta (cf. 3.2.1).

Teniendo en cuenta la complejidad de estos aspectos y su interés para nuestro análisis de las preguntas, en esta sección nos ocuparemos del estudio en corpus de los valores semánticos de ciertos interrogativos. Los interrogativos estudiados son los siguientes: *cuándo*, *dónde*, *cómo*, *quién*, *qué*, *cuál* y una construcción específica con *por qué*. Como veremos, hemos tenido en cuenta las particularidades semánticas de cada interrogativo a la hora de diseñar su análisis. El objetivo principal es determinar qué valores se documentan en nuestros corpus, si alguno de estos valores es más común que los otros, etc.

Hemos dejado fuera de este análisis *cuánto* y *por qué,* por distintas razones. En el caso de *cuánto*, como ya vimos en la sección 3.2.1 del capítulo anterior, el valor semántico de la variable es siempre el mismo y, además, es siempre concreto y

determinable a través de un análisis léxico semántico de la pregunta. Si el interrogativo funciona como determinante, la variable cuantificada está presente en la pregunta (*¿Cuánto* dinero *gana al mes un informático de Google?*); si el interrogativo funciona como pronombre, la variable cuantificada puede deducirse del análisis semántico del verbo (*¿Cuánto* mide *la Torre Agbar?*). El caso de *por qué* es justamente el contrario: la variable que define es altamente abstracta (el concepto de causa), y creemos que, en el actual estado de la cuestión, su estudio para el análisis de la pregunta no puede ir más allá de reconocer el concepto de causa. Consideramos que en el caso de *por qué* lo realmente interesante sería un estudio de la causa en la fase de extracción de la respuesta, tratando de determinar, por ejemplo, los modos lingüísticos para expresar la causa. Para *por qué* solo hemos analizado un aspecto muy concreto: la construcción *por qué + no* con el valor semántico de sugerencia.

En cuanto a la metodología empleada, los distintos valores semánticos para cada interrogativo no se establecían de antemano, sino que se iban diferenciando a partir del análisis exhaustivo de las preguntas seleccionadas en cada uno de los corpus. Para ello, a cada pregunta se le asignaba un valor semántico. Los valores diferenciados siguiendo este procedimiento son (como veremos) bastante homogéneos y se repiten casi siempre en los tres corpus (el corpus que más valores únicos presenta es Wiki).

Diferenciar valores semánticos concretos y clasificar las distintas preguntas de acuerdo a ellos no es una tarea que pueda proporcionar resultados discretos y exactos; en muchos casos, los límites entre un valor semántico u otro pueden ser subjetivos y cuestionables. Es por esta razón por la que nuestra clasificación no pretende presentarse como un modelo de análisis, sino más bien ser una herramienta para determinar hasta qué punto los interrogativos presentan valores de significado distintos en nuestros corpus, cuándo estos valores son imprecisos, si alguno es más común que otro, etc.

Hemos analizado solo preguntas en las que el interrogativo se construye sin preposición. Esta decisión responde al deseo de analizar los valores de los interrogativos en solitario, sin el componente semántico que puede aportar la combinación con la preposición. Debido al grado de detalle del análisis, hemos establecido para cada interrogativo un máximo de 100 preguntas por corpus.

Para cada interrogativo presentaremos, en primer lugar, todos los valores semánticos documentados de forma global y, a continuación, los datos de cada valor en nuestros tres corpus.

3.3.3.3.1 Valores semánticos de *cuándo*

En el capítulo 2 vimos que *cuándo* puede tener dos valores semánticos: un valor temporal (40) y un valor no temporal de carácter hipotético (41).

(40) *¿Cuándo pisó el hombre la Luna por primera vez?*
(41) *¿Cuándo decimos que un ángulo es recto?*

Como ya se explicó en 3.2.2.1, el tipo de evento a localizar y el significado del verbo nos dan pistas sobre el valor concreto de la ubicación temporal que exige *cuándo*. Así en

(42) *¿Cuándo nació bella thorne?*

se pregunta probablemente por una fecha, mientras que en

(43) *¿Cuándo comen los guepardos?*

se pregunta por un período temporal general del tipo «por la mañana» o «por la noche».

En el análisis en nuestros corpus, hemos distinguido, por un lado, un valor general de tipo temporal y un valor general de 'hipótesis' (44).

(44) *¿Cuándo un elemento pertenece a un conjunto?*

Para el valor temporal, se han distinguido además los siguientes subtipos semánticos:

- fecha/año: ubicación de un evento ocurrido en una fecha o año concreto: *el 5 de mayo de 1983, en 1987, en el año 1876*.

  (45) *¿Cuándo fue invadida Jerusalén por el general Titus?*

- período temporal: ubicación de un evento ocurrido en un período temporal no correspondiente a una fecha. El período temporal puede ser histórico (*en el Renacimiento*) o no histórico (*en primavera, por la mañana,* etc.).

  (46) *¿Cuándo terminó el período mesozoico?*
  (47) *¿Cuándo fue domesticado el perro?*
  (48) *¿Cuándo comen los guepardos?*

- casos ambiguos: casos en los que la ubicación temporal por la que se pregunta puede ser de cualquiera de los dos tipos anteriores:
  (49) *¿Cuándo gobernó Inglaterra Henry VIII?*

En (49) podemos responder con un rango de fechas (*entre X e Y*) o con un período temporal (*en el Renacimiento*).

A continuación recogemos los datos relativos a la documentación de cada uno de estos valores en nuestros tres corpus.

DATOS OBTENIDOS

Trivial

Solo hay un caso de pregunta con *cuándo*:

(50) *¿Cuándo son las sombras más cortas, en invierno o en verano?*

El valor semántico en este caso es el de ubicación de un evento en un período temporal (*invierno* o *verano*).

Clef

Hay 28 preguntas con *cuándo*.
Solo se documenta el valor de 'fecha'/'año':

(51) *¿Cuándo murió Lenin?*

Wiki

En los 100 casos seleccionados se documentan los siguientes valores:

- 'fecha'/'año' = 38 casos:
  (52) *¿Cuándo se fundó el SRA?*

- 'período temporal' = 7 casos:
  (53) *¿Cuándo se formó la Tierra?*

- casos ambiguos de ubicación temporal = 39:
  (54) *¿Cuándo hay un eclipse de sol?*

- escenario hipotético en el que se produce el evento indicado en la pregunta = 15 casos:
  (55) *Cuándo un sistema de fuerzas colineales se encuentra en equilibrio?*

En algunos casos, este tipo de preguntas requiere la elaboración de una hipótesis con opinión:
(56) *Cuándo decimos que hay injusticia?*

Conclusiones

La siguiente tabla recoge los valores documentados en los tres corpus:

	Trivial	Clef	Wiki	Total
Fecha	0	28	39	67
P. Temporal	1	0	8	9
Escenario hipotético	0	0	15	15
Casos ambiguos	0	0	38	38
Total	1	28	100	129

Tabla 20: Valores semánticos documentados para *cuándo* en los tres corpus.

El valor temporal es claramente el dominante y, dentro de este, el de ubicación mediante una fecha o año. Si tenemos en cuenta que en los casos ambiguos lo que tenemos siempre es ambigüedad respecto a la ubicación temporal de un evento, se concluye que el valor temporal es incluso más común. Observamos también que el valor temporal se documenta en los tres corpus, mientras que el hipotético solo se documenta en Wiki.

Los casos documentados muestran también que la localización temporal puede apuntar hacia el pasado (57), el futuro (58) o hacia un evento periódico (59):

(57) *Cuándo nació bella thorne?*
(58) *Cuándo saldrá la beta de gears of war 3?*
(59) *Cuándo se celebra el día del trabajador?*

El tiempo del verbo nos aporta información sobre esta cuestión.

Respecto al tiempo verbal, hemos observado también que, como se apuntaba en el capítulo dos (cf. sección 2.3.2.1.1), el valor hipotético se construye en la muestra analizada siempre en presente.

Por otra parte, el trabajo en corpus ha mostrado que determinar el valor temporal concreto de *cuándo* es una tarea compleja incluso para un humano. El tipo de evento y la información semántica del verbo son clave, pero incluso con esta información hay casos que se muestran ambiguos y que solo parecen poder resolverse accediendo a los intereses concretos del usuario. No obstante, hay determinados patrones léxico sintácticos que parecen apuntar a valores temporales concretos de *cuándo*. Por ejemplo, los dos patrones a continuación apuntan casi siempre al valor 'fecha':

- *cuándo* + *nacer*/*morir* + persona
- *cuándo* + *fundarse*/*inventarse* + entidad

Manejar este tipo de patrones en un sistema de BR podría ser útil para determinar el valor temporal concreto de *cuándo*.

3.3.3.3.2 Valores semánticos de *dónde*

En el capítulo 2 vimos que el valor semántico general de *dónde* es también impreciso, y que está determinado por la granularidad.

En esta sección analizaremos los valores semánticos documentados para el interrogativo en nuestros corpus. Estos valores son subtipos semánticos del valor general de tipo locativo al que se asocia el interrogativo.

Los valores concretos documentados son los siguientes:

- 'localización geográfica':

(60) *¿Dónde está el desierto del Gobi?*

- 'localización física no geográfica':

(61) *¿Dónde está el tímpano?*

- 'localización física ambigua' (puede ser o no ser geográfica)

(62) *¿Dónde se ha celebrado alguna Conferencia Mundial de la Mujer?*

- 'localización no física':

(63) *¿Dónde puedo ver películas online gratis?*

DATOS OBTENIDOS

Trivial

Se documentan 6 casos en total.

En todos los casos el valor semántico para *dónde* es el de 'localización geográfica' de una entidad:

(64) *¿Dónde está el Mar de la Tranquilidad?*

Clef

Se documentan 21 casos en total con los siguientes valores:

- 'localización geográfica': = 19 casos:

(65) *¿Dónde está el archipiélago de Svalbard?*

- 'localización física ambigua' = 2 casos:

(66) *¿Dónde se entregan los Oscar?*

Wiki

En los 100 casos seleccionados se documentan los siguientes valores semánticos:

- 'localización geográfica' = 21 casos:

(67) *Dónde se ubica el nilo?*

- 'localización física no geográfica' = 76 casos:
  (68) *Dónde se origina el magma?*

Un caso particular en este sentido es el de la 'localización en un espacio físico ficticio'. En Wiki hay muchos casos de preguntas relativas a videojuegos, por lo que abundan las preguntas sobre la localización de distintas entidades en espacios ficticios:

(69) *Dónde encuentro a los tres regis en pokemon platino?*

Un subtipo específico de este valor es el de identificación del lugar en el que viven determinados animales ('hábitat'), que tiene 30 casos documentados:

(70) *Dónde viven los cisnes?*

- 'localización no física' = 3 casos:
  (71) *Dónde se puedo descargar inazuma eleven 2 en ingles o español?*

Conclusiones

La siguiente tabla recoge los valores semánticos documentados para *dónde* en los tres corpus:

	Trivial	Clef	Wiki	Total
L. Geográfica	6	19	21	46
L. Física no geográfica	0	0	76	76
L. Física ambigua	0	2	0	2
L. no física	0	0	3	3
TOTAL	6	21	100	127

Tabla 21: Valores semánticos documentados para *dónde* en los tres corpus.

Los valores más comunes son los de 'localización geográfica' (documentado en todos los corpus) y 'localización física no geográfica' (documentado solo en Wiki). Los valores para *dónde* que implican una ubicación no física son marginales.

La granularidad de la que hablábamos en el capítulo 2 afecta a todos los subtipos semánticos de *dónde*:

- 'localización geográfica':
  (72) *¿Dónde está el Hermitage? > En San Petesburgo / En Rusia.*

- ‘localización física no geográfica’:
  (73) *¿Dónde está el tímpano? > En el cuerpo / En el oído.*

- ‘localización física ambigua’ (puede ser o no ser geográfica):
  (74) *¿Dónde se ha celebrado alguna Conferencia Mundial de la Mujer? > En Europa / En París / En el Instituto X.*

- ‘localización no física’:
  (75) *¿Dónde puedo ver películas online gratis? > En Internet / En la página X.*

3.3.3.3.3 Valores semánticos de *quién*

En esta sección analizaremos los valores semánticos de *quién* en nuestros tres corpus. En el caso de Wiki trabajaremos otra vez sobre una selección aleatoria de 100 preguntas.

Los valores documentados son los siguientes:

- ‘Nombre propio de un individuo’:
  (76) *¿Quién fue el último premio Nobel de arquitectura?*

- ‘Nombre propio de un conjunto de individuos’:
  (77) *¿Quién hace el viagra?*

- ‘Ambigüedad entre nombre propio de individuo o de conjunto’:
  (78) *¿Quién cometió el atentado en el metro de Tokyo?*

- ‘Descripción de un nombre propio’:
  (79) *¿Quién es Arabella Kiesbauer?*

- ‘Nombre común’:
  (80) *Quién estudia a los fosiles?*

- ‘Definición de un conjunto’:
  (81) *Quiénes eran los mestizos?*

- ‘Identificación no específica’ (no queda claro si se pide nombre propio, nombre común, definición...):
  (82) *Quién era la maxima autoridad de la colonia?*

- ‘No humano’:
  (83) *Quiénes realizan la fotosintesis?*

DATOS OBTENIDOS

Trivial

Se documentan 4 casos en total.

En todos ellos el valor de la variable es el de 'nombre propio de un individuo':

(84) *¿Quién escribió El Diario de Ana Frank?*

Clef

En Clef hay 98 casos con *quién.*

Los valores documentados para la variable son:

- 'Nombre propio de un individuo' = 64 casos.
  (85) *¿Quién es el emperador japonés?*

- 'Nombre propio de un conjunto' = 1 caso.
  (86) *¿Quién otorga la Medalla Fields?*

- 'Ambigüedad entre nombre propio de individuo o de conjunto' = 2 casos.
  (87) *¿Quién fabricaba Windows 95?*

- 'Descripción de un nombre propio' = 30 casos.
  (88) *¿Quién era Bertha von Suttner?*

Wiki

Selección de 100 preguntas.

Los valores documentados para esas 100 preguntas son los siguientes:

- 'Nombre propio' = 68 casos:
  - de una o varias personas = 66 casos:
    (89) *Quién es el creador de ben10?*

  - de uno o varios conjuntos de humanos = 2 casos.
    (90) *Quiénes integran a la onu?*

- 'Nombre común' = 10 casos:
  (91) *Quién debe enseñar a la población a no contaminar las fuentes de agua?*

- 'Descripción de un nombre propio' = 9 casos.
  (92) *Quién fue alfred wegener?*

- 'Definición de un conjunto de seres humanos' = 4 casos.
  (93) *Quién eran los caudillos?*

- 'Identificación no específica de un ser humano' = 5 casos.
  (94) *Quién canta sin compromiso?*

- 'Identificación de entidades no humanas' = 4 casos.
  (95) *Quién rompe la materia en el citoplasma?*

Conclusiones

Los valores que se documentan para *quién* en los tres corpus son los siguientes:

	Trivial	Clef	Wiki	Total
N. Propio de un individuo	4	64	66	134
N. Propio de un conjunto	0	1	2	3
N. Propio ambiguo	0	2	0	2
Descripción de un nombre propio	0	30	9	39
N. Común	0	0	10	10
Definición de un conjunto	0	1	4	5
Identificación no específica	0	0	5	5
No humano	0	0	4	4
Total	4	98	100	202

Tabla 22: Valores semánticos documentados para *quién* en los tres corpus.

Los dos valores más documentados son los que recogíamos para *quién* en el capítulo 2: 'nombre propio de un individuo' y 'descripción de un nombre propio'. El primero se documenta en los tres corpus y el segundo en dos de tres.

El valor de 'nombre propio de conjunto' no se contemplaba en nuestra descripción teórica, donde solo se consideraba el valor de *quién* apuntando a uno o más individuos, pero no a un conjunto de seres humanos.

Resulta llamativo que el valor de 'descripción' solo se documente en Wiki en 9 casos, mientras que en Clef se documenta en 30 ocasiones. Esto se debe probablemente a que, como hemos señalado, el perfil de las preguntas de los corpus es diferente (preguntas factuales en Clef *vs*. preguntas de todo tipo en Wiki).

El valor de *quién* como 'no humano' solo se documenta en Wiki y de forma muy marginal. Creemos que el valor de *quién* como *+humano* es muy claro y que

estos usos se deben más bien a confusiones por parte de los hablantes que a una posible tendencia de uso que se está abriendo camino.

3.3.3.3.4 Valores semánticos de *cuál*

En lo que respecta a *cuál*, hemos analizado en nuestros tres corpus el tipo de entidad a la que apunta el interrogativo.

Hemos analizado todos los casos de Trivial (51 preguntas) y Clef (45 preguntas), y una selección de 100 preguntas de Wiki.

Los valores semánticos que hemos establecido y documentado para la entidad son los siguientes:

- 'Entidad concreta':
  (96) *¿Cuál es la mayor compañía de software del mundo?*

- 'Entidad abstracta':
  (97) *Cuáles son los principios fundamentales de la democracia?*

- 'Valor temporal':
  (98) *¿Cuál es la fecha de comienzo de la compañía de los alimentos de Hershey?*

- 'Cuantificación':
  (99) *¿Cuál es la altura del K2?*

En nuestro análisis hemos englobado bajo 'entidad concreta' distintos tipos de entidades, ya que lo que nos interesaba era este valor en contraposición al de 'entidad abstracta', pues la recuperación de una entidad concreta en un sistema de BR es más simple y factible que la de una entidad abstracta. Dentro de esta categoría de 'entidad concreta' se agrupan todo tipo de entidades:

(100) *¿Cuál es el dedo más sensible de la mano?*
(101) *¿Cuál es la ciudad de las góndolas?*
(102) *¿Cuál es el ingrediente activo en Tylenol?*
(103) *¿Cuál es la moneda irakí?*
(104) *Cuál es la sílaba tónica de campana?*
(105) *Cuáles son los hongos extintos de mexico?*

DATOS OBTENIDOS

Trivial

En Trivial se documentan 51 casos de *cuál* con los siguientes valores:

- 'Entidad concreta' = 50 casos.
  (106) *¿Cuál de los cinco sentidos es menos sensible después de haber comido mucho?*

- 'Entidad abstracta' = 1 caso.
  (107) *¿Cuál es la causa de cada muerte humana?*

Clef

En los 45 casos con *cuál* se documentan los siguientes valores:

- 'Entidad concreta' = 36 casos.
  (108) *¿Qué cargo detenta Ariel Sharon?*

- 'Entidad abstracta' = 1 caso.
  (109) *¿Cuál es la mejor manera de combatir las alergias?*
- 'Cuantificación' = 8 casos.
  (110) *¿Cuál es la distancia entre Braga y Guimarães?*

Wiki

Hemos seleccionado 100 casos al azar de todas las preguntas con *cuál*.
Los valores documentados son los siguientes:

- 'Entidad concreta' = 61 casos:
  (111) *Cuál es el color favorito de cece la de shake it up?*

- 'Entidad abstracta' = 34 casos:
  (112) *Cuál es la importancia de inah para mexico?*

- 'Cuantificación' = 5 casos:
  (113) *Cuál la superficie total que tiene el territorio peruano?*

Conclusiones

En la siguiente tabla mostramos un resumen de los datos obtenidos en los tres corpus:

	Trivial	Clef	Wiki	Total
E. Concreta	49	36	61	146
E. Abstracta	2	0	34	36
Valor Temporal	0	1	0	1
Cuantificación	0	8	5	13
Total	51	45	100	196

Tabla 23: Valores documentados en los tres corpus para la variable correspondiente a *cuál*.

Como queda claro por el análisis, *cuál* puede apuntar a valores semánticos bastante diferentes.

El valor más común con diferencia es el de 'entidad concreta', que se documenta de forma mayoritaria en los tres corpus.

A este valor le sigue el de 'entidad abstracta', que se documenta anecdóticamente en Trivial pero de forma importante en Wiki, mientras que en Clef no se documenta. El hecho de que no se documente en Clef y sí de forma tan numerosa en Wiki es lógico teniendo en cuenta que este valor es el propio de preguntas de opinión o elaboración de un argumento:

(114) *Cuál es la diferencia entre un guion de radio y un guion de teatro?*
(115) *Cuáles son los principios fundamentales de la democracia?*

Estas preguntas sí están presentes en un corpus como Wiki, que recoge todo tipo de preguntas de usuarios reales, pero no en Clef, que, al menos en principio, recoge solo preguntas de tipo factual. Este valor, por tanto, no parece esperable en un entorno de BR factual, pero sí en un sistema de BR que trabaje con preguntas abiertas.

El valor de cuantificación se documenta en los tres corpus y tiene un uso relativo. El temporal, sin embargo, solo se documenta en un corpus con un único caso.

3.3.3.3.5 Valores semánticos de *qué*

En esta sección nos ocupamos de los valores semánticos de *qué* documentados en nuestros tres corpus.

Como hicimos en el análisis de *cuál*, oponemos 'entidad concreta' a 'entidad abstracta' de modo genérico, pues lo que nos interesa es diferenciar ambos tipos de valores generales por su influencia en la mayor o menor dificultad a la hora de determinar el significado concreto de la pregunta y encontrar una respuesta adecuada para ella.

Como con otros interrogativos, hemos utilizado todos los casos de Trivial (73), y una selección al azar de 100 preguntas de Clef y Wiki.

Los valores documentados en nuestros tres corpus son los siguientes:

- 'Entidad concreta':
  (116) *¿Qué animal tiene más músculos: la oruga, el hombre o el elefante?*

- 'Entidad abstracta':
  (117) *¿Qué diferencia hay entre un fósforo común y uno de seguridad?*

- 'Cuantificación':
  (118) *Qué porcentaje de petroleo hay en mexico y en todo el mundo?*

- 'Valor temporal':
  (119) *¿Qué día fue asesinado Rajid Gandhi?*

- 'Definición':
  (120) *¿Qué es Linux?*

DATOS GENERALES OBTENIDOS

Trivial

En Trivial se documentan 73 casos con *qué*. Los valores documentados son:

- 'Entidad concreta' = 67 casos.
  (121) *¿Qué instrumento musical tiene nombre y forma geométricos?*

- 'Entidad abstracta' = 3 casos.
  (257) *¿Qué hacen los gorilas cuando se ponen nerviosos?*

- 'Cuantificación' = 1 caso.
  (122) *¿Qué altura aproximada tiene la Torre Eiffel parisina?*

- 'Temporal' = 2 casos.
  (123) *¿Qué año sigue al 356 antes de Cristo?*

Clef

En las 100 preguntas seleccionadas al azar, se documentan los siguientes valores:

- 'Entidad concreta' = 57 casos.
  (124) *¿Qué cargo detentaba Silvio Berlusconi?*

- 'Cuantificación' = 5 casos.
  (125) *¿Qué altura tiene el Kanchenjunga?*

- 'Valor temporal' = 7 casos.
  (126) *¿Qué año comenzó la Intifada?*

- 'Definición' = 31 casos:
  (127) *¿Qué es la UEFA?*

Wiki

En las 100 preguntas seleccionadas al azar en Wiki, se documentan los siguientes valores:

- 'Entidad concreta' = 58 casos.
  (128) *Qué países tienen las 4 estaciones?*

- 'Entidad abstracta' = 3 casos.
  (129) *Qué factores determinaron el desarrollo de la sabana en el oriente africano Editar?*

- 'Cuantificación' = 2 casos.
  (130) *Qué porcentaje de calor interno de la tierra hay en el mundo?*

- 'Definición' = 37 casos.
  (131) Qué es el sistema muscular?

Conclusiones

La siguiente tabla recoge los datos totales sobre los valores documentados en los tres corpus:

	Trivial	Clef	Wiki	Total
E. Concreta	67	57	58	182
E. Abstracta	3	0	3	6
Cuantificación	1	5	2	8
Valor Temporal	2	7	0	9
Definición	0	31	37	68
Total	73	100	100	273

Tabla 24: Valores documentados en los tres corpus para la variable correspondiente a *qué*.

Como ocurría con *cuál*, el valor de 'entidad concreta' es el más documentado con diferencia (y está presente además en todos los corpus). Por el contrario, el valor de 'entidad abstracta' es muy escaso.

El segundo valor más común es el de 'definición', presente en dos de los tres corpus. Como hemos visto ya, el valor de 'definición' es impreciso y complejo.

Tanto el valor de 'cuantificación' como el temporal son, junto a 'entidad abstracta', muy escasos.

3.3.3.3.6 Valores semánticos de *cómo*

En el caso de *cómo,* el procedimiento que hemos seguido para el análisis de los valores semánticos en nuestros corpus es un poco distinto.

Como vimos en la sección 3.2.1 del capítulo anterior, el valor semántico general para *cómo* es el de 'modo o manera'. Si a este valor general le sumamos el del verbo con el que se combina el interrogativo, obtenemos valores semánticos más precisos. Por ejemplo, en (272):

(132) *¿Cómo se llamaba el actor que hacía de Chanquete en Verano Azul?*

El resultado de combinar *modo* + *llamarse* es 'denominación'. La variable concreta por la que se pregunta en (132) es, por tanto, 'denominación'.

Con el fin de determinar si existen patrones de este tipo especialmente productivos, hemos planteado nuestro análisis de la siguiente forma: en primer lugar hemos extraído de cada corpus los verbos más comunes en las preguntas con *cómo* y hemos determinado el valor semántico de la combinación del interrogativo con cada verbo (sin tener en cuenta el resto de la pregunta, solo verbo e interrogativo). Para este primer análisis hemos utilizado todas las preguntas con *cómo* presentes en cada corpus. En segundo lugar, hemos realizado un análisis paralelo a los que venimos haciendo para los distintos interrogativos: hemos utilizado como máximo 100 preguntas de cada corpus y hemos determinado sus valores semánticos. A continuación presentamos cada uno de los análisis y los datos obtenidos.

3.3.3.3.6.1 Análisis de la variable mediante interrogativo + verbo

Extracción de los verbos

El proceso que hemos seguido para extraer todos los verbos de las preguntas con *cómo* es el siguiente:

1) Hemos extraído todos los verbos de todas las preguntas con *cómo* en cada uno de los corpus. Para ello hemos utilizado los corpus completos:

Trivial – 15 preguntas
Clef – 41 preguntas
Wiki – 21.992 preguntas

2) Hemos determinado el número total de casos por verbo en cada corpus.

3) Hemos analizado los valores semánticos de estos verbos + el valor 'modo', 'manera' aportado por *cómo*.

Para este último paso no hemos tenido en cuenta todos los verbos documentados en los tres corpus. En el caso de Trivial y Clef, hemos analizado solo los verbos con más de una aparición (cf. Apéndice 1). Con Wiki hemos procedido de otra manera: como la lista de verbos documentados llegaba a 739, hemos seleccionado para el análisis los 100 verbos con más casos (cf. Apéndice 1).

Verbos documentados en los corpus

Una ojeada a los Apéndices 2 y 3 revela que el único verbo que se documenta en los tres corpus es además el verbo con más casos en cada uno de ellos: *llamarse*. Ningún otro verbo se documenta en los tres corpus.

Por otra parte, todos los verbos documentados en Trivial y Clef se documentan en Wiki con bastantes casos.

Valores semánticos

A partir del análisis de la combinación de verbos más comunes con *cómo*, los valores que hemos documentado son los siguientes:

- 'Denominación' (se pregunta por nombre o palabra):
  (133) *Cómo se llama al conjunto de gatos?*
  Incluye los casos en los que se pregunta por una palabra en otra lengua, por ejemplo:
  (134) *Cómo se dice abuelo en mixteco?*

- 'Procedimiento':
  (135) *Cómo doy de baja a mi cuenta de yahoo?*

- 'Descripción':
  (136) *Cómo son los organelos unicelulares?*

- 'Modo', 'manera' (no preciso):
  (137) *¿Cómo murió Adolf Hitler?*

Hemos mantenido un valor general de 'modo', 'manera' para aquellos casos en los que no era posible establecer un valor más concreto.

Datos por corpus

Trivial

Se documentan dos valores:

- 'Denominación' = 11 casos.
  (138) *¿Cómo se llama el objeto que se entrega en una carrera de relevos?*

- 'Modo' = 1 caso.
  (139) *¿Cómo se transmiten más rápidamente las ondas sonoras, a través del agua o del aire?*

Clef

Se documentan los siguientes valores:

- 'Denominación' = 21 casos.
  (140) *¿Cómo se llama a la gente que hace fuegos artificiales?*

- 'Procedimiento' = 12 casos.
  (141) *¿Cómo se determina la proporción de deuterio característica de cada vino?*

- 'Descripción' = 1 caso.
  (142) *¿Cómo son los peces espada?*

- 'Modo' = 10 casos.
  (143) *¿Cómo murió Juvénal Habyarimana?*

Wiki

Se documentan los siguientes valores:

- 'Procedimiento' = 9.188 casos.
  (144) *Cómo edito el nombre y el diseño de mi wikia?*

- 'Denominación' = 6.299 casos.
  (145) *Cómo se la llama al conjunto de avestruces?*

- 'Descripción' = 1.758 casos.
  (146) *Cómo son en realidad las estrellas?*

- 'Modo' = 152 casos.
  (147) *¿Cómo crees que acabe lost?*

3.3.3.3.6.2 Análisis detallado de la variable en cada corpus

Con el fin de ser más exhaustivos en nuestro análisis de la variable de *cómo*, hemos procedido a un estudio detallado de las variables por corpus como el que hemos realizado para el resto de los interrogativos (cf. *supra*).

En el caso de Trivial y Clef hemos analizado todas las preguntas con *cómo* (15 y 41, respectivamente); en el caso de Wiki, se han seleccionado 100 preguntas.

Las variables que hemos documentado en este análisis exhaustivo son las mismas que en el análisis general por verbos: 'denominación', 'modo', 'descripción' y 'procedimiento'. A continuación detallamos los datos obtenidos en cada corpus.

Datos de corpus

Trivial

En las 15 preguntas con *cómo* hemos documentado los siguientes valores:

- 'Denominación' = 14 casos.
  (148) *¿Cómo se llama el caballo alado de la mitología griega?*

- 'Modo' = 1 caso.
  (149) *¿Cómo se transmiten más rápidamente las ondas sonoras, a través del agua o del aire?*

Clef

En las 41 preguntas con *cómo* documentadas en el corpus aparecen los siguientes valores:

- 'Denominación' = 21 casos.
  (150) *¿Cómo se llamaba el barco de Cousteau?*

- 'Procedimiento' = 9 casos.
  (151) *¿Cómo se pretende llevar a cabo en Perú la planificación familiar?*

- 'Modo' = 10 casos.
  (152) *¿Cómo se autodefendió Vladimir Zjirinovski contra los manifestantes en Estrasburgo?*

- 'Descripción' = 1 caso.
  (153) *¿Cómo son los peces espada?*

Wiki

En las 100 preguntas de Wiki seleccionadas al azar se documentan los siguientes valores semánticos:

- 'Procedimiento' = 51 casos.
  (154) *Cómo funciona una estufa solar?*

- 'Descripción' = 10 casos.
  (155) *Cómo son los pulmones de la ballena?*

- 'Denominación' = 36 casos.
  (156) *Cómo se llama el gobernador de el estado amazonas?*

- 'Modo' = 3 casos.
  (157) *Cómo se clasifican los conjuntos en termino matematico?*

Conclusiones

En la tabla que sigue recogemos los datos totales del análisis de la variable por verbos en los tres corpus:

	Trivial	Clef	Wiki	Total
Denominación	11	21	6.299	6.331
Descripción	0	1	1.758	1.759
Procedimiento	0	12	9.188	9.200
Modo, manera	1	1	152	154
TOTAL	12	35	17.397	17.444

Tabla 25: Análisis de la variable de *cómo* a través de los verbos en los tres corpus.

Podemos observar que los cuatro valores se documentan en dos de los tres corpus. Este hecho parece un argumento a favor de la idea de establecer patrones sintácticos a partir de la combinación *cómo* + determinado verbo, asociando cada valor semántico a unos determinados patrones.

Respecto a los valores, observamos que el más común es un valor bastante abstracto e impreciso, 'procedimiento'. Le sigue en importancia un valor concreto, 'denominación'. 'Descripción' se aleja de los dos anteriores, aunque también

presenta un número de casos importante; se trata nuevamente de un valor impreciso y abstracto. Finalmente, 'modo' presenta pocos casos en comparación con los otros tres valores.

Recordemos que un único verbo se documenta en los tres corpus, siendo además el verbo con más casos: *llamarse*. Este patrón es, por tanto, especialmente productivo en las preguntas con *cómo*.

Podemos comparar este análisis con el análisis detallado de la variable en los tres corpus, resumido en la siguiente tabla:

	Trivial	Clef	Wiki	Total
Denominación	14	21	36	71
Descripción	0	1	10	11
Procedimiento	0	9	51	60
Modo, manera	1	10	3	14
TOTAL	15	41	100	156

Tabla 26: Valores semánticos documentados para *cómo* en los tres corpus.

Como muestra la tabla, el análisis detallado muestra algunas divergencias con el análisis general por verbos.

En primer lugar, el valor más documentado es ahora 'denominación' y no 'procedimiento'. Por otro lado, 'modo' se documenta más que 'descripción'.

En segundo lugar, en el análisis detallado se documentan en los tres corpus dos de los cuatro valores.

Pese a las divergencias, lo que nos parece interesante es que los valores semánticos documentados en el análisis exhaustivo son los mismos y que su frecuencia es similar.

3.3.3.3.7 Valor de sugerencia para *por qué + no*

Como vimos en el capítulo 2, existe un valor de sugerencia para *por qué* cuando se combina con el adverbio *no*:

(158) *¿Por qué no invertir en bolsa?*

En nuestros tres corpus completos se documentan solo dos casos con *por qué no* que podrían tener el valor de sugerencia, ambos en Wiki:

(159) *Porque los que preguntan sobre pokemon mejor no se meten en la wikidex?*

(160) *¿Por qué no vais poniendo las imagenes de los pokemonde corazon oro y alma plata?*

En (159) la presencia de «mejor» reduce la interpretación a 'sugerencia' (negación externa). En (160) además del valor de sugerencia (negación externa), se podría interpretar el valor causal normal (negación interna).

Los datos muestran por tanto que este uso es muy extraño en preguntas como las de nuestros corpus, lo que sugiere que este valor no es esperable en un entorno de BR (a menos que este maneje sugerencias).

3.4 Conclusiones generales del capítulo

3.4.1 Orden de constituyentes

3.4.1.1 Orden de constituyentes en las totales

Se documentan dos órdenes:

- verbo + sujeto + argumentos;
- sujeto + verbo + argumentos.

El segundo orden (marcado) es más común en Wiki, lo que nos hace pensar que este orden sujeto+verbo puede ser tan común en las preguntas totales como el no marcado.

3.4.1.2 Orden de constituyentes en las parciales

El orden que hemos denominado prototípico es claramente el más común.

Los órdenes alternativos tienen muy baja incidencia.

En una muestra aleatoria de 100 preguntas en los tres corpus, no se documenta ningún tipo de anteposición.

En una muestra aleatoria de 200 preguntas en Wiki, se documenta solo la anteposición tipo B y no la tipo A.

En la anteposición tipo B, la estructura más común es aquella en la que el interrogativo funciona como adjunto y el sujeto como argumento antepuesto.

3.4.2 Negación

La negación se documenta poco, y es mayoritariamente de tipo interno.

El interrogativo con el que más se documenta la negación es *por qué*. Teniendo en cuenta que este es el interrogativo menos común en los tres corpus, este dato parece indicar que la negación en las preguntas es mucho más frecuente con *por qué*.

3.4.3 Partículas interrogativas

3.4.3.1 Incidencia de los interrogativos por corpus

El interrogativo más documentado es *qué* y el menos documentado *por qué*.

En términos generales, la jerarquía en la incidencia en corpus es la siguiente: *qué*, *cómo*, *cuál*, *cuánto*, *dónde*, *quién*, *cuándo*, *por qué*. Los tres interrogativos más documentados son los que presentan más dificultades en la determinación del valor de su variable semántica.

3.4.3.2 Preposiciones

Qué es el interrogativo que admite más preposiciones; *cuándo*, el que menos (solo una, *desde*).

Dónde se documenta combinado con *en*.

3.4.3.3 Valores semánticos de las partículas interrogativas

Cuándo

El valor dominante es el temporal, generalmente, 'fecha'/'año'.

El valor hipotético se documenta solo en Wiki.

El tiempo del verbo se relaciona con determinados valores semánticos: el valor hipotético siempre va en presente; el valor de fecha, generalmente, en pasado. El valor temporal puede combinarse con pasado, presente o futuro.

Complejidad a la hora de determinar el valor temporal. En esta tarea pueden ser útiles patrones léxico sintácticos tipo *cuándo* + *nacer.*

Dónde

'Localización física' es el valor más común y, como subtipo de este, 'localización geográfica'.

La granularidad afecta a todos los valores con *dónde*, de manera que el significado de este interrogativo es siempre impreciso y es necesario recurrir a factores externos a la propia pregunta (usuario, características del sistema) para poder concretarlo.

Quién

Los valores más comunes son 'nombre propio' y 'descripción'. Ambos valores son deducibles por análisis sintáctico más reconocimiento de entidades (NER).

Cuál

Al ser un elemento deíctico vacío de significado puede apuntar a cualquier valor.

En nuestros tres corpus el valor más común es el de 'entidad concreta' (que puede tener varios valores específicos distintos).

Qué

Al igual que *cuál*, presenta varios valores semánticos posibles.

El valor de 'entidad concreta' es también el más común; le sigue el de 'definición'.

Cómo

Documentamos cuatro valores (tanto en el estudio por verbos como en el estudio exhaustivo): 'procedimiento', 'denominación', 'modo, manera' y 'definición'. Los dos primeros son los más comunes y 'definición' el que menos se documenta (en los dos estudios; lo que cambia es la frecuencia de 'procedimiento' y 'denominación').

Los resultados similares del estudio por verbos y del estudio exhaustivo para determinar los posibles valores semánticos de *cómo* parecen apoyar la posibilidad de establecer *clusters* de verbos que se relacionen con cada uno de estos valores (y otros posibles, en caso de haberlos).

Por qué + no

El valor de 'sugerencia' no parece relevante para el perfil de preguntas de nuestros corpus.

Capítulo 4

SpQA

En los capítulos anteriores nos hemos ocupado de los tres ámbitos a partir de los cuales se construye SpQA:

- las necesidades del análisis de la pregunta en un sistema de BR;
- las características de las preguntas en español señaladas por trabajos teóricos;
- las características de las preguntas extraídas de nuestro análisis de corpus.

Estos tres aspectos confluyen para dar forma a SpQA, de cuya descripción nos ocuparemos en el presente capítulo.

En términos generales, SpQA es un *parser* específico para el análisis de preguntas en español en un entorno de BR. Debido a esta especificidad, su principal objetivo es (dentro de los límites del análisis sintáctico semántico que produce) extraer de una pregunta la máxima cantidad de información lingüística relevante para la obtención de una respuesta correcta en un sistema de BR. La cobertura (entendida como la capacidad del sistema para analizar el mayor número de estructuras lingüísticas posibles) y la robustez (entendida como la capacidad del sistema para tratar, al menos parcialmente, estructuras con errores o contempladas solo en parte en la gramática) son también objetivos de SpQA.

El análisis de SpQA maneja información sintáctica y semántica. El análisis sintáctico, definido como la identificación y etiquetado de las relaciones de dependencia de la pregunta[105], es el objetivo básico de SpQA. El análisis semántico se ciñe a la representación de la variable o incógnita presente en la pregunta y se construye sobre información léxica y sintáctica.

En las secciones que siguen nos ocuparemos de la descripción exhaustiva de SpQA. En primer lugar, presentaremos el formalismo con el que se construye, AGFL. A continuación, describiremos brevemente el tipo de *parser* que es SpQA. El siguiente paso será la descripción detallada del modelo de análisis utilizado y de la gramática formal sobre la que se construye el *parser*.

105 Como veremos, el análisis de SpQA es dependencial. No obstante, la representación que genera el *parser* permite también extraer información relativa a los constituyentes oracionales de la pregunta.

4.1 El formalismo AGFL

AGFL[106] (*Affix Grammars over a Finite Lattice*) es un formalismo para la descripción de gramáticas libres de contexto asociado a un generador de analizadores.

Los motivos para la elección del formalismo AGFL en la construcción de SpQA son varios.

En primer lugar, AGFL permite, de forma sencilla e intuitiva, la construcción de gramáticas formales. No es necesaria formación en programación u otras áreas de la informática para poder abordar el formalismo. Esto lo hace accesible para cualquier persona que haya adquirido las bases de su funcionamiento. En palabras de sus propios desarrolladores:

> The user of the AGFL-system needs to have little knowledge about the mechanism of parsing, and does not need to be aware of transformations and optimalizations applied to the grammar; the only thing that counts is the language described by the grammar. (Koster, Seutter, y Seibert, 2008, p. 10)

En segundo lugar, el formalismo presenta una serie de características que lo hacen especialmente interesante para la construcción de analizadores sintácticos. Las principales son (hay más, pero no las citamos, por ser menos relevantes):

- Permite distintos modos de análisis: por oraciones, por segmentos (el sistema analiza el mayor segmento que pueda reconocer), etc.

- Posibilita distintas salidas para el análisis: en forma de árbol de constituyentes o en la forma que el autor de la gramática especifique en sus reglas (mediante *transduction;* volveremos sobre esto más adelante). De este modo, el autor de la gramática puede construir su propio modelo de análisis, según las características que más le convengan (en el caso de SpQA, adaptando ese modelo a las necesidades de un sistema de BR).

- Da la opción de que la salida muestre el mejor análisis (según lo especificado en la gramática), los *n* mejores análisis o todos los análisis para un *input* determinado.

- Permite la extracción de tripletes de dependencias a partir de un tipo concreto de representación del análisis sintáctico (volveremos sobre esto más abajo).

- Presenta facilidades en la construcción de lexicones.

106 El sistema AGFL ha sido desarrollado en el Computer Science Department de la Radboud University of Nijmegen por el equipo liderado por el profesor C.H.A. Koster.
En esta sección presentaremos muy sucintamente el formalismo. Para una descripción más detallada de este y de todo el sistema AGFL, cf.: http://www.agfl.cs.ru.nl/

- Permite la inserción de *fact tables* (cf. *infra*), que posibilitan agregar de un modo sencillo información relativa a, por ejemplo, la subcategorización verbal.

- Posibilita la construcción de gramáticas modulares.

Finalmente, un claro punto a favor de AGFL para la construcción de SpQA, fue la experiencia del grupo de investigación en el que se inserta esta tesis con dicho formalismo[107], además de la estrecha colaboración con el equipo que lo desarrollaba en Nijmegen.

En las secciones siguientes describiremos sucintamente el formalismo y sus principales características.

4.1.1 Reglas en AGFL

AGFL pertenece a la familia de las gramáticas de dos niveles: un primer nivel libre de contexto es extendido en un segundo nivel mediante atributos (*affixes*) que expresan concordancia entre constituyentes sintácticos.

En el primer nivel, las reglas de la gramática son sencillamente reglas libres de contexto, como podemos ver en el ejemplo:

(1)
```
RULE sentence:
  subject, verb.
```

Los elementos que pueden aparecer a izquierda y derecha de la regla se denominan no terminales[108]. Un no terminal es identificado mediante su nombre (*subject*[109], por ejemplo) y el número de atributos que lo determinan (su *arity*). Los valores de dichos atributos (no más de treinta y dos) pueden anidarse y se hacen explícitos en el segundo nivel, mediante reglas que tienen la siguiente forma:

(2)
```
NUMBER :: singular | plural
```

Esta regla define el atributo *NUMBER*[110] como la unión ("|") de los valores *singular*[111] y *plural*.

La combinación de estos dos niveles permite la elaboración de reglas libres de contexto extendidas con atributos que permiten expresar relaciones entre los no terminales como, por ejemplo, la concordancia entre categorías sintácticas:

107 De hecho, como veremos (cf. *infra*), el núcleo inicial de desarrollo de la gramática de SpQA fue una gramática desarrollada en el grupo de Gramática del Español por la D.ª M.ª Paula Santalla del Río.

108 En las gramáticas formales que tratan lenguas naturales los no terminales son generalmente funciones o categorías que constituyen los nodos de los árboles generados de acuerdo con la gramática por medio del analizador o *parser*.

109 Los no terminales se escribirán siempre en cursiva.

110 Los nombres de los atributos se presentarán siempre en mayúscula y cursiva.

111 Los valores de los afijos se muestran en minúsculas y cursiva.

(3)
RULE sentence:
  subject (NUMBER), verb (NUMBER).

Transduction

AGFL permite especificar una *transduction* (en negrita en el ejemplo) para cada una de las reglas:

(4)
Noun (GENDER, NUMBER):
  n (LEMMA, NTYPE, GENDER, NUMBER)
    / "N:", n.

La *transduction* determina una representación concreta del no terminal definido en la regla. Esa representación puede recoger toda la información que aparece en el cuerpo de la regla o simplificarla. En el ejemplo, la información de la regla se simplifica, y la *transduction* solo especifica que el no terminal *Noun* debe ser representado en la salida del *parser* como `N:`[112] más aquello en lo que se reescriba el no terminal *n*.

4.1.2 Salidas del analizador

AGFL permite dos salidas para el analizador: una en forma de árbol de constituyentes (6) y otra que será la especificada en la *transduction* (7). Tomemos la regla de la sección anterior:

(5)
Noun (GENDER, NUMBER):
  n (LEMMA, NTYPE, GENDER, NUMBER)
    / "N:", **n.**

Para la entrada *ventana* las dos salidas[113] de SpQA serían:

112 Los elementos pertenecientes a la representación por *transduction* se marcarán con letra tipo Courier.
113 Todos los ejemplos de análisis se mostrarán en letra Courier 10 Pitch.

(6)
```
# (null) 0 0-7|ventana
PHRASES / @ "\n"@
  PHRASE / @
   NP / @
    NP(third, fem, sing, CASE) / @
     Noun Phrase(third, fem, sing, CASE) / @
     $PENALTY(8)
      Noun Part(DEF, fem, sing, CASE) / @
        Noun Kernel(fem, sing, CASE) / @
         Noun(fem, sing) / "N:" !
          n("ventana", common, fem, sing) "ventana" [1]
            $PENALTY
  $PENALTY(10)
```

(7)
```
# (null) 0 0-7|ventana
N: ventana
```

El análisis de (6) es más detallado y muestra, a menos que se indique lo contrario en la gramática[114], toda la información especificada en las reglas. El análisis de (7) muestra solo aquello que se especifica en la *transduction* de la regla correspondiente.

4.1.3 Uso de penalties: *Best-Only Parsing*

Una de las características más interesantes de AGFL es que permite dar prioridad a unos análisis respecto a otros mediante la adición de *penalties* a las reglas que componen la gramática. Un *penalty* tiene la siguiente forma:

$PENALTY(n)

donde *n* es cualquier número: cuanto más alto sea este número, menor probabilidad tiene la regla o alternativa de ser seleccionada por el sistema para un análisis concreto.

La utilización de *penalties* junto con las frecuencias en los lexicones (cf. *infra*) determinan que ciertos análisis se definan como «mejores» que otros para una determinada secuencia lingüística. Esta es la base de la técnica de AGFL conocida como *Best-Only Parsing* (Koster, Seutter y Seibert, 2007) (el destacado es nuestro):

> What is "best" depends ideally on the semantics of the sentence, but there is no (useful) way to do this automatically. A second best is to take the most probable one, given some probability distribution of sentences. This is what motivates most forms of probabilistic

114 AGFL cuenta con un mecanismo que permite ocultar niveles de análisis en la salida del *parser*.

> parsing. […] A**s the best analysis, the parser takes the longest segment with the lowest penalty level.**
> (Koster et al., 2007, p. 3)

Dada una estructura como *input*, AGFL permite, para la salida del *parser*, seleccionar solo el mejor análisis (*Best-Only*), *n* análisis entre los mejores o todos los análisis posibles. Esto posibilita conocer todos los análisis posibles de esa estructura (lo que puede ser interesante para estudiar la ambigüedad) o solo aquel o aquellos *n* análisis determinados como «mejores» por la gramática.

4.1.4 Arquitectura de las gramáticas en AGFL

Las gramáticas escritas con AGFL constan de un *encabezamiento* y un *cuerpo*. El encabezamiento identifica la gramática con un nombre, mientras que el cuerpo explicita la relación con otras gramáticas y con otros componentes como lexicones y *fact tables*, especifica y define los diferentes objetos (no terminales) manejados y contiene la raíz (*root*) de la gramática. El nombre de la gramática tiene que ser el mismo que el nombre del archivo (con extensión .*gra*) que la contiene. En el siguiente ejemplo podemos observar estos elementos:
(8)

```
GRAMMAR example.                          #  name of the grammar
USES subj, verb.                          #  grammars used

LEXICON adj, noun, adv, misc, verb        #  lexicon files
DEFINES
   ADJE(GRAD),                            #  lexicon nonterminals
   ADJE(GRAD,PREP),
   ADJE_TO(GRAD),
   ADVB,
   NOUN(NUMB),
   DIMENSION,
   VERBI(PREP,TRAN),
   VERBS(PREP,TRAN),
   VERBV(PREP,TRAN),
   VERBG(PREP,TRAN),
   VERBP(PREP,TRAN).

ROOT sentence.                            #  root of the grammar

RULE subject (NUMBER).                    #  specification
subject (NUMBER): pronoun (NUMBER).       #  syntax rule
NUMBER :: plural | singular.              #  affix rule
```

La posibilidad de asociar esta gramática raíz con otras «sub-gramáticas» y lexicones o *fact tables* permite crear de una forma cómoda y sencilla gramáticas modulares.

4.1.5 Lexicones

AGFL también permite la construcción y manejo de lexicones en la gramática. Los lexicones contienen terminales y sus descripciones y tienen extensión *.dat*. Pueden ocuparse, como en el caso de SpQA, de la información morfosintáctica asociada a las palabras. A continuación presentamos un pequeño fragmento del lexicón de sustantivos de SpQA:

(9)

"abrigo"	N ("abrigo", common, masc, sing)
"ábrigo"	N ("ábrigo", common, masc, sing)
"abrigos"	N ("abrigo", common, masc, plu)
"ábrigos"	N ("ábrigo", common, masc, plu)
"abriles"	N ("abril", common, masc, plu)

Tomemos la primera entrada, «abrigo»: la forma de la palabra se sitúa en la columna de la izquierda. En la columna de la derecha se sitúa, en primer lugar, la clase de palabra (*N*), seguida de su lema («abrigo»), y determinadas características de la categoría en cuestión (en este ejemplo, categoría del nombre, género y número) que se corresponden con afijos que se han definido en la gramática (*common* sería un valor del afijo *NTYPE*; *masc* sería un valor de *GENDER* y *plu* sería un valor de *NUMBER*).

AGFL permite además incluir datos sobre frecuencias léxicas en las entradas de los lexicones, información que, junto con los *penalties* en las reglas (cf. *supra*), determina prioridades en el análisis. Esta característica permite lidiar, por ejemplo, con homógrafos que puedan generar ambigüedad léxica.

4.1.6 *Fact tables*

Desde la versión 2.8 AGFL permite otro componente en sus gramáticas denominado *fact tables*:

> From version 2.8 of AGFL, grammars may specify fact tables. A fact is a special nonterminal that never generates output i.e. is a condition, that is implemented through a fast lookup mechanism. It will take a number of input affixes from the flat domains INT and TEXT to produce other affix values. This lookup mechanism has been implemented as part of the lexicon mechanism. A typical application of fact tables should be in stemming.
> (Koster el al., 2008, p. 13)

Las *fact tables* permiten, por tanto, establecer condiciones o restricciones en la gramática de un modo eficaz. Más adelante volveremos sobre ellas al describir la gramática de SpQA.

4.2 Breve descripción técnica del *parser*[115]

SpQA es un *parser/transducer*[116] (Koster, 2011) basado en reglas (Carroll, 2005). Dichas reglas están escritas en el formalismo AGFL y en conjunto constituyen una gramática de atributos[117]. El generador de analizadores del sistema AGFL compila esta gramática y genera el *parser*.

En Koster et al. (2008, p. 10) se nos da una descripción técnica del tipo de *parsers* que genera AGFL:

> The parsers generated by the AGFL-system are top-down recursive backup parsers [Koster, 1974], [Meijer, 1986] implementing nondeterministic top-down parsing [Nederhoff, 1993] using continuations and positive memoization. This parsing method is capable of dealing with (slightly restricted) left-recursive grammars. It can produce, on demand, one or all parsings for ambiguous input. […] Especially when using Best-Only parsing, the generated parsers can be very fast.

El analizador generado es por tanto un *top-down chart parser* que usa la heurística *Best-Only* (Koster et al., 2007) (cf. *supra*).

4.3 El análisis de SpQA

Hemos visto al presentar el formalismo que AGFL permite dos tipos de salida: una en forma de árbol de constituyentes, que recoge toda la información de las reglas de la gramática, y otra que corresponde a la representación especificada en la *transduction*.

En SpQA construimos la representación del análisis sintáctico semántico por medio de la *transduction*. En los párrafos que siguen detallaremos las características de este modelo de representación y los motivos por los que lo hemos elegido para nuestro sistema.

4.3.1 Representación en SpQA: grafo de dependencias

El tipo de representación construido a partir de *transduction* en SpQA consiste en un grafo (10) que expresa relaciones de tipo dependencial:

> a directed acyclic graph whose nodes are marked with words and whose arcs are marked with relators. Each arc is directed from one node (its head) to another (its tail). (Koster, 2011)

115 En esta sección aportamos algunos datos técnicos sobre el *parser* y el modo en el que se genera. Para detalles más concretos cf. (Koster, 2011), donde se describe el *parser* para el holandés DUPIRA, construido con el mismo formalismo.

116 A partir de ahora, por simplificar la exposición, nos referimos a SpQA solo como *parser* o analizador.

117 http://en.wikipedia.org/wiki/Attribute_grammar

(10)
```
# (null) 0 0-12|Bisbal canta
[V: cantar third sing present <SUBJ PN: Bisbal]
```

En el grafo del ejemplo, los nodos son `V: cantar` y `PN: Bisbal`, y el *relator* es `<SUBJ`. El *relator* establece el tipo de relación sintáctica entre los nodos mediante una etiqueta (*SUBJ* = sujeto), y la dirección de la dependencia mediante los signos "<", ">" (el símbolo apunta hacia la *head*). En el grafo del ejemplo se nos dice, por tanto, que `PN: Bisbal` es dependendiente de `V: cantar` en una relación de sujeto.

A partir de este grafo, AGFL permite la extracción de tripletes de dependencias (11):

> A dependency graph (or tree) can be unnested into a set of dependency triples, one of each arc in the graph. By a dependency triple we mean a triple of the form <head, relator, tail>.
> (Koster, 2011)

(11) *Bisbal canta.*
```
[V: cantar third sing present, SUBJ, PN: Bisbal]
```

Este modelo de representación del análisis está basado en las ideas del profesor C.H. Koster[118] (Koster, 2011)[119]. Podemos encontrar este mismo modelo de representación en otros *parsers* construidos con AGFL, como, por ejemplo, en el *parser*

para el holandés DUPIRA[120] (Koster, 2011) o en el *parser* para el inglés EP4IR[121] (Koster et al., 2007).

Los elementos de los que consta el grafo son los siguientes:

- **Nodos**: son los elementos entre los que se establecen las relaciones sintácticas; se corresponden (generalmente) con las distintas clases de palabras contempladas en la gramática[122].

El grafo permite la inclusión de información relativa a atributos de los nodos. En (12), por ejemplo, el grafo incluye los valores de los atributos de persona, número y tiempo del verbo *cantar* (en negrita en el ejemplo):

118 http://www.cs.ru.nl/~kees/
119 Por supuesto no nos referimos al modelo dependencial, que podemos encontrar en muchos *parsers*, sino al grafo de dependencias entendido y construido tal y como detallamos en este capítulo.
120 http://www.agfl.cs.ru.nl/DUPIRA/index.html
121 http://www.agfl.cs.ru.nl/EP4IR/index.html
122 Al final de la sección recogemos todas las clases de palabras que pueden funcionar como nodos en SpQA.

(12) `# (null) 0 0-12|Bisbal canta`
`[V: cantar` **`third sing present`** `<SUBJ PN: Bisbal]`

Entre dos corchetes, se establece siempre como *head* aquel nodo no dependiente de otros nodos hacia el que apuntan los *relators* de los restantes nodos del grafo (en el ejemplo anterior, `V: cantar`). Volveremos sobre esto más abajo.

- ***Relators***: establecen la función sintáctica y la dirección de la dependencia. Las funciones sintácticas[123] expresadas por los *relators* se sitúan al nivel de la oración (13) y al nivel de la frase (14).

(13) `# (null) 0 0-20|Ratzinger renunció.`
`[V: renunciar third sing past` **`<SUBJ`** `PN: Ratzinger]`

TRIPLETES
`[V: renunciar third sing past, SUBJ, PN: Ratzinger]`

(14) `# (null) 1 0-17|hombre hermoso`
`[N: hombre` **`<ATTR`** `A: hermoso]`

TRIPLETES
`[N: hombre, ATTR, A: hermoso]`

El *relator* puede ir colocado tanto a la izquierda (14) como a la derecha del nodo (15).

`# (null) 2 0-14|hombre hermoso`
`[A: hermoso` **`>ATTR`** `N: hombre]`

TRIPLETES
`[N: hombre, ATTR, A: hermoso]`

- **Corchetes**: los corchetes delimitan el ámbito de las dependencias en el grafo. A nivel oracional se usan para marcar los límites de las cláusulas (16), (17).

123 Al final de la sección recogemos todas las funciones sintácticas de SpQA.

(16) `# (null) 2 0-26|Rubalcaba censuró a Rajoy.`
`[V: censurar third sing past <OBJ PN: Rajoy <SUBJ PN: Rubalcaba]`

(17) `# (null) 4 0-32| Quiero que compres manzanas hoy.`
`1[V: querer first sing present <OBJ 2[V: comprar second sing present <OBJ N: manzanas <CIRC X: hoy]2 ]1`

TRIPLETES
`[V: comprar second sing present, CIRC, X: hoy]`
`[V: comprar second sing present, OBJ, N: manzanas]`
`[V: querer first sing present, OBJ, V: comprar second sing present]`

En (17), los corchetes marcados con **1** establecen un primer nivel en el que la *head* es `V: querer`, ya que este es el único nodo no dependiente de otros y el terminal hacia el que apuntan el resto de nodos mediante el indicador de dependencia <. A su vez, los corchetes marcados con un **2** establecen un segundo ámbito o plano de dependencias. En ese plano, el nodo que no depende de otros es `V: comprar`, y hacia él apuntan el *OBJ* y el *CIRC* mediante <. Al estar marcados los ámbitos de acción de cada *head*, los dependientes se asignan correctamente.

Al nivel de la frase, se utilizan siempre que hay estructuras con más de una posible *head* para una dependencia (18) (frases nominales modificadas por frases preposicionales con otra frase nominal en su interior, frases nominales modificadas por frases adjetivas con modificación, etc.).

(18) `# (null) 5 0-20|El boli de la chica.`
`1[N: boli <MODde 2[N: chica <DET D: la]2 <DET D: el]1`

TRIPLETES
`[N: boli, DET, D: el]`
`[N: boli, MODde, N: chica]`
`[N: chica, DET, D: la]`

En el ejemplo, sin los corchetes señalados con el **2**, los dos determinantes se asociarían a la misma *head*, *boli*:

(19) `*[N: boli <MODde N: chica <DET D: la <DET D: el]`

TRIPLETES
[N: boli, DET, D: el]
[N: boli, DET, D: la]
[N: boli, MODde, N: chica]

En los ejemplos anteriores se puede observar que en el grafo de dependencias las relaciones sintácticas se establecen entre palabras[124]. Sin embargo, como veremos más adelante en detalle, hay casos en SpQA en los que ciertas palabras de la oración no se muestran en el análisis (como el nexo «que» en las cláusulas subordinadas sustantivas (20), cf. 4.2.3.2.3), o bien no funcionan como nodos sino como parámetros del *relator*, como en el caso de las preposiciones (21) (cf. 4.2.2.5).

(20) # (null) 2 0-56|Bárcenas dejó **que el periódico publicase sus papeles.**
[V: dejar third sing past <OBJ **[V: publicar second sing imperfect <OBJ [N: papeles <DET D: sus] <SUBJ [N: periódico <DET D: el]]** <SUBJ PN: Bárcenas]

(21) # (null) 3 0-23|Rajoy huyó **de Génova.**
[V: huir third sing past **<CIRCde** PN: Génova <SUBJ PN: Rajoy]

Por otra parte, aunque las relaciones de dependencia se establecen entre palabras, la presencia de los corchetes en el grafo permite delimitar los límites de los constituyentes a nivel oracional:

(22) # (null) 0 0-31|El rey tropezó en la reunión.
[V: tropezar third sing past <**CIRC**en **[**N: reunión <DET D: la**]** <**SUBJ [**N: rey <DET D: el**]**]

4.3.2 Normalizaciones en la representación

El análisis por *transduction* posibilita ciertas normalizaciones en la representación de las estructuras sintácticas. Un ejemplo de estas normalizaciones sería la «despasivización», que consiste en pasar a voz activa estructuras en voz pasiva. De esta manera, (23) y (24) tienen la misma representación en SpQA (25):

(23) *El juez acusó a Bárcenas.*
(24) *Bárcenas fue acusado por el juez.*

124 A nivel oracional, el verbo siempre funciona como *head*.

(25)

```
# (null) 0 0-27|El juez acusó a Bárcenas.
[V: acusar third sing past <OBJ PN: Bárcenas <SUBJ [N: juez <DET
D: el]]


# (null) 1 0-33|Bárcenas fue acusado por el juez.
[V: acusar third sing past <OBJ PN: Bárcenas <SUBJ [N: juez <DET
D: el]]
```

Más adelante volveremos sobre las distintas normalizaciones que se llevan a cabo en SpQA.

4.3.3 Utilización del grafo de dependencias en SpQA

Esta representación se ha utilizado, como ya hemos mencionado, en otros *parsers*. En esos *parsers*, la mayor parte de la información que representa el grafo concierne a relaciones de tipo sintáctico.

En nuestro caso tomamos el grafo como estructura de representación y lo utilizamos para recoger aquella información que consideramos relevante para el análisis preguntas en BR. Como iremos viendo al detallar las características de nuestra gramática y del análisis que se propone en ella, el grafo de SpQA recoge información sintáctica y semántica (en menor grado). La información sintáctica concierne a los distintos constituyentes de la oración (a nivel oracional y de la frase) y sus relaciones de dependencia. La información semántica se limita al carácter de la incógnita o variable presente en las preguntas (cf. sección 4.3).

4.3.4 Motivos para la elección de este modelo

La elección de este modelo de representación responde a diferentes motivos.

En primer lugar, este modelo se ha utilizado satisfactoriamente en otros *parsers* orientados a tareas de Recuperación de Información (Koster et al., 2007).

En segundo lugar, se considera que el grafo permite recoger toda la información sintáctica (y cierta información semántica) que consideramos pertinente para un sistema de BR.

La representación que el grafo ofrece es además compacta, simple y potencialmente fácil de utilizar por otras aplicaciones o módulos de un sistema de este tipo.

Finalmente, el grafo permite la extracción de tripletes de dependencias que, como vimos en el capítulo 1, constituyen el método de representación sintáctica más utilizado en el área de la BR.

4.3.5 Elementos del grafo

En la presentación de los módulos de la gramática iremos detallando las características del análisis por *transduction* de SpQA. Como paso previo, para que los

ejemplos de análisis en la exposición sean comprensibles, presentaremos los elementos que pueden funcionar como nodos en el grafo así como las principales funciones sintácticas a nivel oracional (argumentos y adjuntos) y de la frase contempladas en la gramática.

Elementos que pueden funcionar como nodos en SpQA

- Verbo: se representa como `V`.

(26)
```
# (null) 4 0-19|Zapatero claudicó.
[V: claudicar third sing past <SUBJ PN: Zapatero]
```

- Nombre común: se representa como `N`.

(27)
```
# (null) 5 0-19|El presidente duda.
[V: dudar third sing present <SUBJ [N: presidente <DET D: el]]
```

- Nombre propio: se representa como `PN`.

(28)
```
# (null) 8 0-17|Bárcenas miente.
[V: mentir third sing present <SUBJ PN: Bárcenas]
```

- Adjetivo: se representa como `A`.

(29)
```
# (null) 9 0-22|Su viejo amigo llegó.
[V: llegar third sing past <SUBJ [N: amigo <ATTR A: viejo <DET
D: su]]
```

- Adverbio: se representa como `X`.

(30)
```
# (null) 11 0-11|Vive lejos.
[V: vivir third sing present <CIRC X: lejos]
```

- Determinante: se representa como `D`.

(31)
```
# (null) 10 0-9|La peste.
[N: peste <DET D: la]
```

- Cuantificador: se representa como `Q`.

(32)
```
# (null) 11 0-20|Bastantes problemas.
[N: problemas <QUANT Q: bastantes]
```

- Clítico verbal: de objeto directo (se representa como `lo`); de objeto indirecto (se representa como `le`); reflexivo (se representa como `se`).

(33)
```
# (null) 12 0-13|Me lo contó.
[V: contar third sing past <OBJ lo third masc|neut sing <IOBJ le
first GENDER sing]
```

(34)
```
# (null) 1 0-10|Se mienten
[V: mentir third plu present <REF se third GENDER plu]
```

Las palabras interrogativas también pueden funcionar como nodos en el grafo, pero tienen una representación especial, ya que su lema nunca se muestra en la representación. En lugar de ello, cuando el interrogativo constituye la única unidad de la frase interrogativa, se muestra su contenido semántico (35); cuando funciona como determinante o modificador, se muestra generalmente la palabra a la que determina/modifica (36).

(35) `# (null) 3 0-27|¿`**`Cuándo`** `nació Leo Messi?`
`whQ [V: nacer third sing past <SUBJ PN: Leo Messi <QCIRC` **`TIME`**`]`

(36) `# (null) 4 0-40|¿`**`Qué planeta`** `está más cerca del Sol?`
`whQ [V: estar third sing present <CIRC [[X: cerca <MODde [N: sol <DET D: el]] <QUANT más] <QSUBJ` **`N: planeta`**`]`

Estas cuestiones se explicarán en detalle cuando se presente el módulo de las interrogativas (cf. sección 4.3.4.2).

Principales relaciones sintácticas de dependencia a nivel de la frase en SpQA

- `ATTR`: adjetivo funcionando como modificador de un nombre.

(37) `# (null) 0 0-34|Bárcenas escribía con pluma azul.`
`[V: escribir third sing imperfect <CIRCcon [N: pluma <`**`ATTR`** `A: azul] <SUBJ PN: Bárcenas]`

- `MOD`: función general para el resto de modificadores.

(38) `# (null) 1 0-43|Verne escribió muchas novelas de aventuras.`
`[V: escribir third sing past <OBJ [N: novelas <`**`MODde`** `N: aventuras <QUANT Q: muchas] <SUBJ PN: Verne]`

- `DET`: determinante.

(39) `# (null) 1 0-26|Adora la ópera.`
`[V: adorar third sing present <OBJ [N: ópera <`**`DET`** `D: la]]`

- `QUANT`: determinante (1) o adverbio (2) cuantificador.

(40) `# (null) 2 0-32|El presidente debe muchas explicaciones.`
`[V: deber third sing present <OBJ [N: explicaciones <`**`QUANT`** `Q: muchas] <SUBJ [N: presidente <DET D: el]]`

(41) `# (null) 3 0-50|Madrid está bastante lejos del mar.`
`[V: estar third sing present <CIRC [[X: lejos <MODde [N: mar <DET D: el]] <`**`QUANT`** `X: bastante] <SUBJ PN: Madrid]`

- `COMP`: expresa la relación de comparación en las comparativas (cf. 4.2.2.6).

(42) `# (null) 4 0-37|Rajoy es más alto que Zapatero.`
`[V: ser third sing present <PRED [A: alto <`**`COMP`**`másque PN: Zapatero] <SUBJ PN: Rajoy]`

- AUX: expresa la relación de auxiliar en las perífrasis.

(43) `# (null) 5 0-40|Las avestruces no pueden volar.`
`[[V: volar third plu present <AUXposib V: poder] <SUBJ [N: avestruces <DET D: las] <NEG X: no]`

Principales relaciones sintácticas de dependencia a nivel oracional en SpQA

- SUBJ: sujeto.

(44) `# (null) 6 0-18|Arguiñano cocina.`
`[V: cocinar third sing present <SUBJ PN: Arguiñano]`

- OBJ: objeto directo. En su representación no aparece la preposición *a*.

(45) `# (null) 7 0-31|Cervantes escribió El Quijote.`
`[V: escribir third sing past <OBJ PN: El Quijote <SUBJ PN: Cervantes]`

- IOBJ: objeto indirecto. En su representación no aparece la preposición *a*.

(46) `# (null) 8 0-31|Debe dinero al banco.`
`[V: deber third sing present <OBJ N: dinero <IOBJ [N: banco <DET D: el]]`

- PRED: predicativo.
  El predicativo engloba predicativos (1) y atributos (2).

(47) `# (null) 9 0-22|Ella terminó agotada.`
`[V: terminar third sing past <PRED A:agotada <SUBJ P: ella]`

(48) `# (null) 10 0-21|La mujer es mi amiga.`
`[V: ser third sing present <PRED [N: amiga <DET D: mi] <SUBJ [N: mujer <DET D: la]]`

- PC: complemento preposicional.
  Se refiere a complementos preposicionales regidos por el verbo.

(49) `# (null) 11 0-20|Habla de tonterías.`
`[V: hablar third sing present <PCde N: tonterías]`

- CIRC: complemento circunstancial.

(50) `# (null) 12 0-29|El famoso viajó a Barcelona.`
`[V: viajar third sing past <CIRCa PN: Barcelona <SUBJ [A: famoso <DET D: el]]`

- NEG: relación de dependencia específica para el adverbio de negación, *no*.

(51) `# (null) 20 0-26|No bajarán los impuestos.`
`[V: bajar third plu future <OBJ [A:impuestos <DET los] <NEG X: no]`

El complemento agente, aunque se contempla como argumento en la gramática (*agent*), no se representa en el grafo de dependencias donde, como hemos visto, se realiza «despasivización», de manera que el agente pasa a ser sujeto:

(52) `# (null) 13 0-36|Berlusconi fue castigado por la ley.`
`[V: castigar third sing past <OBJ PN: Berlusconi` **`<SUBJ`** `[N: ley <DET D: la]]`

4.4 Descripción de la gramática

Como hemos visto, el *parser* de SpQA se genera a partir de una gramática de atributos y sus lexicones y *fact tables* asociados. La gramática contiene una serie de reglas que se ocupan de la descripción de la sintaxis, mientras que los lexicones contienen la información sobre las clases de palabras manejadas.

El núcleo inicial de desarrollo de la gramática de SpQA está en ASPIRA, una gramática (en desarrollo) orientada a tareas de RI, fruto de la colaboración de la autora con el profesor C.H. Koster[125]. A este núcleo inicial corresponden parte de los lexicones de la gramática y secciones de algunos módulos que describen tipos de frases[126]. La arquitectura modular de la gramática también responde al diseño inicial de ASPIRA, y es similar a las de otras gramáticas construidas con el formalismo AGFL (como la del *parser* para el holandés DUPIRA, Koster, 2011). En este diseño, una serie de módulos se encargan de las reglas que definen las clases de palabras; otra serie de módulos de las reglas que definen los distintos tipos de frases; finalmente, otra serie de módulos se encarga de las reglas que definen los distintos tipos de cláusulas (Rojo, 1978; Rojo, y Jiménez, 1989).

SpQA está orientado al análisis de preguntas y no al análisis del lenguaje en general. Por esta razón, el nivel de profundidad de los módulos destinados a la identificación de clases de palabras o frases es aquel que hemos juzgado suficiente para dar una cobertura adecuada a esa tarea específica. La variedad de estructuras lingüísticas posibles en una pregunta es mucho menor que aquella de la que sería necesario dar cuenta en una gramática de propósito general. Además, en un entorno de BR, la complejidad esperable para las preguntas a nivel estructural es menor que la que podemos encontrar en las preguntas del lenguaje común. En la evaluación de SpQA (cf. capítulo 5) podremos observar los resultados de este planteamiento.

En las secciones que siguen describiremos detalladamente la gramática de SpQA. En lo que respecta a la descripción de la gramática, nuestro objetivo no es tanto la exhaustividad en la presentación de todos y cada uno de los elementos que la conforman sino la delimitación y descripción clara de las principales piezas que entran en juego en el análisis de las preguntas. Con este objetivo en mente, presentaremos en primer lugar los lexicones de la gramática, para introducirnos a continuación en los

125 ASPIRA, a su vez, surgió de una versión de la gramática de propósito general AVALON (Álvarez, C. et al., 1998), desarrollada por Mª. Paula Santalla del Río en la Universidad de Santiago de Compostela.

126 Estas primeras versiones de los módulos sufrieron bastantes cambios en el desarrollo posterior de la gramática.

distintos módulos gramaticales que contienen las reglas sintácticas. En la descripción de los módulos, trataremos en primer lugar los módulos generales, que definen las clases de palabras, los tipos de frases y los tipos de cláusula. A continuación nos centraremos en aquello que constituye el núcleo de este trabajo: el módulo que se encarga de las oraciones interrogativas[127].

4.4.1 Lexicones

Los lexicones de la gramática recogen las distintas formas léxicas contempladas por esta. La gramática de SpQA consta de cinco lexicones:

- *n.dat*: recoge todas las formas de los sustantivos reconocidos por la gramática.
- *a.dat*: recoge todas las formas de los adjetivos reconocidos por la gramática.
- *v.dat*: recoge todas las formas de los verbos reconocidos por la gramática.
- *x.dat*: recoge todas las formas de los adverbios y locuciones adverbiales reconocidos por la gramática.
- *o.dat*: recoge el resto de clases de palabras reconocidas por la gramática (determinantes, cuantificadores, preposiciones, conjunciones, palabras interrogativas y clíticos verbales).

4.4.2 Módulos generales de la gramática

4.4.2.1 Clases de palabras

En esta sección trataremos las distintas clases de palabras antes enumeradas. Distinguimos dos grupos:

- Clases de palabras principales de la gramática (definen tipos de frases en SpQA): sustantivos, adjetivos, adverbios, verbos y partículas interrogativas.
- Otras clases de palabras: determinantes, cuantificadores, pronombres, preposiciones, conjunciones y clíticos verbales.

4.4.2.1.1 Sustantivos

La categoría se define con el no terminal *Noun* e incluye:

- Nombres comunes: les corresponde el siguiente no terminal en la gramática:

Noun(GENDER, NUMBER)

En la *transduction* se representan como `N:` más el lema.

(53) `# (null) 0 0-11|el gobierno`
`[N: gobierno <DET D: el]`

127 Como venimos haciendo hasta ahora, para facilitar la lectura, diferenciaremos tipográficamente los elementos pertenecientes a la gramática (en cursiva), de los elementos propios del grafo de dependencias (en letra tipo Courier).

En la gramática distinguimos, dentro de los nombres comunes, cuatro subtipos semánticos con no terminales específicos: *tempNoun*, *locNoun*, *modalNoun* y *causeNoun*. Cada uno de estos subtipos de nombres comunes está relacionado con un contenido semántico concreto:

- *tempNoun*: expresión de la temporalidad (*día*, *año*, *mes*, etc.);
- *locNoun*: expresión de la localización geográfica (*país*, *estado*, *continente*, etc.);
- *modalNoun*: expresión de la modalidad (*modo*, *forma*, etc.);
- *causeNoun*: expresión de la causalidad (*causa*, *razón*).

Todos se representan en el grafo mediante `N:` más lema. Su diferenciación en la gramática obedece a cuestiones relacionadas con el procesamiento de la frase interrogativa que veremos más adelante.

Para los nombres comunes, los afijos de la categoría son:

- *GENDER*: define el género de la palabra, que puede ser masculino (*masc*) o femenino (*fem*).
- *NUMBER*: define el número de la palabra, que puede ser singular (*sing*) o plural (*plu*).

Como veremos, estos afijos son comunes a otras clases de palabras.

- Nombres propios: les corresponde el siguiente no terminal en la gramática: *Noun(proper)*

En la *transduction* se representan como `PN:` más lema.

(54)
```
# (null) 1 0-6|la ONU
[PN: ONU <DET D: la]
```

Para los nombres propios tenemos un solo afijo, *proper*, que no tiene valor morfológico y solo indica que el sustantivo es propio.

Como se puede ver en los ejemplos, en el caso de los nombres (comunes y propios) ninguno de los afijos citados aparece en la *transduction*.

NER y BR

Vimos en el capítulo 1 que el reconocimiento de entidades nombradas y su clasificación (NER/NEC) es uno de los aspectos prioritarios en BR. Por esta razón, en SpQA se ha prestado especial cuidado a las reglas para el reconocimiento de entidades nombradas. Con todo, la eficacia del *parser* en esta tarea es limitada.

La tarea de NER ha experimentado un gran desarrollo en los últimos años debido a su importancia en el área de la recuperación y extracción de información. Por esta razón, existen muchas herramientas específicas para llevarla a cabo. Una de las herramientas libres disponibles para el español es la del sistema *Freeling* (Padró et al., 2010).

En una pequeña evaluación para determinar la eficacia de SpQA frente a *Freeling* en el reconocimiento de entidades en preguntas, comprobamos que el rendimiento de SpQA era inferior al módulo de *Freeling* en este aspecto. Por esta razón decidimos implementar en la gramática una serie de mecanismos que permiten, en el caso de desearlo, utilizar como preprocesado para las preguntas el módulo de NER de *Freeling*, aplicando posteriormente el análisis sintáctico semántico propio de SpQA.

Las entidades por las que se pregunta en BR también pueden ser nombres comunes, por lo que el reconocimiento de los nombres comunes también es importante en SpQA. Para paliar posibles deficiencias del lexicón de sustantivos (*n.dat*) se han implementado reglas que permiten el reconocimiento robusto de sustantivos. Cuando el sustantivo no está en el lexicón, pero se identifica como tal, se le aplica la etiqueta `UNKNOWN N:`:

(55) `# (null) 0 0-29|¿Qué es un rabdobiosarcoma?`
`whQ [V: ser third sing present <QPRED DEFINITION` **`<SUBJ [UNKNOWN N: rabdobiosarcoma?`**[128] `<DET D: un]]`

4.4.2.1.2 Adjetivos

La categoría se define en la gramática con el no terminal *Adj*:

Adj(GENDER,NUMBER)

Y se representa en el grafo como `A:` más lema:

(56) `# (null) 0 0-10|muy bonito`
`[`**`A: bonito`** `<QUANT X: muy]`

Se consideran adjetivos los propiamente dichos (*coche* rojo), los participios funcionando como tales (*hombre* agotado) y los ordinales (primer *árbol*).

Los afijos de la categoría son los mismos que los de los sustantivos: *GENDER* y *NUMBER*.

Ninguno de los afijos aparece en la representación.

Cuando el adjetivo funciona como modificador de un sustantivo la función que le corresponde en el *relator* es `ATTR`:

(57) `# (null) 1 0-14|el niño guapo`
`[N: niño` **`<ATTR A: guapo`** `<DET D: el]`

128 Al no reconocer «rabdobiosarcoma» como sustantivo, SpQA integra como parte de él el signo ortográfico de cierre de la interrogación «?».

(58) `# (null) 3 0-14|hombre agotado`
`[N: hombre` **`<ATTR A:agotado`**`]`

(59) `# (null) 5 0-13|primer árbol`
`[N: árbol` **`<ATTR A: primer`**`]`

4.4.2.1.3 Adverbios

La categoría se define en la gramática con el no terminal *Adverb*:

Adverb (AVTYPE,PREP,DEGREE,AVPREF)

En el grafo se representa como `X:` más lema.

(60)`# (null) 1 0-5|lejos`
`[X: lejos]`

La categoría engloba los distintos tipos de adverbios y también las locuciones adverbiales:

(61)`# (null) 2 0-14|a base de bien`
`[X: a base de bien]`

Los afijos de la categoría son los siguientes:

- *AVTYPE*: define el tipo de adverbio.
- *PREP*: hace referencia a la posible combinación del adverbio con preposiciones.
- *DEGREE*: define el grado del adverbio.
- *AVPREF*: define las preferencias de combinación del adverbio (verbo, adverbio, nombre o adjetivo).

Ninguno de estos afijos aparece en la representación.

4.4.2.1.4 Verbos

Se definen en la gramática mediante no terminales con una misma denominación, *Verb*, pero con distintos afijos dependiendo del tipo de verbo.

De forma general, todas las formas verbales se representan en el grafo mediante `V:` más el lema en infinitivo. En el caso de los verbos en forma personal, se añade además la información correspondiente a los afijos de persona (*PERSON*), número (*NUMBER*) y tiempo (*TENSE*):

(62)`# (null) 0 0-6|llegó`
`[V: llegar third sing past]`

(63)`# (null) 1 0-6|bailar`
`[V: bailar]`

(64)`# (null) 2 0-8|llorando`
`[V: llorar]`

(65)`# (null) 3 0-8|olvidado`
`[V: olvidar]`

Los valores posibles para el afijo *NUMBER* ya los hemos visto (*sing* y *plu*). Los valores de los afijos *PERSON* y *TENSE* son los siguientes:

- *PERSON*: primera, *first* (*canto*, *cantamos*); segunda, *second* (*cantas*, *cantáis*); tercera, *third* (*canta*, *cantan*).

- *TENSE*:
  Formas simples: presente, *present* (*vamos*); perfecto, *past* (*fui*); imperfecto, *imperfect* (*había*); futuro, *future* (*irás*); condicional, *conditional* (*vivirían*).
  Formas compuestas: presente perfecto, *present_perfect* (*has viajado*); pretérito perfecto, *past_perfect* (*hubo venido*); pluscuamperfecto, *pluperfect* (*había soñado*); futuro perfecto, *future_perfect* (*habremos llegado*); condicional perfecto, *conditional_perfect* (*habríais llamado)*.

Cuando la forma verbal está constituida por más de una unidad la representación es la siguiente:

- **Tiempos compuestos**: se representan con el lema del participio, sin información del auxiliar:

(66)`# (null) 0 0-10|ha soñado`
`[V: soñar third sing present_perfect]`

- **Perífrasis**: no se recoge la información relativa al nexo (67) pero sí la relativa al auxiliar (67), (68). La subordinación del auxiliar al verbo principal se marca mediante la relación de dependencia `AUX`.
  Se recoge también información relativa al aspecto mediante el afijo *PERTYPE*; los tipos de perífrasis son:

  - **Aspectuales**: *inmin* para las ingresivas (*voy a irme*); *begin* para las incoativas (*empezó a cantar*); *freq* para las habituales o consuetudinarias (*suelo estudiar*); *repet* para las reiterativas (*volvió a fumar*); *finish* para las egresivas (*terminó de estudiar*); *durat* para las durativas (*está viajando*); *result* para las resultativas (*tengo estudiado*).

- **Modales**: *oblig* para obligación (*tengo que ir*); *probab* para probabilidad (*debe de ser*).

La información del afijo *PERTYPE* se representa como paramétro del relator AUX.

(67)# (null) 1 0-24|había empezado **a** cantar
[V: cantar first|third sing pluperfect **<AUXbegin** V: empezar]

(68)# (null) 2 0-10|debemos ir
[V: ir first plu present **<AUXoblig** V: deber]

- **Clíticos**: los clíticos siempre aparecen en la representación, indicando en cada caso su función sintáctica.
  El pronombre reflexivo tiene un indicador de función propio, REF. Como hemos visto ya, en la representación el clítico siempre se representa como lo en el caso del objeto directo, se en el caso de los reflexivos y le en el caso del objeto indirecto, junto a la información de persona, género y número.

(69)# (null) 3 0-7|dámelo
[V: dar second sing present **<OBJ lo third masc sing <IOBJ le first GENDER sing**]

(70)# (null) 4 0-10|se lo dijo
[V: decir third sing past **<OBJ lo third masc sing <IOBJ le third GENDER NUMBER**]

(71)# (null) 5 0-8|me peino
[V: peinar first sing present **<REF se first GENDER sing**]

Dentro de la categoría general *Verb* se distinguen mediante no terminales específicos dos subtipos semánticos de verbos:

- *VerbQuant*: corresponde a un grupo de verbos que implican cuantificación. Distinguimos a su vez dos tipos de *VerbQuant*: una para los verbos que implican cuantificación de una medida (*medir*, *pesar*) y otra para los verbos que implican cuantificación de dinero (*costar*, *ganar*).
- *VerbDenom*: corresponde a un grupo de verbos relacionado con la denominación de entidades (*llamar*, *denominar*, *nombrar*, etc.).

Las dos clases semánticas están relacionadas con el procesamiento de la frase interrogativa, por lo que volveremos sobre ellas más adelante. En el grafo se representan como todos los verbos: `V:` más lema.

4.4.2.1.4.1 Subcategorización verbal

Los verbos tienen un afijo que maneja la subcategorización (o valencia): *VSUB*. *VSUB* puede tener los siguientes valores:

- *intran*: verbos intransitivos (*morir*);
- *ditran*: verbos ditransitivos (*regalar*);
- *cplxtran*: verbos que rigen complemento directo y predicativo (*considerar*);
- *pred*: verbos que rigen predicativo (*ponerse*);
- *ppc*: verbos que rigen complemento preposicional (*tender a*);
- *io*: verbos que rigen complemento indirecto (*sonreír*);
- *do+ppc*: verbos que rigen complemento directo más complemento preposicional (*convencer* (*de*));
- *do*: verbos transitivos (*comer*).

Las preferencias de subcategorización se recogen en una *fact table*: *vsub.fct*. Estas preferencias son aquellas a las que la gramática da prioridad para el análisis de un verbo determinado.

El listado de verbos de *vsub.fct*[129] no engloba todos los lemas del lexicón de verbos *v.dat*. Por esta razón no todos los verbos tienen en la gramática unas preferencias de subcategorización preestablecidas. Para los verbos sin una subcategorización preestablecida son los argumentos con los que el verbo aparece en la oración a analizar los que determinan el valor de *VSUB* en cada caso.

La información relativa a la subcategorización no aparece en el grafo de dependencias.

4.4.2.1.5 Partículas interrogativas

De la descripción de las partículas interrogativas nos ocuparemos cuando presentemos el módulo de las interrogativas (cf. sección 4.4.3.4.2).

4.4.2.1.6 Determinantes

La categoría se define en la gramática mediante el no terminal *Determiners*: *Determiners(CASE,DEF,GENDER,NUMBER)*

En el grafo se representan como `D:` más lema. La relación de determinación entre determinante y núcleo se expresa mediante la función `DET:`

129 La información de subcategorización se extrajo de un lexicón de verbos construido por el Grupo de Gramática del Español (http://gramatica.usc.es/?lang=es) a partir de datos de la BDS (http://www.bds.usc.es/).

(72)# (null) 0 0-9|la estufa
[N: estufa **<DET D: la**]

(73)# (null) 1 0-13|algún sueño
[N: sueño **<DET D: algún**]

Se distinguen varios subtipos de determinantes: demostrativos, artículos, posesivos, indefinidos, etc., aunque todos ellos se representan de igual manera en la *transduction*: `D:` más lema.

4.4.2.1.6.1 Cuantificadores

Se distingue como categoría propia un subtipo de determinantes: aquellos relacionados con la cuantificación. Esta categoría engloba los cardinales (dos *libros*) y los indefinidos cuantificativos (bastantes *cosas*).

La categoría se define en la gramática mediante el no terminal *Quant* :

Quant(RANK,GENDER,NUMBER)

En el grafo se representa como `Q:` más lema, y la relación de cuantificación se expresa mediante la función `QUANT:`

(74)# (null) 0 0-14|cinco árboles
[N: árboles **<QUANT Q: cinco**]

(75)# (null) 1 0-16|bastantes libros
[N: libros **<QUANT Q: bastantes]**

4.4.2.1.7 Pronombres

La categoría se define mediante los terminales:

P(PTYPE,PERSON,GENDER,NUMBER,CASE)
P(PTYPE,PERSON,GENDER,NUMBER,CASE,CFUNC)

Como con los determinantes, en la gramática se distinguen (mediante el afijo *PTYPE*) varios tipos de pronombres: personales (*yo*), demostrativos (*esta*), indefinidos (*algunos*), globales (*todos*), etc. Ninguno de estos subtipos aparece en el grafo, donde todos los pronombres se muestran como `P:.`

(76)# (null) 1 0-14|todos llegaron
[V: llegar third plu past <SUBJ **P: todos**]

(77)# (null) 3 0-16|algunos soñaron
[V: soñar third plu past <SUBJ **P: algunos**]

4.4.2.1.8 Preposiciones

La categoría se define en la gramática mediante el terminal *Prep*:
Prep(PREP)

Las preposiciones no tienen representación autónoma en el grafo de dependencias ya que son un parámetro del *relator* que indica la función de la frase preposicional correspondiente.

(78)# (null) 4 0-29|el descubrimiento de América
[N: descubrimiento **<MODde** PN: América <DET D: el]

(79)# (null) 6 0-23|Cañizares es de Madrid.
[V: ser third sing present **<PREDde** PN: Madrid <SUBJ PN: Cañizares]

Volveremos sobre esta representación cuando tratemos la frase preposicional (cf. sección 4.4.2.2.5).

En los casos de contracción de preposición más determinante se deshace la contracción en la representación:

(80)# (null) 2 0-16|El salmón viene del río.
[V: venir third sing present <CIRC**de** [N: río <DET D: **el**] <SUBJ [N: salmón <DET D: el]]

4.4.2.1.9 Conjunciones

La categoría se define mediante el terminal *Conj*:

Conj(CJTYPE)

La representación de las conjunciones varía dependiendo del tipo de conjunción:

- En las **coordinadas** la conjunción no se representa. En su lugar se utiliza el elemento "|" para marcar conjunción de elementos:

(81)# (null) 0 0-40|Barça y Madrid compitieron en la final.
[V: competir third plu past <CIRCen [N: final <DET D: la] **<SUBJ PN: Barça | PN: Madrid**]

TRIPLETES
[N: final, DET, D: la]
[V: competir third plu past, CIRCen, N: final]
[V: competir third plu past, SUBJ, PN: Barça]
[V: competir third plu past, SUBJ, PN: Madrid]

- En las subordinadas con *que* la conjunción se omite porque se considera que no aporta información a la representación:

(82)# (null) 1 0-53|El Pentágono decidió que EEUU entrara en la guerra.
[V: decidir third sing past **<OBJ [V: entrar third sing imperfect <SUBJ PN: EEUU <CIRCen [N: guerra <DET D: la]]** <SUBJ PN: El Pentágono][130]

- En las subordinadas con *cuando* se mantiene la conjunción porque aporta contenido semántico a la representación ('ubicación temporal'). Su representación es paralela a la de las preposiciones:

(83)# (null) 2 0-56|La victoria de los aliados llegó cuando EEUU intervino.
[V: llegar third sing past **<CIRCcuando [V: intervenir third sing past <SUBJ PN: EEUU]** <SUBJ [N: victoria <MODde [N: aliados <DET D: los] <DET D: la]]

- Para la disyunción, la representación es algo más compleja. Cuando tenemos una estructura disyuntiva, la conjunción no se muestra. Los miembros de la disyunción se representan como dependientes de un nodo denominado `disjunction` con el que mantienen una relación de dependencia etiquetada como `DISJ`:

(84)# (null) 0 0-24|El ibuprofeno se comercializa en pastilla o en polvo.
[V: comercializar third sing present <REF se third GENDER sing <CIRC **[disjunction <DISJen N: pastilla <DISJen N: polvo]** <SUBJ [UNKNOWN N: ibuprofeno <DET D: el]]

130 Volveremos sobre la representación de las subordinadas en la sección 4.4.2.3.2.3.

TRIPLETES
[DISJUNCTION, DISJen, N: pastilla]
[DISJUNCTION, DISJen, N: polvo]
[UNKNOWN N: ibuprofeno, DET, D: el]
[V:comercializar third sing present, CIRC, DISJUNCTION]
[V:comercializar third sing present, REF, se third GENDER sing]
[V:comercializar third sing present, SUBJ, UNKNOWN N: ibuprofeno]

En el ejemplo, cada miembro de la disyunción se muestra como parte del circunstancial, pero en una relación de disyunción. Esta representación es similar a la del analizador de *DepPattern* para el español.

Otras gramáticas escritas con AGFL como DUPIRA o EP4IR no distinguen en su representación entre coordinación y disyunción de elementos. Veamos un ejemplo de análisis del *parser* para el inglés, EP4IR:

(85) He bought milk and juice.
{P: he, SUBJ [V: bought, OBJ [N: milk | N: juice]]}

TRIPLETES
[he, SUBJ, bought]
[bought, OBJ, juice]
[bought, OBJ, milk]

(86) He bought milk or juice.
{P: he, SUBJ [V: bought, OBJ [N: milk | N: juice]]}

TRIPLETES
[he, SUBJ, bought]
[bought, OBJ, juice]
[bought, OBJ, milk]

Con esta representación no se distingue el caso en el que el objeto consta de dos elementos (85) del caso en el que el objeto consta de un solo elemento posible entre dos (86).

Frente a este tipo de análisis, nosotros consideramos que debe distinguirse la relación de coordinación de la de disyunción, pues la diferencia de significado entre ambos tipos de relación es relevante a la hora de recuperar y extraer información. Esta afirmación se verá más clara en el caso de las preguntas disyuntivas y volveremos sobre ella más adelante (cf. sección 4.4.3.4.3).

4.4.2.2 Tipos de frases en SpQA

En la gramática de SpQA se distinguen los siguientes tipos de frases:

- Frase Nominal: no terminal *NP* (*Noun Phrase*).

Ejemplos: *el niño, la casa grande.*

- Frase Adjetiva: no terminal *AP* (*Adjective Phrase*).

Ejemplos: *muy bonito, bastante guapo.*

- Frase Adverbial: no terminal *XP* (*Adverbial Phrase*).

Ejemplos: *cerca, muy bien.*

- Frase Verbal: no terminal *VP* (*Verbal Phrase*).

Ejemplos: *viaja, ha dormido, se lo ha dicho, tiene que estudiar.*

- Frase Preposicional: no terminal *PP* (*Prepositional Phrase*).

Ejemplos: *de Barcelona, para muy lejos, con arrojo.*

- Frase Comparativa: no terminal *CP* (*Comparative Phrase*).

Ejemplos: (*más celoso*) *que Otelo,* (*tan amigo*) *como tú.*

- Frase Interrogativa

Ejemplo: *¿Qué famoso presidente latinoamericano... ?*

Como vimos cuando presentamos el modelo de análisis de SpQA, en la representación no se muestran las categorías a nivel de frase (aunque en ocasiones sí se marcan sus límites mediante corchetes, cf. *supra*):

(87) `# (null) 0 0-24|El gobierno convocará un referéndum urgente.`
`[V: convocar third sing future <OBJ` **`[N: referéndum <ATTR A: urgente <DET D: un]`** `<SUBJ [N: gobierno <DET D: el]]`

De este modo, en el ejemplo anterior no se dice que

```
[N: referéndum <ATTR A: urgente <DET D: un]
```

es una *NP*[131].

Pese a todo, aunque la información sobre los distintos tipos de frases no se muestra en la representación, sí está debajo del análisis que genera esa representación[132]. Por esa razón, en las secciones siguientes presentaremos sucintamente cada uno de los tipos de frases de SpQA.

131 Aunque esta información pueda deducirse del hecho de que la *head* de la estructura es un sustantivo.
132 Esto permite, además, incorporar este aspecto a la representación en el futuro si fuese necesario.

4.4.2.2.1 Frase nominal

Estructura básica:

- Núcleo: generalmente es un sustantivo (*Noun*), aunque la gramática también contempla las posibilidades de que sea un pronombre (todos *vinieron*) o un adjetivo (*el* alto).
- Frase sustantiva: no terminal *Noun Part*. Núcleo más posibles modificadores (frases preposicionales, adjetivas, etc.).
- Frase nominal: no terminal *Noun Phrase.* Posibles determinantes (*Determiners*) más una frase sustantiva (*Noun Part*).

Además de esta *NP* general se distinguen en la gramática varias *NP* con valores semánticos específicos:

- *NPquant*: frase nominal de tipo cuantitativo:

(88)
```
# (null) 2 0-21|ocho escritores de París
[N: escritores <MODde PN: París <QUANT Q: ocho]
```

- *NPdate:* todas las *NP* que expresan una fecha (*5 de julio, tres de agosto de 1983, 05-02-1995,* etc.) o una ubicación temporal clara similar a una fecha (*en 1989, el miércoles, en agosto*).
  Esta *NP* tiene un indicador de función específico, *DATE*, que también se utiliza con las frases preposicionales con valor de fecha (cf. *infra*).
  Con estas *NP* se realiza además una normalización en la *transduction*, de manera que todas las fechas se representan con el mismo formato:

(89)
```
# (null) 0 0-20| Einstein nació el 2 de abril.
[V: nacer third sing past <DATE 2-04 <SUBJ PN: Einstein]
```

(90)
```
# (null) 2 0-21|Einstein nació el 02 de abril.
[V: nacer third sing past <DATE 02-04 <SUBJ PN: Einstein]
```

(91)
```
# (null) 4 0-18|Einstein nació el 2 del 04.
[V: nacer third sing past <DATE 2-04 <SUBJ PN: Einstein]
```

(92)
```
# (null) 8 0-32|Lenin nació el tres de abril de 1870.
[V: nacer third sing past <DATE 03-04-1870 <SUBJ PN: Lenin]
```

(93)
```
# (null) 9 0-30|Lenin nació el 03 de abril de 1870.
[V: nacer third sing past <DATE 03-04-1870 <SUBJ PN: Lenin]
```

- *tempNoun Part*: frase sustantiva que tiene como núcleo un *tempNoun* (cf. *supra*).

  ¿En qué año *se inició la Segunda Guerra Mundial?*

- *locNoun Part*: frase sustantiva que tiene como núcleo un *locNoun* (cf. *supra*).

  ¿En qué país *de Europa se legalizó primero el matrimonio homosexual?*

- *modalNoun Part*: frase sustantiva que tiene como núcleo un *modalNoun* (cf. *supra*).

  ¿De qué forma *se mide un perímetro?*

- *causeNoun Part*: frase sustantiva que tiene como núcleo un *causeNoun* (cf. *supra*).

  ¿Por qué motivo *el cielo es azul?*

Cuando presentemos la frase interrogativa (cf. sección 4.4.3.4.3) volveremos sobre estos tipos especiales de *NP*.

En la frase nominal, en el grafo no se representa el nominal diferenciado mediante corchetes:

(94) `# (null) 0 0-24|un gobierno incompetente`
`[N: gobierno <ATTR A: incompetente <DET D: un]`

No obstante, esta información podría incorporarse a la gramática si fuera necesario gracias a la arquitectura de la frase nominal en SpQA.

Cuando la frase nominal presenta varios determinantes no se representa la jerarquía de determinación sino que todos los determinantes se sitúan al mismo nivel:

(95) `# (null) 1 0-18|todos estos chicos`
`[N: chicos <DET D: todos | D: estos]`

EJEMPLOS DE ANÁLISIS

(96) `# (null) 2 0-15|la niña bonita`
`[N: niña <ATTR A: bonita <DET D: la]`

(97) `# (null) 3 0-39|todos los hombres buenos del presidente`
`[N: hombres <ATTR A: buenos <MODde [N: presidente <DET D: el] <DET D: todos | D: los]`

4.4.2.2.2 Frase adjetiva

La estructura básica es la misma que la de la frase nominal:

- Núcleo: cualquier categoría que pueda funcionar como adjetivo (adjetivos, ordinales y participios).
- *Adj Part*: puede constar de un solo adjetivo (*guapo*), o de un adjetivo modificado (*libre de culpa*).
- *Adj Phrase*: consta de una *Adj Part* sola o modificada (*muy guapa, bastante más guapa, más guapa que tonta*).

EJEMPLOS DE ANÁLISIS

(98) `# (null) 5 0-9|muy guapa`
`[A: guapa <QUANT X: muy]`

(99) `# (null) 6 0-19|bastante más guapa`
`[A: guapa <QUANT [X: más <QUANT X: bastante]]`

(100) `# (null) 7 0-15|guapa de verdad`
`[A: guapa <MODde N: verdad]`

4.4.2.2.3 Frase adverbial

La estructura de la frase adverbial es paralela a la de la frase adjetiva:

- Núcleo: adverbio.
- *Adverb Part*: puede estar constituida por un adverbio (*lejos*) o un adverbio modificado (el modificador va después del adverbio: *lejos de aquí, aquí mismo*).
- *Adverb Phrase*: puede estar constituida por una *Adverb Part* (*allí*) o una *Adverb Part* pre-modificada (*muy lejos, bastante más lejos de aquí, más lejos que cerca*).

EJEMPLOS DE ANÁLISIS

(101) `# (null) 9 0-9|muy lejos`
`[X: lejos <QUANT X: muy]`

(102) `# (null) 0 0-18|muy lejos de aquí`
`[X: lejos <MODde X: aquí <QUANT X: muy]`

(103) `# (null) 1 0-28|bastante más lejos de aquí`
`[[X: lejos <MODde X: aquí] <QUANT [X: más <QUANT X: bastante]]`

4.4.2.2.4 Frase verbal

Está constituida por una forma verbal (verbo solo, perífrasis verbal o verbo más clíticos).

EJEMPLOS DE ANÁLISIS

(104)
```
# (null) 2 0-6|murió
[V: morir third sing past]
```

(105)
```
# (null) 0 0-17|habíamos pensado
[V: pensar first plu pluperfect]
```

(106)
```
# (null) 4 0-14|comenzó a ver
[V: ver third sing past <AUXbegin V: comenzar]
```

(107)
```
# (null) 5 0-7|dámelo
[V: dar second sing present <OBJ lo third masc sing <IOBJ le
first GENDER sing]
```

(108)
```
# (null) 6 0-15|me las pagarás
[V: pagar second sing future <OBJ lo third fem plu <IOBJ le
first GENDER sing]
```

4.4.2.2.5 Frase preposicional

Constituida por una preposición (*Prep*) y su término, que puede ser:

- una frase: *NP* (*Voy a* tu casa), *AP* (*Viste de* rojo), *XP* (*Vino de* muy lejos);

- una cláusula subordinada: relativa (*Es del* que conoces), cláusulas de infinitivo (*Con* estudiar *no es suficiente*), cláusulas con «que» (*Estoy contenta de* que hayas venido), cláusulas relativas temporales (*Esto es de* cuando tu padre aún vivía), etc.

Como ya vimos al tratar las preposiciones, la representación de la frase preposicional es distinta a la de los otros tipos de frases. En este caso la preposición no constituye el núcleo de la frase preposicional, sino que se adhiere como parámetro al indicador de función sintáctica que le corresponde a esta: `CIRC` para las frases preposicionales funcionando como circunstancial (109), `MOD` para las frases preposicionales funcionado como modificador (110), `PC` para las frases preposicionales funcionando como complemento preposicional (111) y `PRED` para las frases preposicionales con *de* funcionando como predicativo (112).

(109) `# (null) 4 0-20|Rajoy vive en Madrid .`
`[V: vivir third sing present` **`<CIRCen PN: Madrid`** `<SUBJ PN: Rajoy]`

(110) `# (null) 5 0-24|el ensanche de Barcelona`
`[[N: ensanche` **`<MODde PN: Barcelona ]`** `<DET D: el]`

(111) `# (null) 6 0-18|Habla de política`
`[V: hablar third sing present` **`<PCde N: política`**`]`

(112) `# (null) 7 0-31|Fernando Alonso es de Asturias.`
`[V: ser third sing present` **`<PREDde PN: Asturias`** `<SUBJ PN: Fernando Alonso]`

Este tratamiento de la frase preposicional es similar al de otros *parsers,* como el de Stanford[133], para el inglés:

(113) *Paul Auster lives* ***in New York****.*
(*Typed dependencies, collapsed*)
nn(Auster-2, Paul-1)
nsubj(lives-3, Auster-2)
root(ROOT-0, lives-3)
nn(York-6, New-5)
prep_in(lives-3, York-6)

o al de DepPattern[134] para varias lenguas, entre ellas, el español:

(114) *Paul Auster vive en Nueva York.*
(SubjL;vivir_VERB_1;Paul@Auster_NOUN_0)
(**CircR/en**_PRP_2;vivir_VERB_1;Nueva@York_NOUN_3)

Distinguimos un tipo semántico especial de frase preposicional: *PPdate*. Como su nombre indica, esta frase preposicional tiene como término una fecha. Este tipo de frase preposicional tiene el mismo indicador de función que la *NPdate*: `DATE`.

(115) `# (null) 8 0-37|El desastre nuclear ocurrió` **`en 1989.`**
`[V: ocurrir third sing past` **`<DATEen 1989`** `<SUBJ [N: desastre <ATTR A: nuclear <DET D: el]]`

(116) `# (null) 4 0-21|Estamos a 5 de julio.`
`[V: estar first plu present` **`<DATEa`** `5-07]`

Como se puede ver en los ejemplos, en el caso de la *PPdate* también recogemos la información relativa a la preposición en el grafo de dependencias.

133 http://nlp.stanford.edu/software/lex-parser.shtml
134 http://gramatica.usc.es/pln/tools/deppattern.html

4.4.2.2.6 Frase comparativa

Se corresponde con el segundo término de la comparación en las estructuras comparativas:
más alto que tú, *menos listo* que yo, *tan hombre* como tú**.**

La estructura comparativa en su totalidad está formada en SpQA por:

- Un adverbio comparativo: *más*, *menos*, *tan*.
- El término comparado: un adjetivo (*más* alto *que tú*), un sustantivo (*más* pan *que leche*), una frase preposicional (*más* de fiesta *que de diario*), un adverbio (*más* cerca *que lejos*) o un verbo (trabaja *más que estudia*).
- La frase comparativa: conjunción (*como*, *que*) más una frase nominal (*más cotilla que* la vecina), adjetiva (*más guapa que* lista), adverbial (*más lejos que* cerca), preposicional (*más de pueblo que* de ciudad) o verbal (*trabaja más que* estudia).

En la representación de la frase comparativa esta se considera un modificador del término comparado y la relación de comparación se expresa a través del indicador de función `COMP`. Pegado al indicador de función y como parámetro se coloca el nexo comparativo (constituido por el adverbio comparativo más la conjunción) y a continuación el núcleo de la frase comparativa:

(117) `# (null) 10 0-49|El Empire State es más alto que la Torre Eiffel.`
`[V: ser third sing present <PRED` **`[A: alto <COMPmásque [N: torre <MOD PN: Eiffel <DET D: la]`** `<SUBJ [PN: Empire State <DET D: el]]`

4.4.2.3 Módulos clausales

En esta sección nos ocuparemos de la descripción de las cláusulas no interrogativas tratadas en SpQA.

Como ocurre con las frases, en la representación de SpQA se muestran los límites de las cláusulas mediante corchetes pero no se marcan los distintos tipos de cláusulas. No obstante, como estos tipos están debajo del análisis que realiza el *parser*, los presentaremos a continuación de modo general.

Diferenciamos dos tipos de cláusulas: las declarativas no subordinadas y las subordinadas.

4.4.2.3.1 Cláusulas declarativas no subordinadas

En SpQA distinguimos los siguientes tipos de cláusulas declarativas no subordinadas:

- Cláusula en voz activa con sujeto elíptico: descrita mediante el no terminal *VOC Phrase*.

(118)
```
# (null) 11 0-29|Venció a todos sus enemigos.
[V: vencer third sing past <OBJ [N: enemigos <DET D: todos | D: sus]]
```

- Cláusula en voz activa con sujeto expreso: *SVOC Phrase*.

(119)
```
# (null) 12 0-27|Bárcenas viajó a Canadá.
[V: viajar third sing past <CIRCa PN: Canadá <SUBJ PN: Bárcenas]
```

- Cláusula con verbo impersonal: *VOC_IMP Phrase*.
  Excepto en los casos de verbos inherentemente impersonales (*llover*, *haber*, etc.), en la representación se marca la impersonalidad a través de un sujeto impersonal que tiene como nodo la fórmula `IMP`.

(120)
```
# (null) 13 0-23|Se lee poco en España.
[V: leer third sing present <SUBJ IMP <CIRC X: poco <CIRCen PN: España]
```

- Cláusula en voz Pasiva: *Pass Clause*.
  Como sabemos, las oraciones pasivas se someten a «despasivización» a través de la *transduction*, de manera que en el grafo de dependencias no se diferencian de una oración en voz activa:

(121)
```
# (null) 14 0-48|La misión Apollo XIII fue abortada por la Nasa.
[V: abortar third sing past <OBJ [N: misión <MOD PN: Apollo XIII <DET D: la] <SUBJ [PN: Nasa <DET D: la]]
```

- Cláusulas copulativas con sujeto elíptico: *xP Phrase*.

(122)
```
# (null) 15 0-30|Fue un error fatal del estado.
[V: ser third sing past <PRED [N: error <ATTR A: fatal <MODde [N: estado <DET D: el] <DET D: un]]
```

- Cláusulas copulativas con sujeto expreso: *SxP Phrase*.

(123)
```
# (null) 16 0-20|Merkel es rubia.
[V: ser third sing present <PRED A: rubia <SUBJ PN: Merkel]
```

4.4.2.3.2 Cláusulas subordinadas

La representación de las cláusulas subordinadas en SpQA varía dependiendo del tipo de subordinada. A continuación las presentaremos una a una, deteniéndonos en los detalles relativos a su representación.

4.4.2.3.2.1 Cláusula de relativo: *Relative Phrase*

Cláusula subordinada constituida por:

- un pronombre relativo;
- una cláusula declarativa no subordinada.

Se distinguen dos tipos básicos:

- Cláusula de relativo que funciona como modificador de una frase nominal: *Zapatero no apoyó la reforma* que propuso el ministro de Educación. (124).
- Cláusula de relativo que funciona como una frase nominal: El que fue elegido por el pueblo egipcio *fue derogado por el ejército.* (125).

(124) `# (null) 17 0-68|Zapatero no apoyó` **`la reforma que propuso el ministro de Educación.`**
`[V: apoyar third sing past` **`<OBJ [[N: reforma >OBJ [V: proponer third sing past <SUBJ [N: ministro <MODde PN: Educación <DET D: el]]] <DET D: la]`** `<SUBJ PN: Zapatero <NEG X: no]`

Como vemos en (124), en el caso de la frase relativa funcionando como modificador de una frase nominal no aparece el relativo en la representación. Por otra parte, no se muestra la relación de modificación de la subordinada en el interior de la frase nominal; en lugar de eso, se expresa la función que desempeña la frase nominal respecto al verbo principal y al verbo subordinado.

(125) `# (null) 18 0-63|`**`El que fue elegido por el pueblo egipcio`** `fue derogado por el ejército.`
`[V: derogar third sing past` **`<OBJ [P: el que >OBJ [V: elegir third sing past <SUBJ [N: pueblo <ATTR A: egipcio <DET D: el]]]`** `<SUBJ [N: ejército <DET D: el]]`

En el caso de la relativa funcionando como frase nominal sí se muestra el relativo, y se señala la función que desempeña ese relativo respecto al verbo subordinado y (en representación de toda la subordinada) respecto al principal.

Para mostrar con más claridad las relaciones de dependencia señaladas, mostramos a continuación los tripletes de dependencias que corresponden a las dos oraciones anteriores:

(126) # (null) 17 0-68|Zapatero no apoyó la reforma que propuso el ministro de Educación.
[N: ministro, DET, D: el]
[N: ministro, MODde, PN: Educación]
[N: reforma, DET, D: la]
[V: apoyar third sing past, NEG, X: no]
[V: apoyar third sing past, OBJ, N: reforma]
[V: apoyar third sing past, SUBJ, PN: Zapatero]
[V: proponer third sing past, OBJ, N: reforma]
[V: proponer third sing past, SUBJ, N: ministro]

(127) # (null) 18 0-63|El que eligió el pueblo egipcio fue derogado por el ejército.
[N: ejército, DET, D: el]
[N: pueblo, ATTR, A: egipcio]
[N: pueblo, DET, D: el]
[V: derogar third sing past, OBJ, el que]
[V: derogar third sing past, SUBJ, N: ejército]
[V: elegir third sing past, OBJ, el que]
[V: elegir third sing past, SUBJ, N: pueblo]

La decisión de representación que consiste en mostrar la relación de la entidad modificada (*reforma*) en (124) respecto al verbo subordinado, en lugar de mostrar la relación de modificación de la relativa sobre esa misma entidad, responde al hecho de que consideramos esta representación más útil en un entorno de BR. Veamos un ejemplo de los tripletes extraídos de una pregunta con una frase de relativo:

(128)# (null) 5 0-46|¿Qué país que tiene petróleo es muy pobre?
whQ [V: ser third sing present <PRED [A: pobre <QUANT X: muy] <QSUBJ [N: país >SUBJ [V: tener third sing present <OBJ N: petróleo]]]

TRIPLETES
[A: pobre, QUANT, X: muy]
[V: tener third sing present, SUBJ, N: país]
[V: ser third sing present, PRED, A: pobre]
[**V: ser third sing present, QSUBJ, N: país]**
[V: tener third sing present, OBJ, N: petróleo]

Según nuestro análisis, consideramos más útil la relación de sujeto de *país* respecto a *tener* que la relación de modificación de *tener* respecto a *país*. Nuestra opción nos da más información sobre la entidad por la que se está preguntando (un país). Mostrar la relación de modificación implicaría, además, tener que añadir otra relación de dependencia en la que *ser* funcionaría como modificador de *país*. De esta manera se multiplicarían por dos las relaciones de dependencia y no se añadiría información útil al grafo.

4.4.2.3.2.2 Cláusula de infinitivo: *Infinitive Clause*

Cláusula subordinada formada por un infinitivo y sus posibles argumentos.
La cláusula funciona como sujeto (129) u objeto (130) del verbo principal:

(129) # (null) 0 0-45|**Legalizar la marihuana** es un asunto complejo.
[V: ser third sing present <PRED [N: asunto <ATTR A: complejo <DET D: un] **<SUBJ [V: legalizar <OBJ [A: marihuana <DET la]]**]

(130) # (null) 1 0-58|Wert pretende **modificar por completo el sistema educativo.**
[V: pretender third sing present **<OBJ [V: modificar <OBJ [N: sistema <ATTR A: educativo <DET D: el] <CIRC X: por completo]** <SUBJ PN: Wert]

Como se observa en los ejemplos, la relación de dependencia se establece entre los verbos de las cláusulas; podemos verlo más claro en la representación por tripletes:

(131) # (null) 0 0-45|**Legalizar la marihuana** es un asunto complejo.
[A: marihuana, DET, la]
[N: asunto, ATTR, A: complejo]
[N: asunto, DET, D: un]
[V: legalizar, OBJ, A: marihuana]
[V: ser third sing present, PRED, N: asunto]
[V: ser third sing present, SUBJ, V: legalizar]

(132) # (null) 1 0-58|Wert pretende **modificar por completo el sistema educativo.**
[N: sistema, ATTR, A: educativo]
[N: sistema, DET, D: el]
[V: modificar, CIRC, X: por completo]
[V: modificar, OBJ, N: sistema]
[V: pretender third sing present, OBJ, V: modificar]
[V: pretender third sing present, SUBJ, PN: Wert]

4.4.2.3.2.3 Cláusula subordinada con *que: THAT Clause*

Está formada por el nexo *que* más una cláusula.
En la representación no se muestra el nexo y la relación de dependencia se establece entre los verbos:

(133)# (null) 2 0-33|EEUU intentó **que Israel cediese.**
[V: intentar third sing past <OBJ [**V: ceder third sing imperfect** <SUBJ PN: Israel] <SUBJ PN: EEUU]

TRIPLETES
[V: ceder third sing imperfect, SUBJ, PN: Israel]
[V: intentar third sing past, OBJ, V: ceder third sing imperfect]
[V: intentar third sing past, SUBJ, PN: EEUU]

4.4.2.3.2.4 Cláusula subordinada temporal: *Temp Clause*

Cláusula subordinada con valor temporal, constituida por un nexo de valor temporal (*cuando*, *después de*, *antes de*, etc.) y una cláusula.

Siempre funciona como adjunto. La *transduction* es similar a la de la frase preposicional, con el nexo como parámetro del *relator*. La relación de dependencia se establece entre los predicados:

(134)# (null) 3 0-50|España fue intervenida **cuando el FMI lo decidió.**
[V: intervenir third sing past <OBJ PN: España **<CIRCcuando [V: decidir third sing past <OBJ lo third masc sing <SUBJ [PN: FMI <DET D: el]]**]

TRIPLETES
[PN: FMI, DET, D: el]
[V: decidir third sing past, OBJ, lo third masc sing]
[V: decidir third sing past, SUBJ, PN: FMI]
[V: intervenir third sing past, CIRCcuando, V: decidir third sing past]
[V: intervenir third sing past, OBJ, PN: España]

Esta representación de las subordinadas temporales mediante una relación entre predicados es similar también a la del *parser* de Stanford:

SpQA:

(135)# (null) 4 0-54|Bárcenas habló **cuando El Mundo publicó sus papeles.**
[V: hablar third sing past **<CIRCcuando [V: publicar third sing past <OBJ [N: papeles <DET D: sus] <SUBJ PN: El Mundo]** <SUBJ PN: Bárcenas]

TRIPLETES
[N: papeles, DET, D: sus]
[V: hablar third sing past, CIRCcuando, V: publicar third sing past]
[V: hablar third sing past, SUBJ, PN: Bárcenas]
[V: publicar third sing past, OBJ, N: papeles]
[V: publicar third sing past, SUBJ, PN: El Mundo]

Stanford Parser:
(136) *Bárcenas spoke **when El Mundo published his notes**.*
nsubj(spoke-2, Bárcenas-1)
root(ROOT-0, spoke-2)
advmod(published-6, when-3)
nn(Mundo-5, El-4)
nsubj(published-6, Mundo-5)
advcl(spoke-2, published-6)
poss(notes-8, his-7)
dobj(published-6, notes-8)

4.4.2.3.2.5 Cláusula condicional: *COND Clause*

Cláusula subordinada con valor condicional, constituida por la conjunción *si* más una cláusula.

El indicador de función que se le asigna es COND; la relación de dependencia se establece entre los predicados:

(137) # (null) 0 0-39|Llegaría lejos **si entrara en política**
[V: llegar third sing conditional **<COND [V: entrar third sing imperfect <CIRCen N: política]** <CIRC X: lejos]

TRIPLETES
[V: entrar third sing imperfect, CIRCen, N: política]
[V: llegar third sing conditional, CIRC, X: lejos]
[V: llegar third sing conditional, COND, V: entrar third sing imperfect]

4.4.3 Módulo de las preguntas

En las secciones siguientes trataremos en detalle todo lo concerniente al análisis de las preguntas en SpQA.

En primer lugar presentaremos de modo general los tipos de preguntas que contempla la gramática e introduciremos su representación básica.

A continuación entraremos de lleno en aspectos relativos a la formalización en

SpQA. Primero, nos ocuparemos de la formalización de los aspectos lingüísticos señalados como relevantes para el análisis de preguntas en capítulos anteriores (cf. capítulo 2, sección 2.4). En las secciones posteriores, profundizaremos en la formalización de cada tipo de preguntas, tratando todos los elementos importantes para su análisis y representación.

4.4.3.1 Tipos de preguntas en SpQA y representación básica

En el capítulo 2 definimos qué son las preguntas en SpQA: oraciones interrogativas directas cuya finalidad es la demanda de información en un contexto no discursivo (cf. capítulo 2).

Teniendo en cuenta las características gramaticales y semánticas de las preguntas así entendidas, en SpQA distinguimos tres tipos de preguntas, cada una con una representación propia:

- **Preguntas totales**: en la representación se identifican con la etiqueta `ynQ` (abreviatura de *yes/no questions*).
  Tienen las características gramaticales señaladas en el capítulo 2.
  En ellas la incógnita se corresponde con el carácter afirmativo o negativo de aquello que se dice en la oración. El ámbito de la interrogación[135], por tanto, afecta a toda la oración. Por esta razón, en la representación de las totales no se marca ningún ámbito de acción para la incógnita o variable interrogativa (sobrentendiéndose que esta afecta a toda la pregunta).

(138) `# (null) 6 0-29|¿La marihuana gasta la piel?`
`ynQ` **`[V: gastar third sing present <OBJ [N: piel <DET D: la] <SUBJ [N: marihuana <DET D: la]]`**

- **Preguntas parciales:** en la representación se identifican con la etiqueta `whQ` (abreviatura de *wh questions*).
  Tienen las características gramaticales señaladas en el capítulo 2.
  En ellas el ámbito de la interrogación solo afecta a la frase interrogativa, de manera que esta constituye la incógnita que debe ser resuelta. Este hecho se manifiesta en la representación, donde el constituyente que funciona como frase interrogativa se marca con una `Q`, que se coloca antes del indicador de función sintáctica.
  SpQA lleva a cabo además una serie de transformaciones para hacer explícito el valor semántico de la incógnita.

(139) `# (null) 7 0-38|¿Dónde está el castillo de Windsor?`
`whQ [V: estar third sing present <SUBJ [N: castillo <MODde PN: Windsor <DET D: el] <`**`QCIRC LOCATION]`**

Más adelante volveremos en profundidad sobre este aspecto.

135 En los ejemplos de análisis marcamos el ámbito de la interrogación en negrita.

- **Preguntas con disyunción**: engloban las disyuntivas propiamente dichas (140) y las parciales con disyunción (141) (cf. capítulo 2).
  (140) *¿Pasteur descubrió la penicilina o la cocaína*
  (141) *¿Qué descubrió Pasteur: la penicilina o la cocaína?*
  En la representación, ambos tipos de preguntas se etiquetan como `disjQ` (abreviatura de *disjunction question*).
  Tienen las características señaladas en el capítulo 2 para las disyuntivas y las parciales con disyunción.
  En ellas el ámbito de la interrogación afecta solo a la disyunción. Este hecho se manifiesta en su representación, donde ambos tipos de preguntas se homogeneizan, marcándose la disyunción con una `Q` precediendo al indicador de función sintáctica:

(140a) `#(null) 0 0-49|¿Pasteur descubrió la penicilina o la cocaína?`
`disjQ [V: descubrir third sing past` **`<QOBJ [disjunction <DISJ [N: penicilina <DET D: la] <DISJ [N: cocaína <DET D: la]]`** `<SUBJ PN: Pasteur]`

(141a) `# (null) 1 0-55|¿Qué descubrió Pasteur: la penicilina o la cocaína?`
`disjQ [V: descubrir third sing past` **`<QOBJ [disjunction <DISJ [N: penicilina <DET D: la] <DISJ [N: cocaína <DET D: la]]`** `<SUBJ PN: Pasteur]`

Con esta representación, se pone de manifiesto que en ambas estructuras el significado subyacente es el mismo.

Como se puede observar en los ejemplos anteriores, la representación de SpQA hace explícita la relación entre foco e interrogación (recordemos que el foco se corresponde con la incógnita; cf. capítulo 2, sección 2.2.3), que es distinta para cada uno de los tipos de pregunta.

Por otra parte, la representación se sitúa en la línea de aquella que, como sabemos (cf. capítulo 2, sección 2.3.1.3), está debajo de la mayoría de las aproximaciones de la semántica formal a las preguntas parciales y totales[136]:

TOTALES
(142) *Did Lee Harvey Oswald kill John F. Kennedy?*
(142a) ?kill' (lee_harvey_oswald' , john_f_kennedy')

PARCIALES
(143) *Who killed John F. Kennedy?*
(143a) ?x1 kill' (x1, john_f_kennedy')

136 Como vimos en el capítulo 2, las aproximaciones semánticas no se han ocupado de las disyuntivas.

En el caso de las parciales SpQA enriquece la representación de (143a) al indicar los rasgos semánticos de x. Esta característica de la representación sigue el segundo postulado de Hamblin (cf. capítulo 2, sección 2.3.1.1.1) que, recordemos, sostenía que determinar el significado de las preguntas consiste en saber qué cuenta como respuesta. En SpQA representamos el significado de una pregunta parcial a través de la representación del valor semántico de su incógnita junto a toda la información sintáctica correspondiente a la pregunta; la combinación de ambos elementos determina «qué cuenta como respuesta».

El análisis propuesto también tiene conexiones con el enfoque categorial (cf. sección 2.3.1.1.3 del capítulo 2), si bien en el análisis de SpQA no se predetermina un tipo de constituyente sintáctico para la pregunta, sino que se definen una serie de características semánticas y sintácticas que debe cumplir ese constituyente.

Además, la representación de SpQA puede conectarse también con lo que explicábamos en el capítulo 2 sobre la relación entre respuesta tipo constituyente y respuesta tipo oración: para una parcial como (143), la respuesta tipo constituyente se correspondería con el valor semántico determinado por la frase interrogativa, representado en SpQA como `LOCATION` («ubicación», por ejemplo: *en Inglaterra*), mientras que la respuesta tipo oración retomaría todo el material no focalizado (*El castillo de Windsor está*) más el constituyente que «encajase» con la frase interrogativa (*en Inglaterra*).

4.4.3.2 Formalización de las preguntas en SpQA: aspectos lingüísticos señalados en capítulos anteriores

En esta sección presentaremos el tratamiento dado en SpQA a aquellos aspectos lingüísticos señalados como importantes para el análisis de preguntas en los capítulos 2 y 3.

4.4.3.2.1 Negación

En el capítulo 2 vimos que el carácter interno o externo de la negación implica significados distintos para una pregunta. Vimos también que, sin acudir a la semántica, solo es posible distinguir entre negación interna o externa cuando hay términos de polaridad positiva (externa) o negativa (interna) presentes en la pregunta.

Por otra parte, en el capítulo 3 constatamos que la negación se documenta poco en las preguntas (aparece especialmente con *por qué*) y que, en la gran mayoría de los casos en los que aparece, es de tipo interno.

En el análisis de SpQA hemos tomado la decisión de que la negación sea siempre de tipo interno:

(144)
```
# (null) 2 0-42|¿Por qué las avestruces no pueden volar?
whQ [[V: volar third plu present <AUXposib V: poder] <SUBJ [N:
avestruces <DET D: las] <CIRC X: no <QCIRC CAUSE]
```

(145) # (null) 3 0-84|¿**No** será Iker Casillas el portero oficial en el próximo partido de la selección?
ynQ [V: ser third sing future <SUBJ PN: Iker Casillas **<NEG X: no** <PRED [N: portero <ATTR A: oficial <DET D: el] <CIRCen [N: partido <MODde [N: selección <DET D: la] <ATTR A: próximo <DET D: el]]

Esta decisión en la formalización responde, en primer lugar, al hecho de que sin términos de polaridad adicionales no es posible distinguir la negación interna y externa. En segundo lugar, responde a que, como los datos demuestran (cf. capítulo 3, sección 3.3.2) la negación externa es poco esperable en un entorno de BR, especialmente en uno de tipo factual (sería distinto, por ejemplo, en un entorno de BR del ámbito de la opinión y las creencias).

4.4.3.2.2 Orden de constituyentes

4.4.3.2.2.1 Totales

Con el sujeto explícito hay dos órdenes posibles:

- verbo + sujeto + argumentos;
- sujeto + verbo + argumentos.

En el capítulo 3 vimos que el segundo tipo, con el sujeto antepuesto, es también muy común y que no parecía implicar ningún significado especial para la pregunta.

En SpQA se recogen los dos órdenes posibles sin ningún matiz en la representación que los diferencie:

- verbo + sujeto + argumentos:

(148) # (null) 0 0-52|¿Dimitirá Mariano Rajoy de su cargo de presidente?
ynQ [V: dimitir third sing future <PCde [N: cargo <MODde N: presidente <DET D: su] <SUBJ PN: Mariano Rajoy]

- sujeto + verbo + complementos:

(149) # (null) 1 0-52|¿Mariano Rajoy dimitirá de su cargo de presidente?
ynQ [V: dimitir third sing future <PCde [N: cargo <MODde N: presidente <DET D: su] <SUBJ PN: Mariano Rajoy]

4.4.3.2.2.2 Parciales

En el capítulo 2 vimos que las parciales pueden presentar un orden prototípico o varios órdenes alternativos. El estudio de corpus del capítulo 3 nos mostró que los

órdenes alternativos tienen muy baja incidencia. Entre ellos, el más documentado es la anteposición tipo B.

En SpQA se recogen todos los órdenes documentados:

- **Orden prototípico**:

(150) `# (null) 2 0-34|¿Cuándo será juzgado Bárcenas?`
`whQ [V: juzgar third sing future <OBJ PN: Bárcenas <QCIRC TIME]`

- **Anteposición tipo A**: un análisis[137] de los casos de Gayo (2010) muestra que la anteposición tipo A se documenta allí con el sujeto (83,7%), el complemento preposicional (11,6%) y el complemento indirecto (4,6%). En nuestros corpus, por otra parte, se documentaba la tematización del sujeto, el complemento indirecto y el circunstancial (cf. capítulo 3, sección 3.3.1.2.2.1)
  A la luz de los datos, en SpQA hemos optado por la «sobregeneración» y hemos establecido reglas para todos los argumentos en posición tematizada con cualquier interrogativo.

(151) `# (null) 3 0-26|¿`**`Rubalcaba`** `cuándo hablará?`
`whQ [V: hablar third sing future` **`<SUBJ PN: Rubalcaba`** `<QCIRC TIME]`

(152) `# (null) 4 0-26|¿`**`En Madrid`** `cuándo nieva?`
`whQ [V: nevar third sing present <QCIRC TIME` **`<CIRCen PN: Madrid`**`]`

(153) `# (null) 5 0-36|¿`**`Al político`** `quién le paga los trajes?`
`whQ [V: pagar third sing present <IOBJ le third GENDER sing` **`<IOBJ [N: político <DET D: el]`** `<QSUBJ NAME +PERSON sing <OBJ [N: trajes <DET D: los]]`

(154) `# (null) 6 0-56|¿`**`La cabeza`** `de qué se la llenó el político al pueblo?`
`whQ [V: llenar third sing past <OBJ lo third fem sing <IOBJ le third masc sing` **`<OBJ [N: cabeza <DET D: la]`** `<IOBJ [N: pueblo <DET D: el] <QPCde ENTITY <SUBJ [N: político <DET D: el]]`

(155) `# (null) 7 0-35|¿`**`Culpable`** `de la crisis quién fue?`
`whQ [V: ser third sing past` **`<PRED [A: culpable <MODde [N: crisis <DET D: la]]`** `<QSUBJ NAME +PERSON sing]`

137 Datos no publicados.

(156) # (null) 8 0-38|¿**De política** cuándo habla la gente?
whQ [V: hablar third sing present **<PCde N: política** <SUBJ [N: gente <DET D: la] <QCIRC TIME]

- **Anteposición tipo B**: en Gayo (2010) se anteponen el sujeto (mayoría de los casos) y el objeto directo, siempre con el interrogativo funcionando como adjunto. En nuestros corpus se documenta solo el sujeto tematizado, con el interrogativo funcionando como *QCIRC* (cf. capítulo 3, sección 3.3.1.2.2.2).
  En este caso, teniendo en cuenta la mayor homogeneidad que a la luz de los datos parece presentar la anteposición tipo B, hemos implementado el sujeto y el objeto directo antepuestos tras el *QCIRC*.

(157) # (null) 9 0-43|¿Desde cuándo **Portugal** es una república?
whQ [V: ser third sing **<SUBJ PN: Portugal** <PRED [N: república <DET D: una] <QCIRCdesde TIME]

(158) # (null) 11 0-50|¿Por qué **la revolución** la hacen los marginados?
whQ [V: hacer third plu present <OBJ lo third fem sing **<OBJ [N: revolución <DET D: la]** <SUBJ [A: marginados <DET los] <QCIRC CAUSE]

Los adverbios antes del interrogativo (*aproximadamente*, *exactamente*, *actualmente*) se analizan como *CIRC* antepuesto (anteposición tipo A). No se mantiene la distinción que hicimos en el capítulo 3 (sección 3.3.1.2.2.3) entre «adverbios propiamente tematizados»:

(159) # (null) 13 0-41|¿**Actualmente** quién gobierna en España?
whQ [V:gobernar third sing present <CIRCen PN: España <QSUBJ NAME +PERSON sing **<CIRC X: actualmente**]

y adverbios colocados al inicio de la estructura que modifican de algún modo el interrogativo (*aproximadamente*, *exactamente*):

(160) # (null) 16 0-55|¿**Exactamente** cuándo será el próximo eclipse de sol?
whQ [V: ser third sing future <SUBJ [N: eclipse <MODde N: sol <ATTR A: próximo <DET D: el] <QCIRC TIME **<CIRC X: exactamente**]

Hemos considerado cualquier adjunto antes del interrogativo como *CIRC* porque creemos que la distinción citada no aporta nada significativo a la representación de la pregunta.

El adjetivo *cada* antes del interrogativo lo hemos implementado solo con *cuánto*

funcionando como *QCIRC*. Su representación es igual a la de las preposiciones o a la de la conjunción temporal *cuando*, es decir, como parámetro del *relator*:

(161) `# (null) 17 0-60|¿**Cada** cuánto tiempo hay un eclipse de sol aproximadamente?`
`whQ [V: haber third sing present **<QCIRCcada** [QUANTITY <MODde N: tiempo] <OBJ [N: eclipse <MODde N: sol <DET D: un] <CIRC X: aproximadamente]`

4.4.3.2.3 Rasgos morfosintácticos de las partículas interrogativas

Este aspecto lo trataremos en profundidad cuando presentemos el módulo de las parciales (cf. sección 4.4.3.4.2.1).

4.4.3.2.4 Subordinación a distancia

Vimos en el capítulo 2 que en algunos casos la partícula interrogativa puede asociarse con el verbo principal y con el verbo subordinado, y que solo en determinadas construcciones es posible saber de cuál depende realmente. Recordemos:

- **Interrogativo funcionando como adjunto**: siempre que existe compatibilidad semántica entre el interrogativo y los verbos de principal y subordinada no es posible saber a qué verbo se asocia el interrogativo.
  En SpQA asociamos por defecto el interrogativo al verbo principal:

(162) `# (null) 18 0-53|¿Cuándo dijo Rajoy que el gobierno subiría el IVA?`
`whQ [V: decir third sing past <OBJ [V: subir third sing conditional <OBJ [PN: IVA <DET D: el] <SUBJ [N: gobierno <DET D: el]] <SUBJ PN: Rajoy <QCIRC TIME]`

TRIPLETES
`[N: gobierno, DET, D: el]`
`[PN: IVA, DET, D: el]`
`[V: decir third sing past, OBJ, V: subir third sing conditional]`
`[V: decir third sing past, QCIRC, TIME]`
`[V: decir third sing past, SUBJ, PN: Rajoy]`
`[V: subir third sing conditional, OBJ, PN: IVA]`
`[V: subir third sing conditional, SUBJ, N: gobierno]`

- **Interrogativo funcionando como sujeto**: siempre que se da concordancia entre el interrogativo y los verbos, se produce ambigüedad. En SpQA en estos casos asociamos el sujeto siempre al verbo principal.

(163) # (null) 20 0-58|¿Quién dijo que movería el mundo con un punto de apoyo?
whQ [V: decir third sing past <OBJ [V: mover third sing conditional <OBJ [N: mundo <DET D: el] <CIRCcon [N: punto <MODde N: apoyo <DET D: un]] <QSUBJ NAME +PERSON sing]

TRIPLETES
[N: mundo, DET, D: el]
[N: punto, DET, D: un]
[N: punto, MODde, N: apoyo]
[V: decir third sing past, OBJ, V: mover third sing conditional]
[V: decir third sing past, QSUBJ, NAME +PERSON sing]
[V: mover third sing conditional, CIRCcon, N: punto]
[V: mover third sing conditional, OBJ, N: mundo]

Lógicamente, si se da concordancia solo con uno de los verbos, el interrogativo se asocia a ese verbo:

(164) # (null) 21 0-81|¿Quiénes se pensaba que ganarían las últimas elecciones generales en España?
whQ [V: pensar third sing imperfect <SUBJ IMP <OBJ [[V: ganar third plu conditional <OBJ [N: elecciones <ATTR A: generales <ATTR A: últimas <DET D: las] <CIRCen PN: España] <QSUBJ NAME +PERSON plu]]

TRIPLETES
[N: elecciones, ATTR, A: generales]
[N: elecciones, ATTR, A: últimas]
[N: elecciones, DET, D: las]
[V: ganar third plu conditional, CIRCen, PN: España]
[V: ganar third plu conditional, OBJ, N: elecciones]
[V: ganar third plu conditional, QSUBJ, NAME +PERSON plu]
[V: pensar third sing imperfect, OBJ, V: ganar third plu conditional]
[V: pensar third sing imperfect, SUBJ, IMP]

- **Interrogativo funcionando como indirecto**: como el indirecto interrogado puede asociarse con cualquiera de los predicados, en SpQA se asocia por defecto al verbo principal:

(165) `# (null) 22 0-29|¿A quién dijo que sonriera?`
`whQ [V: decir third sing <QIOBJ PERSON <OBJ V: sonreír third sing]`

TRIPLETES
`[V: decir third sing, OBJ, V: sonreír third sing]`
`[V: decir third sing, QIOBJ, PERSON]`

En conclusión: excepto cuando tenemos alguna marca explícita de que el interrogativo se asocia a un verbo concreto (concordancia del sujeto), el interrogativo se asocia por defecto al verbo principal.

4.4.3.2.5 Verbo no finito

Para las parciales hemos implementado la posibilidad de tener un verbo en infinitivo como verbo principal:

(166) `# (null) 23 0-39|¿Dónde comprar medicamentos en India?`
`whQ [V: comprar <OBJ N: medicamentos <CIRCen PN: India <QCIRC LOCATION]`

(167) `# (null) 6 0-34|¿Hasta cuándo tomar el biberón?`
`whQ [V: tomar <OBJ [N: biberón <DET D: el] <QCIRChasta TIME]`

4.4.3.3 Aspectos semánticos generales

Además de en el procesamiento de la frase interrogativa, SpQA maneja información semántica en el reconocimiento de tres tipos de estructuras:

- **Entidades Nombradas** (NER): hemos reiterado la importancia de NER en las tareas de BR. Por esta razón, este es uno de los aspectos semánticos a los que se le ha dado prioridad en SpQA.

En cuanto a la implementación, en la sección en la que describimos la frase nominal en SpQA vimos que en nuestra representación las entidades nombradas se marcan con la categoría `PN` (cf. sección 4.4.2.2.1).

(168) `# (null) 0 0-42|¿Cuándo hay rebajas en El Corte Inglés?`
`whQ [V: haber third sing present <QCIRC TIME <OBJ N: rebajas <CIRCen` **`PN: El Corte Inglés`**`]`

Hemos explicado también que además de sus propias reglas para el reconocimiento de entidades nombradas, SpQA cuenta con mecanismos que le permiten manejar la salida del sistema de NER de *Freeling*, más eficiente que el de nuestra gramática (cf. sección 4.4.2.2.1).

- **Estructuras que implican cuantificación**: en el capítulo 1 también hemos tratado la importancia del manejo de la cuantificación en los sitemas de BR (cf. capítulo 1, sección 2.4.1.3).
  Como hemos visto en secciones anteriores, la cuantificación se marca en SpQA mediante dos elementos: una clase de palabras propia para los determinantes cuantificadores, `Q` y un indicador de función para las relaciones de cuantificación, `QUANT`.

(169) `# (null) 1 0-49|¿Qué` **`tres países`** `no participaron en la cumbre?`
`whQ [V: participar third plu past <PCen [N: cumbre <DET D: la] <NEG X: no <QSUBJ [`**`N: países <QUANT Q: tres`**`]]`

- **Estructuras que implican ubicación temporal**: al igual que el reconocimiento de entidades nombradas y de relaciones de cuantificación, la identificación de estructuras que ubican temporalmente un evento también es importante en BR.
  Hemos visto que en SpQA contamos con un indicador de función específico para la ubicación temporal, `DATE`. Recordemos que `DATE` se utiliza cuando tenemos una *NPdate* (cf. sección 4.4.2.2.1) o una *PPdate* (cf. sección 4.4.2.2.5) y que las fechas (día + mes, día + mes + año) se homogeneizan en un único formato mediante la *transduction*.

(170) `# (null) 2 0-41|¿Qué ocurrió` **`el 14 de octubre de 1492`**`?`
`whQ [V: ocurrir third sing past` **`<DATE 14-10-1492`** `<QSUBJ ENTITY]`

(171) `# (null) 0 0-37|¿Qué ocurrió` **`el 14 del 10 de 1492`**`?`
`whQ [V: ocurrir third sing past` **`<DATE 14-10-1492`** `<QSUBJ ENTITY]`

4.4.3.4 Módulos específicos para cada tipo de pregunta en SpQA

4.4.3.4.1 Las preguntas totales

Se identifican como `ynQ` en la representación (172).

La incógnita se refiere al carácter afirmativo o negativo de la predicación, así que en SpQA la incógnita abarca toda la estructura (172).

(172) `# (null) 1 0-68|¿Es verdad que a Vilardo le inocularon un rabdomiosarcoma olfativo?`
`ynQ [V: ser third sing present <PRED N: verdad <SUBJ [V: inocular third plu past <IOBJ le third GENDER sing <IOBJ PN: Vilardo <OBJ [UNKNOWN N: rabdomiosarcoma <ATTR A: olfativo <DET D: un]]`

El módulo de las totales permite cualquier combinación de argumentos posverbales:

(173) `# (null) 2 0-51|¿El presidente le debe explicaciones al pueblo en España?`
`ynQ [V: deber third sing present <IOBJ le third masc sing <OBJ N: explicaciones <IOBJ [N: pueblo <DET D: el] <CIRCen PN: España <SUBJ [N: presidente <DET D: el]]`

(174) `# (null) 1 0-59|¿Le debe explicaciones al pueblo en España el presidente?`
`ynQ [V: deber third sing present <IOBJ le third masc sing <OBJ N: explicaciones <IOBJ [N: pueblo <DET D: el] <CIRCen PN: España <SUBJ [N: presidente <DET D: el]]`

4.4.3.4.2 Las preguntas parciales

El tratamiento de las parciales es bastante más complejo que el de las totales. Como sabemos, en las parciales la incógnita de la pregunta se define en la frase interrogativa. Es por eso por lo que este constituyente y, en especial, las partículas interrogativas, constituyen la clave en el procesamiento de las preguntas parciales en SpQA.

En las secciones siguientes nos ocuparemos de los distintos aspectos implicados en el análisis de las preguntas parciales en SpQA. En primer lugar trataremos la formalización de los aspectos morfosintácticos de las partículas interrogativas. A continuación nos ocuparemos de los aspectos semánticos y de la representación de la frase interrogativa en SpQA.

4.4.3.4.2.1 Aspectos morfosintácticos de las partículas interrogativas

Al final del capítulo 2 concluíamos que todos los aspectos morfosintácticos señalados nos parecían relevantes para la implementación de las partículas interrogativas en SpQA. En esta sección retomaremos esos aspectos, deteniéndonos en los detalles de su formalización.

4.4.3.4.2.1.1 Clases de palabras

Las partículas interrogativas implementadas son las que hemos tratado hasta el momento: *qué*, *quién*, *cuál*, *cuánto* (con la variante *cuán*), *cómo*, *dónde* (más la preposición *a,* con la variante *adónde*), *cuándo* y *por qué*.

En cuanto a la clase de palabras correspondiente a cada partícula interrogativa, en SpQA distinguimos los siguientes casos:

- Pronombres: *quién*, *qué*, *cuánto* y *cuál*.
- Determinantes: *qué*, *cuánto* y *cuál*.
- Adjetivo (cuantificador): *cuán* (Real Academia Española, 2009, 22.1f y ss.).
- Adverbios: *cuánto*, *cuándo*, *cómo*, *dónde* (con la variante *adónde*) y *por qué*.

En SpQA las partículas del último grupo siempre funcionan como adjuntos (excepto *cuánto*), mientras que las de los dos primeros grupos funcionan prototípicamente como argumentos. *Cuán* funciona siempre como modificador. Más adelante volveremos sobre estas cuestiones.

4.4.3.4.2.1.2 Tipos de frases interrogativas

Las partículas interrogativas, combinadas con otras unidades, pueden constituir los siguientes tipos de frases en SpQA:

- **Frase nominal**:
  - Interrogativo funcionando como núcleo (pronombre): *quién*, *qué*, *cuál* y *cuánto*. El interrogativo puede ser modificado por una frase preposicional o una frase relativa.

(175) ¿Qué que es fundamental en la vida actual *descubrió Pasteur?*
(176) ¿Quién *inventó el teléfono?*
(177) ¿Cuántos de los ganadores del Tour *fueron acusados de dopaje?*

  - Interrogativo funcionando como determinante: *qué*, *cuál* y *cuánto*.

(178) ¿Qué coche *alcanza una velocidad más alta hoy en día?*
(179) ¿Cuál río[138] *es el más largo del mundo?*
(180) ¿Cuánto ácido fólico *debe tomar una embarazada diariamente?*

La frase nominal suele tener funciones de tipo argumental.

138 Hemos incluido en la gramática el uso de *cuál* como determinante a pesar de que, como vimos en el capítulo 1 (sección 2.4.3.1), este uso en Europa es poco común y más propio de América.

- **Frase preposicional**: todos los interrogativos excepto *por qué*[139] pueden ir precedidos de preposición[140].

(181) ¿Con cuántas carabelas *partió Colón?*
(182) ¿Hasta dónde *se extendía el imperio de Alejandro Magno?*

En la combinación de preposición e interrogativo nos hemos inclinado, excepto en casos concretos, por la «sobregeneración», es decir: hemos considerado que cualquier preposición se puede combinar con cualquier interrogativo. Esta decisión no genera problemas y es más sencilla para la implementación (sobre todo teniendo en cuenta que la restricción combinatoria no aporta nada a la gramática). En ciertos casos, no obstante, hemos limitado las preposiciones que se pueden combinar con un interrogativo para un determinado valor semántico; trataremos estas combinaciones en la sección relativa a los aspectos semánticos (cf. *infra*).

La frase preposicional funciona siempre como adjunto (183), modificador (184), complemento preposicional (185) o predicativo (186):

(183) ¿Para qué *subió el IVA el gobierno?*
(184) ¿*Dentro* de dónde *se encuentra la Mona Lisa?*
(185) ¿A qué *tiende el gobierno español?*
(186) ¿De dónde *era Colón?*

- **Frase adverbial y adjetiva**: son posibles con la variante de *cuánto*, *cuán*:

(187) ¿Cuán largo *era ese pasaje?*
(188) ¿Cuán lejos *está el Auditorio de la Facultad de Filología?*

La construcción «cómo así» (cf. capítulo 2, sección 2.2.4.3.1), propia de zonas de Centroamérica y probablemente más próxima al lenguaje oral no se ha implementado en la gramática[141].

4.4.3.4.2.1.3 Frases interrogativas – función sintáctica

Para cada interrogativo hemos recogido las posibilidades de función sintáctica que señalábamos en el capítulo 2:

Qué

Puede desempeñar todas las funciones sintácticas:

- Sujeto: como frase nominal.
  (189) ¿Qué ejército *ocupó Haití?*
  (190) ¿Qué *provocó las inundaciones de 1976?*

139 Hemos incluido, por tanto, la construcción *a cómo*, aunque nos parezca más bien propia del lenguaje oral. Cf. *infra* para más detalles.
140 Con *preposición* nos referimos también a las locuciones preposicionales del tipo *a partir de*.
141 Previamente hemos comprobado que no se recogía en ninguno de nuestros corpus de preguntas, ni siquiera en Wiki.

- Objeto directo: como frase nominal o frase preposicional con *a*.
  (191) ¿Qué institución *constituyeron Brasil, Portugal, Angola, Mozambique, Santo Tomé y Príncipe, Guinea-Bissau y Cabo Verde en Brasilia?*
  (192) ¿A qué político *ha denunciado Bárcenas?*

- Objeto indirecto: como frase preposicional con *a*.
  (193) ¿A qué banco *le debe más dinero la diputación de Pontevedra?*

- Complemento preposicional: como frase preposicional.
  (194) ¿De qué *hablaron Rajoy y Obama en su último encuentro?*

- Predicativo: como frase nominal o frase preposicional con *de*.
  (195) ¿Qué *es la quinua?*
  (196) ¿De qué *es el helado de* straciatella*?*

- Circunstancial: como frase preposicional o nominal.
  (197) ¿En qué ciudad *explotó una carta bomba?*
  (198) ¿Qué día *se celebra la independencia de EEUU?*

- Modificador: como frase preposicional.
  (199) ¿De qué *están* hechos *los implantes de mama?*
  (200) *¿Dentro* de qué estado *está Columbus?*

Quién

También puede desempeñar todas las funciones sintácticas:

- Sujeto: como frase nominal.
  (201) ¿Quién *dirigió El Pianista?*
- Objeto directo: como frase preposicional con *a.*
  (202) ¿A quién *acusó El País de estafa recientemente?*

- Objeto indirecto: como frase preposicional con a.
  (203) ¿A quién *le mintió descaradamente el PP?*

- Complemento preposicional: como frase preposicional.
  (204) ¿De quién *le habló Rajoy a Bush?*

- Circunstancial: como frase preposicional.
  (205) ¿Con quién *viajó a París Carolina de Mónaco en 1998?*

- Predicativo: como frase nominal o frase preposicional con *de.*
  (206) ¿Quién *es Ulises?*
  (207) ¿De quién *es El Ulises?*

- Modificador: como frase preposicional con *de*.
  (208) *¿Detrás* de quién *está Brad Pitt en la lista de los más guapos del mundo?*

Cuál

Puede desempeñar todas las funciones sintácticas:

- Sujeto: como frase nominal.
  (209) ¿Cuál de los participantes *ganó el último concurso de El precio justo?*
  (210) ¿Cuál director *ganó el Oso de oro en la Berlinale de 2012?*

- Objeto directo: como frase nominal o frase preposicional con *a*.
  (211) ¿Cuál empresa *llevó a la ruina Ruiz Mateos?*
  (212) ¿A cuál *actor eligió Roman Polanski para protagonizar El pianista?*

- Objeto indirecto: como frase preposicional con *a*.
  (213) ¿A cuál banco *le debe más dinero el PSOE?*

- Predicativo: como frase nominal o frase preposicional con *de*.
  (214) ¿Cuál de los políticos del PP *es Soraya Sáez de Santamaría?*
  (215) ¿De cuál de los escritores de la Generación del 27 *es el Romancero Gitano?*

- Complemento preposicional: como frase preposicional.
  (216) ¿Con cuál película *ganó Spielberg su primer Óscar?*

- Circunstancial: como frase preposicional.
  (217) ¿En cuál país *está Buruti?*

- Modificador: como frase preposicional con *de*:
  (218) ¿De cuál empresa *es* fundador *Amancio Ortega?*
  (219) *¿Dentro* de cuál estado *está Chicago?*

Cuánto

Al igual que las partículas anteriores puede desempeñar todas las funciones sintácticas:

- Sujeto: como frase nominal.
  (220) ¿Cuántos países *dependen de la Troika?*
  (221) ¿Cuántos de los miembros *de la UE están en crisis?*

- Objeto directo: como frase nominal o frase preposicional con *a*.
  (222) ¿Cuánto dinero *gastó Santiago en la ciudad de la cultura?*
  (223) ¿A cuántos políticos *acusó Garzón durante su mandato?*

- Objeto indirecto: como frase preposicional con *a*.
  (224) ¿A cuántas empresas *les debe dinero la Diputación de Pontevedra?*

- Predicativo: frase nominal, frase adjetiva (*cuán*) o frase preposicional con *de*.
  (225) ¿Cuánto de todo lo que ocurre actualmente en España *es responsabilidad del gobierno?*
  (226) ¿Cuán preocupante *es la situación actual en nuestro país?*
  (227) ¿De cuántos Beatles *es la canción A Day in the Life?*

- Complemento preposicional: como frase preposicional.
  (228) ¿De cuántos problemas *habló Rajoy en su última comparecencia?*

- Circunstancial: como frase preposicional.
  (229) ¿En cuántos países *es el euro la moneda oficial?*

- Modificador: como frase preposicional con *de*:
  (230) ¿De cuántas empresas *es fundador Amancio Ortega?*
  (231) ¿Dentro de cuántos años *serán posibles los viajes en el tiempo?*

Dónde

Solo funciona como adjunto: como frase adverbial o frase preposicional.
(232) ¿Dónde *se iniciaron las manifestaciones de la llamada Primavera Árabe?*
(233) ¿Hacia dónde *se dirigía Colón en un principio?*

La frase preposicional *a + dónde* tiene la variante *adónde*:
(234) ¿A dónde *se dirigía Colón?* / ¿Adónde *se dirigía Colón?*

Cuándo

Solo funciona como adjunto: como frase adverbial o frase preposicional.
(235) ¿Cuándo se *descubrió la Antártida?*
(236) ¿Desde cuándo *España es una democracia?*

Cómo

Puede desempeñar las siguientes funciones:

- Circunstancial: frase adverbial o frase preposicional.
  (237) ¿Cómo *se prepara una paella?*
  (238) ¿A cómo *estaba el euro respecto al dólar el 5 de julio de 2013?*

- Predicativo: frase adverbial.
  (239) ¿Cómo *es Leonardo Di Caprio?*
  (240) ¿Cómo de alto *es un Citroen Picasso?*

Por qué

Solo funciona como adjunto en frases adverbiales:
(241) ¿Por qué *se produjo la crisis?*

Ambigüedad

Lógicamente, el hecho de que una misma construcción pueda desempeñar distintas funciones sintácticas genera ambigüedad. La ambigüedad es estructural y semántica en ciertos casos:

(242) *¿Qué causó la crisis?* > dos interpretaciones igualmente válidas:
(242a) Algo causó la crisis. (Frase interrogativa = sujeto)
(242b) La crisis causó algo. (Frase interrogativa = objeto directo)

En otros casos la ambigüedad es solo estructural (243) y podría resolverse manejando cierta información semántica[142]:

(243) *¿Qué empresa fabrica el Escarabajo?* > la interpretación correcta (243a) solo es posible manejando información semántica (a saber, que generalmente son las empresas las que fabrican algo); sin esta información, cabe también (243b).

(243a) La empresa X fabrica el Escarabajo. (Frase interrogativa = sujeto)
(243b) El Escarabajo fabrica la empresa X. (Frase interrogativa = objeto)

En esta posible confluencia de funciones en una misma estructura destaca muy especialmente la ambigüedad entre sujeto y objeto directo, sobre todo para las construcciones con *qué + fn* (cf. capítulo 5, sección 5.3.2).

4.4.3.4.2.2 Aspectos semánticos y representación de la frase interrogativa

4.4.3.4.2.2.1 Estructura general de la representación

El análisis de la frase interrogativa es el que más se aleja de la representación puramente sintáctica para acercarse a la representación semántica. En primer lugar, en nuestra representación la partícula interrogativa como unidad léxica no aparece en el grafo de dependencias.

Como vimos en los capítulos anteriores, en las interrogativas parciales la partícula interrogativa genera una incógnita, un vacío que solo puede ser llenado por una estructura (la respuesta) con unas determinadas características semánticas (más o menos concretas dependiendo del interrogativo). Esta es, como sabemos, la base de la representación de las interrogativas parciales en la mayoría de los planteamientos de semántica formal.

142 Que SpQA no maneja.

En nuestro análisis intentamos determinar el valor semántico que denota el interrogativo, es decir: el valor semántico de la variable o incógnita que representa el interrogativo. Este valor coincide además con el del tipo de respuesta esperada (punto en el que seguimos el segundo postulado de Hamblin, cf. *supra*). Por esta razón, para la representación del significado de la pregunta, el mero valor léxico del interrogativo no nos parece relevante, ya que no añade nada a dicha representación. Lo que nos parece relevante de las partículas interrogativas son los valores semánticos asociados a ellas además de su función sintáctica. Estas dos informaciones son, en consecuencia, las que recoge nuestra representación de las frases interrogativas: función sintáctica y valor semántico de la frase interrogativa.

La representación de la frase interrogativa en SpQA consta por tanto de dos partes:

- indicador de función sintáctica;
- representación sintáctico semántica de la frase interrogativa = núcleo semántico de la frase interrogativa + información sintáctica en ciertas construcciones.

Indicador de función sintáctica

El indicador de función sintáctica va primero, integrado en el *relator* y precedido de la letra `Q`. Así, por ejemplo, la frase interrogativa funcionando como objeto directo pasaría de `OBJECT` a `QOBJECT`. De esta manera indicamos en la representación que el constituyente así marcado es la frase interrogativa, el elemento que contiene la incógnita, la variable que se ha de determinar. Al indicador de función sintáctica le sigue la representación sintáctico semántica de la frase interrogativa. Veamos un ejemplo sencillo:

(244) `# (null) 4 0-43|¿`**`Dónde`** `están las cataratas del Niágara?`
`whQ [V: estar third plu present <SUBJ [N: cataratas <MODde [PN: Niágara <DET D: el] <DET D: las]` **`<QCIRC LOCATION]`**

En el ejemplo, el indicador de función sintáctica es `QCIRC`, y la representación sintáctico semántica que le corresponde a *dónde* es `LOCATION`.

Cuando la frase interrogativa funciona como modificador:

(245) ¿*Dentro* de qué *vive el cangrejo ermitaño?*

colocamos la `Q` al nivel del constituyente principal:

(245a) `# (null) 5 0-44|¿`**`Dentro de qué`** `vive el cangrejo ermitaño?`
`whQ [V: vivir third sing present <SUBJ [N: cangrejo <ATTR A: ermitaño <DET D: el] <`**`QCIRC [X: dentro <MODde ENTITY]`**`]`

Utilizamos esta representación porque consideramos que en este caso la incógnita abarca todo el constituyente y no solo la frase interrogativa.

Si la frase interrogativa va precedida de preposición, esta se muestra en general como parámetro del *relator*:

(246) `# (null) 6 0-46|¿`**`En dónde`** `están las cataratas del Niágara?`
`whQ [V: estar third plu present <SUBJ [N: cataratas <MODde [PN: Niágara <DET D: el] <DET D: las]` **`<QCIRCen LOCATION`**`]`

Cuando tratemos los valores semánticos de los interrogativos veremos que en una serie de casos muy concretos la función sintáctica se especifica también semánticamente. Observemos un ejemplo:

(247) `# (null) 7 0-39|¿En qué país nació Albert Einstein?`
`whQ [V: nacer third sing past <SUBJ PN: Albert Einstein` **`<QCIRC_LOCATIONen`** `N: país]`

En el ejemplo anterior se especifica el valor semántico de la función sintáctica general `QCIRC` mediante la adición del valor semántico `LOCATION`. En la siguiente sección detallaremos los casos en los que usamos esta representación.

Representación sintáctico semántica de la frase interrogativa

La representación de la frase interrogativa consta de:

- información sintáctica: información relativa a determinantes o modificadores del núcleo sintáctico semántico;
- núcleo semántico: expresa el valor semántico de la frase interrogativa (`LOCATION` en (246)).

Si el interrogativo es modificado (por una frase preposicional o relativa), el modificador aparece en la representación:

(248) `# (null) 8 0-58|¿`**`Cuál de los presidentes españoles`** `gobernó más años?`
`whQ [V: gobernar third sing past <OBJ [N: años <QUANT Q: más]` **`<QSUBJ [ENTITY <MODde [N: presidentes <ATTR A: españoles <DET D: los]]`**`]`

Si el modificador es una frase de relativo, su representación es la que hemos visto en la sección 4.2.3.2.1.

(249) `# (null) 9 0-42|¿Cuánto que se sepa ha robado Bárcenas?`
`whQ [V: robar third sing present_perfect <QOBJ [QUANTITY` **`>OBJ [V: saber third sing present]`**`] <SUBJ PN: Bárcenas]`

Los determinantes que se combinan con el interrogativo también se muestran en la representación:

(250) `# (null) 0 0-44|¿Con qué` **`otro`** `nombre se conoce a Maradona?`
`whQ [V: conocer third sing present <SUBJ IMP <QCIRCcon [N: nombre` **`<DET D: otro`**`] <OBJ PN: Maradona]`

Núcleo semántico: interrogativo como núcleo *vs.* interrogativo como determinante

La representación del núcleo semántico varía dependiendo de si el interrogativo funciona como determinante o como núcleo sintáctico (pronombre o adverbio).

Cuando el interrogativo funciona como núcleo sintáctico (251) la representación consiste en el valor semántico al que el interrogativo se asocia. Por ejemplo, *qué* se asocia de modo general al valor semántico `ENTITY` (entidad):

NÚCLEO SINTÁCTICO

(251) `# (null) 10 0-41|¿`**`Qué`** `causa miles de muertes en África?`
`whQ [V: causar third sing present <OBJ [N: miles <MODde N: muertes]` **`<QSUBJ ENTITY`** `<CIRCen PN: África]`

En algunos casos, el valor del núcleo semántico se concreta más mediante la adición de ciertas especificaciones de significado:

(252) `# (null) 46 0-26|¿Quién es Penélope Cruz?`
`whQ [V: ser third sing present <QPRED` **`DESCRIPTION +PERSON`** `<SUBJ PN: Penélope Cruz]`

Cuando esto ocurre, el valor semántico central o nuclear es el que va primero (`DESCRIPTION` en el ejemplo), mientras que la información precedida de un «+» añade matices de significado a ese valor principal (en este caso, `PERSON`). Volveremos sobre esto más abajo.

Cuando el interrogativo funciona como determinante (252), excepto en el caso de *cuánto* (*vid*. *infra*), la representación de la frase interrogativa consiste en la palabra que constituye el núcleo sintáctico de la frase interrogativa y no se muestra ninguna información sobre el interrogativo. Siguiendo con el ejemplo de *qué*:

DETERMINANTE

(253) # (null) 11 0-51|¿**Qué ciudad** en Australia tiene selvas tropicales?
whQ [V: tener third sing present <CIRCen PN: Australia <OBJ [N: selvas <ATTR A: tropicales] **<QSUBJ N: ciudad**]

Esta representación se aplica a *qué* y *cuál*.

La elección de no representar la partícula interrogativa con *qué* y *cuál* funcionando como determinante se debe a que consideramos que esta no aporta ningún contenido a la representación de la frase interrogativa, pues en estos casos la entidad por la que se pregunta está explícita:

(254) # (null) 12 0-37|¿**Qué torneos** ganó Andrei Medvedev?
whQ [V: ganar third sing past **<QOBJ N: torneos** <SUBJ PN: Andrei Medvedev]

(2545 # (null) 13 0-48|¿**Cuál ciudad de España** tiene más habitantes?
whQ [V: tener third sing present <OBJ [N: habitantes <QUANT Q: más] **<QSUBJ [N: ciudad <MODde PN: España]**]

Como vimos en el capítulo 2, en estos casos el interrogativo solo aporta el contenido semántico de «identificación de la entidad explícita». Por eso hemos optado por no representarlo.

Para *cuánto* funcionando como determinante la representación es distinta. En estos casos, consideramos que el núcleo semántico de la frase interrogativa es el concepto de 'cantidad' (QUANTITY). Por eso, el núcleo sintáctico de la frase interrogativa lo representamos como modificador (en forma de frase preposicional) de este núcleo semántico representado por QUANTITY:

(256) # (null) 14 0-69|¿**Cuánto ácido fólico** debe tomar una mujer embarazada diariamente?
whQ [[V: tomar third sing present <AUXoblig V: deber] **<QOBJ [QUANTITY <MODde PN: ácido fólico]** <SUBJ [N: mujer <ATTR A: embarazada <DET D: una] <CIRC X: diariamente]

Optamos por esta representación porque, como decíamos, consideramos que el núcleo semántico de la frase interrogativa es en realidad el concepto de cantidad y no el propio núcleo sintáctico. Comparemos los siguientes ejemplos:

(257a) # (null) 15 0-37|¿**Qué torneos** ganó Andrei Medvedev?
whQ [V: ganar third sing past **<QOBJ N: torneos** <SUBJ PN: Andrei Medvedev]

(257b) # (null) 16 0-41|¿**Cuántos torneos** ganó Andrei Medvedev?
whQ [V: ganar third sing past **<QOBJ [QUANTITY <MODde N: torneos]** <SUBJ PN: Andrei Medvedev]

En (256) se nos pide la identificación de los torneos que ganó Andrei Medvedev, mientras que en (257) se nos pide el número de torneos que ganó Andrei Medvedev, es decir, el núcleo semántico de la incógnita no son los torneos en sí mismos sino el *número de torneos*. Nuestra representación muestra esta diferencia, representando *cuánto* + fn (*cuántos torneos*) como «número de» + fn (*número de torneos*).

En la sección siguiente presentaremos en detalle los distintos valores semánticos asociados a cada una de las partículas interrogativas.

4.4.3.4.2.2.2 Valores semánticos asociados a cada partícula interrogativa[143]

En los capítulos dos y tres vimos que cada partícula interrogativa puede asociarse a una serie de valores semánticos.

El valor semántico general de cada interrogativo tiene un origen léxico y se corresponde con los valores que vimos en el capítulo 2. Estos valores semánticos generales pueden, a su vez, ser especificados a través de otras unidades con las que se combina el interrogativo: dentro de la frase interrogativa, sustantivos y preposiciones; a nivel oracional, el verbo. Por otra parte, las construcciones sintácticas específicas en las que el interrogativo se integra también pueden definir más explícitamente su valor semántico.

A continuación veremos como estos tres factores (valor léxico inherente, valor léxico de otras unidades, construcción sintáctica) interactúan para construir los distintos valores semánticos de las partículas interrogativas en SpQA. Como hemos hecho en otras secciones, iremos presentando los interrogativos con sus valores semánticos uno a uno.

a) *CUÁNDO*

Apuntábamos en el capítulo 2 que el valor semántico general asociado a *cuándo* es el de 'ubicación temporal'. Ese valor lo representamos en SpQA mediante el nodo TIME.

(258) # (null) 17 0-66|¿**Cuándo** ganó por primera vez las elecciones José María Aznar?
whQ **[V: ganar** third sing **past** <OBJ [N: elecciones <DET D: las] <SUBJ PN: José María Aznar **<QCIRC TIME** <CIRC X: por primera vez]

143 En el Apéndice 2 recogemos un compendio de todos los valores expuestos a continuación.

Para el valor temporal el tiempo del verbo es especialmente interesante, ya que sitúa la localización del evento en el pasado, en el presente o en el futuro. Como vemos en el ejemplo, en la representación de SpQA recogemos esa información.

Distinguir subtipos de localización temporal es, como decíamos en el capítulo 2, una tarea que requiere un procesamiento semántico que maneje conocimiento del mundo. Este tipo de procesamiento está más allá de las posibilidades de SpQA. Por esta razón, en la gramática se distingue simplemente un valor general de ubicación temporal.

Hemos visto también que para *cuándo* existe otro valor que no es de tipo temporal: el *cuándo* hipotético (259).

(259) *¿Cuándo un triángulo es equilátero?*

En el capítulo 2 observábamos también que distinguir el valor temporal del hipotético solo es posible mediante un análisis semántico profundo y manejando conocimiento del mundo. Por esta razón el valor hipotético tampoco se ha codificado en la gramática.

EJEMPLOS DE ANÁLISIS

- Circunstancial:

(260) `# (null) 5 0-43|¿Cuándo será el próximo eclipse de sol?`
`whQ [V: ser third sing future <SUBJ [N: eclipse <MODde N: sol <ATTR A: próximo <DET D: el] **<QCIRC TIME**]`

(261) `# (null) 19 0-55|¿**Desde cuándo** es presidente de España Mariano Rajoy?`
`whQ [V: ser third sing present <PRED [N: presidente <MODde PN: España] <SUBJ PN: Mariano Rajoy **<QCIRCdesde TIME**]`

b) *DÓNDE*

El valor semántico general de *dónde* es el de 'ubicación espacial'. Como vimos en el capítulo 3, el tipo de localización más común para *dónde* es 'localización física'.

Para *dónde* hemos codificado dos posibles valores semánticos:

- un valor general de 'localización': `LOCATION`.

(262) `# (null) 20 0-29|¿**Dónde** se produce la bilis?`
`whQ [V: producir third sing present <OBJ [N: bilis <DET D: la] <SUBJ IMP **<QCIRC LOCATION**]`

- un valor específico de 'localización geográfica': `LOCATION +GEO`.

(263) `# (null) 21 0-29|¿**Dónde** está Sierra Morena?`
`whQ [V: estar third sing present <SUBJ PN: Sierra Morena **<QCIRC LOCATION +GEO**]`

En la representación, LOCATION constituye el valor semántico central, especificado mediante +GEO.

El valor de localización geográfica se restringe a preguntas con *dónde* en las que la entidad por la que se pregunta es una entidad nombrada (NE) funcionando como sujeto. Esta restricción se sitúa en la línea de lo que vimos en el capítulo 3: la mayoría de los casos de localización geográfica implican la ubicación de una NE.

Sabemos, sin embargo, que también hay casos de localización geográfica en los que la entidad es un nombre común:

(264) *¿Dónde está* la sede de la Interpol*?*

Estos casos se analizan en SpQA como LOCATION, sin ninguna especificación del tipo concreto de localización física:

(264a) `# (null) 22 0-38|¿`**`Dónde`** `está la sede de la Interpol?`
`whQ [V: estar third sing present <SUBJ [N: sede <MODde [PN: Interpol <DET D: la] <DET D: la]` **`<QCIRC LOCATION`**`]`

Esta decisión de análisis se debe a que, con los datos que maneja la gramática, no es posible distinguir en las preguntas en las que la entidad es un nombre común aquellos casos que implican localización geográfica (265) de aquellos que implican otro tipo de localización física (266).

(265) ¿Dónde *está* la sede de la Interpol*?*
(266) ¿Dónde *se produce* la bilis*?*

De esta manera, LOCATION (*dónde* + nombre común) indica una localización general, que puede ser o no una localización geográfica (aunque, a la luz de los datos, en la mayoría de los casos será probablemente una localización física de tipo no geográfico, cf. capítulo 3, sección 3.3.3.3.2), mientras que LOCATION +GEO (*dónde* + NE) identifica un subtipo de localización, la geográfica.

Consideramos que de esta forma se aprovechan todas las posibilidades gramaticales y semánticas que puede procesar SpQA para precisar al máximo el significado concreto de la frase interrogativa con *dónde*.

EJEMPLOS DE ANÁLISIS

LOCATION

- Circunstancial:

(267) `# (null) 23 0-23|¿`**`Dónde`** `vive el koala?`
`whQ [V: vivir third sing present <SUBJ [N: koala <DET D: el]` **`<QCIRC LOCATION`**`]`

(268) # (null) 24 0-34|¿**En dónde** están las pirámides?
whQ [V: estar third plu present <SUBJ [N: pirámides <DET D: las] **<QCIRCen LOCATION**]

- Modificador:

(269) # (null) 25 0-37|¿**Dentro de dónde** vive el oso pardo?
whQ [V: vivir third sing present <SUBJ [N: oso <ATTR A: pardo <DET D: el] <QCIRC **[X: dentro <MODde LOCATION]**]

LOCATION +GEO

- Circunstancial:

(270) # (null) 26 0-28|¿**Dónde** está el Reichstag?
whQ [V: estar third sing present <SUBJ [PN: Reichstag <DET D: el] **<QCIRC LOCATION +GEO**]

(271) # (null) 27 0-25|¿**En dónde** está Moscú?
whQ [V: estar third sing present <SUBJ PN: Moscú **<QCIRCen LOCATION +GEO**]

c) *CÓMO*

Cómo presenta un valor general de 'modo, manera'. Sumando al valor semántico del interrogativo el valor del verbo, vimos en el capítulo 3 que podemos llegar a distinguir hasta tres subtipos semánticos (cf. capítulo 3, sección 3.3.3.3.6): 'denominación', 'procedimiento' y 'descripción'.

En la formalización hemos distinguido un valor general de 'modo, manera' y dos de estos tres valores específicos: 'denominación' y 'descripción'.

El valor semántico 'denominación' está formalizado como NAME. Este valor semántico se construye a partir de la información del interrogativo, el verbo y la construcción sintáctica. Se corresponde con la siguiente construcción:

Cómo + *VerbDenom* + entidad (sujeto u objeto directo)

(273) # (null) 29 0-44|¿**Cómo** se denomina científicamente la sal?
whQ [V: denominar third sing present <SUBJ [N: sal <DET D: la] **<QCIRC NAME** <CIRC X: científicamente]

El valor semántico 'descripción' está formalizado como DESCRIPTION. Como en el caso anterior, el valor semántico nace de la combinación del valor semántico del interrogativo con una construcción sintáctica concreta, en esta caso, una construcción copulativa:

Cómo + verbo *ser* + sujeto

(274) #(null) 31 0-22|¿**Cómo** es un zigurat?
whQ [V: ser third sing present **<QPRED DESCRIPTION** <SUBJ [N: zigurat <DET D: un]]

En cuanto al tercer valor que distinguíamos en el capítulo 3, 'procedimiento', la cuestión es más compleja. Como vimos allí, la cantidad de verbos que dan pie a este valor es enorme. Además, en bastantes casos la frontera entre 'procedimiento' y 'modo' es difícil de establecer solo con la información de interrogativo más verbo. Por estas razones, no incluimos este subtipo semántico en SpQA. Este valor lo englobamos en el tipo semántico más general MANNER, que abarca todos los casos con *cómo* que no se corresponden con alguna de las construcciones anteriores.

(275)# (null) 32 0-29|¿**Cómo** murió Jimmy Hendrix?
whQ [V: morir third sing past <SUBJ PN: Jimmy Hendrix **<QCIRC MANNER**]

(276)# (null) 33 0-28|¿**Cómo** se corta en juliana?
whQ [V: cortar third sing present <SUBJ IMP **<QCIRC MANNER** <CIRCen N: juliana]

Además de estos tres valores, *cómo* puede tener valor cuantitativo en dos construcciones distintas; en ambos casos el interrogativo se representa como QUANTITY:

- *a* + *cómo* + verbo + argumentos
  En este caso el valor QUANTITY se especifica como QUANTITY +MONEY:

(277) # (null) 34 0-44|¿**A cómo** vendió sus acciones el Santander?
whQ [V: vender third sing past <OBJ [N: acciones <DET D: sus] <SUBJ [PN: Santander <DET D: el] **<QCIRCa QUANTITY +MONEY**]

En la representación el valor semántico central o nuclear es el que va primero: QUANTITY. La información precedida de un "+" (MONEY) añade matices de significado (cantidad **de dinero**) a ese valor principal.

- *cómo* + frase preposicional[144] + verbo + complementos

(278) # (null) 35 0-34|¿**Cómo de alta** es la Torre Agbar?
whQ [V: ser third sing present **<QPRED [QUANTITY <MOD de A: alta]** <SUBJ [PN: Torre Agbar <DET D: la]]

144 Cuyo término es una frase adjetiva (*cómo de alta*) o adverbial (*cómo de lejos*).

La primera construcción es más bien propia del lenguaje oral, pero la hemos incluido porque su implementación era simple, no chocaba con otros análisis y, por lo tanto, no causaba problemas y añadía información a la gramática.

Vimos también (cf. capítulo 2) que *cómo* puede tener valor causal en dos construcciones:

- *cómo* + *no*:
  (279) *¿Cómo no aumenta el número de becas el gobierno?*

En SpQA no hemos dado cuenta de este valor por dos razones: la primera es que no se puede distinguir de cualquier otro de los valores presentados para *cómo* con negación interna. La segunda es que esta construcción con este valor no se documenta en nuestros corpus.

- *cómo* + apódosis de período condicional:
  (280) ¿Cómo *dices que te interesa si no le prestas la menor atención?*

Esta estructura no la hemos incluido porque nos parece más bien propia del lenguaje oral y su implementación era más costosa (en términos de formalización) que, por ejemplo, la de *a cómo* con valor cuantitativo.

EJEMPLOS DE ANÁLISIS

NAME

- Circunstancial.

(281) # (null) 36 0-60|¿**Cómo** se llamaba la carabela que divisó tierra americana?
whQ [V: llamar third sing imperfect <SUBJ [N: carabela >SUBJ [V: divisar third sing past <OBJ [N: tierra <ATTR A: americana]] <DET D: la] **<QCIRC NAME**]

MANNER

- Circunstancial.

(282)# (null) 37 0-38|¿**Cómo** se prepara una buena bechamel?
whQ [V: preparar third sing present <OBJ [N: bechamel <ATTR A: buena <DET D: una] <SUBJ IMP **<QCIRC MANNER**]

DESCRIPTION

- Predicativo.

(283)# (null) 38 0-36|¿**Cómo** es un triángulo isósceles?
whQ [V: ser third sing present **<QPRED DESCRIPTION** <SUBJ [N: triángulo <ATTR A: isósceles <DET D: un]]

QUANTITITY + MONEY

- Circunstancial.

(284) # (null) 39 0-40|¿**A cómo** están las acciones de Endesa?
whQ [V: estar third plu present <SUBJ [N: acciones <MODde PN: Endesa <DET D: las] **<QCIRCa QUANTITY +MONEY**]

QUANTITY

- Predicativo.

(285) # (null) 40 0-33|¿**Cómo de grande** es el Camp Nou?
whQ [V: ser third sing present **<QPRED [QUANTITY <MODde A: grande]** <SUBJ [PN: Camp Nou <DET D: el]]

d) *POR QUÉ*

El valor semántico general de *por qué* es el de causa. En SpQA le corresponde el nodo CAUSE:

(286) # (null) 41 0-37|¿**Por qué** subió el IVA el gobierno?
whQ [V: subir third sing past <OBJ [PN: IVA <DET D: el] <SUBJ [N: gobierno <DET D: el] **<QCIRC CAUSE**]

En el capítulo 2 vimos que además del valor causal, la construcción *por qué* + *no* podía presentar el valor de sugerencia. En SpQA hemos prescindido de este valor por dos razones:

- Primero, porque no hay forma de distinguir el *por qué no* de sugerencia (negación externa) del *por qué no* con negación interna.
  En el capítulo 3, cuando analizamos la negación en corpus, vimos además que el interrogativo con el que más se usa *no* es *por qué*, y que la negación en este caso es casi siempre interna. Por eso creemos que es preferible asociar en la gramática la construcción *por qué no* con una negación interna y no con una construcción de sugerencia.

- Segundo, porque las construcciones de sugerencia no parecen esperables en un perfil de BR, o al menos no en uno que no maneje la opinión.

Los casos de *por qué no* son, por tanto, analizados siempre como casos de negación interna.

(287) `# (null) 0 0-45|¿`**`Por qué no`** `baja el gobierno los impuestos?`
`whQ [V: bajar third sing present` **`<NEG X: no`** `<OBJ [A:impuestos <DET los] <SUBJ [N: gobierno <DET D: el]` **`<QCIRC CAUSE`**`]`

EJEMPLOS DE ANÁLISIS

`CAUSE`

- Circunstancial.

(288) `# (null) 43 0-47|¿`**`Por qué`** `gira la Luna alrededor de la Tierra?`
`whQ [V: girar third sing present <SUBJ [PN: Luna <DET D: la]` **`<QCIRC CAUSE`** `<CIRC [X: alrededor <MODde [PN: Tierra <DET D: la]]]`

(289) `# (null) 44 0-39|¿`**`Por qué no`** `sale Grecia de la crisis?`
`whQ [V: salir third sing present <NEG X: no <SUBJ PN: Grecia` **`<QCIRC CAUSE`** `<PCde [N: crisis <DET D: la]]`

e) *QUIÉN*

El valor semántico general de *quién* es el de 'persona'.

Como hemos visto (cf. capítulo 3, sección 3.3.3.3.3), dentro de este valor general los dos valores concretos más comunes son 'nombre propio' y 'descripción'. Ambos valores se asocian con patrones léxico sintácticos:

- Nombre propio: *quién* + verbo no copulativo + argumentos.
  (290) *¿Quién escribió El Decamerón?*

- Descripción: *quién* + verbo *ser* + entidad nombrada.
  (291) *¿Quién es el Capitán Nemo?*

En la gramática hemos implementado estos dos valores junto a otro más, también ligado a una construcción específica: *de* + *quién* + verbo *ser* + sujeto. Este tercer valor es el de 'propiedad':

(292) *¿De quién es La Bamba?*

Los valores semánticos implementados para *quién* son, por lo tanto, tres:

- para nombre propio: `NAME +PERSON`. Este valor se asocia al patrón sintáctico: *quién* + verbo no copulativo + argumentos. En la representación se añade la información relativa al afijo *NUMBER* (*sing*/*plu*) para el interrogativo.

(293) `# (null) 45 0-26|¿`**`Quién`** `dirigió Titanic?`
`whQ [V: dirigir third sing past <OBJ PN: Titanic` **`<QSUBJ NAME +PERSON sing`**`]`

- para descripción: `DESCRIPTION +PERSON`. Este valor se asocia al patrón sintáctico: *quién* + *ser* + entidad nombrada y funciona siempre como predicativo.

(294) `# (null) 46 0-26|¿`**`Quién`** `es J.J. Ramírez?`
`whQ [V: ser third sing present` **`<QPRED DESCRIPTION +PERSON`** `<SUBJ PN: J.J. Ramírez]`

- para propiedad: `NAME +PERSON +PROPERTY`. Este valor se asocia con la construcción: *de* + *quién* + *ser* + sujeto. Su función, por tanto, es siempre la de predicativo.

(295) `# (null) 47 0-24|¿`**`De quién`** `es La Bamba?`
`whQ [V: ser third sing present` **`<QPREDde NAME +PERSON +PROPERTY`** `<SUBJ PN: La Bamba]`

Como siempre que se especifica el valor semántico (cf. *supra*), en la representación el valor semántico central o nuclear es el que va primero: `NAME` o `DESCRIPTION`. La información precedida de un "+" añade matices de significado a ese valor principal: en un caso `PERSON` y en otro `PERSON` y `PROPERTY`.

Con `+PERSON` indicamos entonces que el valor `NAME` tiene el rasgo *+humano*. De esta manera la representación trata de recoger que con *quién* se pregunta generalmente por un valor asociado a un humano, que no tiene, sin embargo, que corresponderse siempre exactamente con un humano. Recordemos el ejemplo de Clef analizado en el capítulo 3 (cf. sección 3.3.3.3.3):

(296) *¿Quién fabricaba Windows 95?*
(296a) *Bill Gates / Microsoft.*

Lo mismo ocurre con `+PROPERTY`: con este matiz semántico indicamos que el nombre por el que se pregunta se corresponde con el de alguien o algo (asociado de alguna manera a una entidad humana, *vid.* ejemplo anterior) que es el propietario o autor de aquella entidad presente en la pregunta que funciona como sujeto.

EJEMPLOS DE ANÁLISIS

`NAME +PERSON`

- Sujeto:

(297) `# (null) 48 0-46|¿`**`Quién`** `ha ganado el último Premio Pulitzer?`
`whQ [V: ganar third sing present_perfect <OBJ [PN: Premio Pulitzer <MOD A: último <DET D: el]` **`<QSUBJ NAME +PERSON sing`**`]`

(298) # (null) 50 0-65|¿**Quién de los candidatos** ganó las últimas primarias del PSOE?
whQ [V: ganar third sing past <OBJ [N: primarias <ATTR A: últimas <MODde [PN: PSOE <DET D: el] <DET D: las] **<QSUBJ [NAME +PERSON sing <MODde [N: candidatos <DET D: los]]**]

- Objeto directo:

(299) # (null) 0 0-77|¿**A quién** se eligió ganador de la última edición de Operación Triunfo?
whQ [V: elegir third sing past <SUBJ IMP **<QOBJ NAME +PERSON sing** <PRED [N: ganador <MODde [N: edición <ATTR A: última <DET D: la] <MODde PN: Operación Triunfo]]

(300) # (null) 2 0-94|¿**A quién que nadie conocía** eligió Carlos Saura para protagonizar su famosa película?
whQ [V: elegir third sing past **<QOBJ [NAME +PERSON sing >OBJ [V: conocer third sing imperfect <SUBJ P: nadie]]** <SUBJ PN: Carlos Saura <CIRCpara [V: protagonizar <OBJ [N: película <ATTR A: famosa <DET D: su]]]

(301) # (null) 3 0-78|¿**Quién** fue elegido primer presidente de España tras la dictadura de Franco?
whQ [V: elegir third sing past <PRED [N: presidente <MODde PN: España <ATTR A: primer] **<QOBJ NAME +PERSON sing** <CIRCtras [N: dictadura <MODde PN: Franco <DET D: la]]

- Objeto indirecto:

(302) # (null) 4 0-64|¿**A quién** no le dio la mano Sarkozy en la última cumbre europea?
whQ [V: dar third sing past <IOBJ le third GENDER sing <NEG X: no **<QIOBJ NAME +PERSON sing** <OBJ [N: mano <DET D: la] <SUBJ PN: Sarkozy <CIRCen [N: cumbre <ATTR A: europea <ATTR A: última <DET D: la]]

(303) # (null) 1 0-65|**¿A quién de los empleados** le falló la empresa con sus medidas?
whQ [V: fallar third sing past <IOBJ le third GENDER sing **<QIOBJ [NAME +PERSON sing <MODde [N: empleados <DET D: los]]** <SUBJ [N: empresa <DET D: la] <CIRCcon [N: medidas <DET D: sus]]

- Circunstancial:

(304) # (null) 6 0-52|¿**Por quién** se evitó el golpe de estado de Tejero?
whQ [V: evitar third sing past <OBJ [N: golpe <MODde A:estado <MODde PN: Tejero <DET D: el] <SUBJ IMP **<QCIRCpor NAME +PERSON sing**]

- Complemento preposicional:

(305) # (null) 7 0-49|¿**De quién** dependen los presupuestos del estado?
whQ [V: depender third plu present **<QPCde NAME +PERSON sing** <SUBJ [N: presupuestos <MODde [N: estado <DET D: el] <DET D: los]]

- Modificador:

(306) # (null) 8 0-76|¿**Detrás de quién** se sienta el presidente en el Congreso de los diputados?
whQ [V: sentar third sing present <REF se third GENDER sing <SUBJ [N: presidente <DET D: el] **<QCIRC [X: detrás <MODde NAME +PERSON sing]** <CIRCen [N: congreso <MODde [N: diputados <DET D: los] <DET D: el]]

NAME +PERSON +PROPERTY

- Predicativo:

(307)# (null) 0 0-40|¿**De quién** es la canción *Wonderwall*?
whQ [V: ser third sing present **<QPREDde NAME +PERSON +PROPERTY** <SUBJ [N: canción <MOD PN: Wonderwall <DET D: la]]

(308) # (null) 1 0-69|¿**De quién de los futbolistas del Madrid** es la mansión más grande?
whQ [V: ser third sing present **<QPREDde [NAME +PERSON +PROPERTY <MODde [N: futbolistas <MODde [PN: Madrid <DET D: el] <DET D: los]]** <SUBJ [N: mansión <ATTR [A: grande <QUANT X:más] <DET D: la]]

DESCRIPTION +PERSON

- Predicativo.

(309) # (null) 2 0-27|¿**Quién** fue Benedicto XVI?
whQ [V: ser third sing past **<QPRED [DESCRIPTION +PERSON]** <SUBJ PN: Benedicto XVI]

f) *CUÁNTO*

Como hemos visto, desde el punto de vista semántico *cuánto* es uno de los interrogativos más simples ya que siempre apunta a una cantidad, tanto si funciona como núcleo sintáctico de la frase interrogativa como si no.

El valor general que le corresponde en SpQA es QUANTITY.

Cuando funciona como determinante, la representación de *cuánto* se aleja de lo puramente sintáctico para acercarse a lo semántico: se representa la noción de cantidad como núcleo de la construcción en virtud de su calidad de núcleo semántico:

(310) # (null) 3 0-40|¿**Con cuántas carabelas** partió Colón?
whQ [V: partir third sing past <SUBJ PN: Colón **<QCIRCcon [QUANTITY <MODde N: carabelas]**]

Este valor general QUANTITY se especifica semánticamente en tres casos:

- cuando tenemos un objeto directo de persona (con la preposición *a*): QUANTITY +PERSON.

(311) # (null) 4 0-41|¿**A cuántos actores** despidió Hitchcock?
whQ [V: despedir third sing past <**QOBJ [QUANTITY +PERSON** <MODde N: actores] <SUBJ PN: Hitchcock]

- con *cuánto* funcionando como pronombre más un verbo de tipo cuantitativo (*VerbQuant*), la entidad puede ser:

QUANTITY +MONEY cuando el verbo cuantifica dinero:

(312) # (null) 1 0-37|¿**Cuánto costó** el fichaje de Messi?
whQ [V: costar third sing past **<QOBJ QUANTITY +MONEY** <SUBJ [N: fichaje <MODde PN: Messi <DET D: el]]

QUANTITY +MEASURE cuando el verbo cuantifica una medida:

(313) # (null) 2 0-37|¿**Cuánto pesa** de media un cachalote?
whQ [V: pesar third sing present **<QOBJ QUANTITY +MEASURE** <SUBJ [N: cachalote <DET D: un] <CIRCde N: media]

EJEMPLOS DE ANÁLISIS

QUANTITY

- Sujeto.

(314) # (null) 3 0-72|¿**Cuántos futbolistas** siguen jugando con 30 años en la liga española?
whQ [[V: jugar third plu present <AUXdurat V: seguir] <CIRCcon [N: años <QUANT Q: 30] **<QSUBJ [QUANTITY <MODde N: futbolistas]** <CIRCen [N: liga <ATTR A: española <DET D: la]]

(315) # (null) 4 0-47|¿**Cuántos políticos de España** son madrileños?
whQ [V: ser third plu present <PRED A: madrileños **<QSUBJ [QUANTITY <MODde [N: políticos <MODde PN: España]]**]

(316) # (null) 5 0-40|¿**Cuánto que se sepa** robó Mario Conde?
whQ [V: robar third sing past **<QOBJ [QUANTITY >OBJ[V:saber third sing present]]** <SUBJ PN: Mario Conde]

- Objeto directo:

(317) # (null) 6 0-69|¿**Cuántos de los jugadores del Barcelona** fichó en Holanda Van Gaal?
whQ [V: fichar third sing past **<QOBJ [QUANTITY <MODde [N: jugadores <MODde [PN: Barcelona <DET D: el] <DET D: los]]** <CIRCen PN: Holanda <SUBJ PN: Van Gaal]

- Objeto indirecto:

(318) # (null) 7 0-39|¿**A cuántos** les debe dinero Bárcenas?
whQ [V: deber third sing present <IOBJ le third GENDER plu **<QIOBJ QUANTITY** <OBJ N: dinero <SUBJ PN: Bárcenas]

- Predicativo.

(319) # (null) 9 0-53|¿**De cuántos bits** era el primer procesador de Intel?
whQ [V: ser third sing imperfect **<QPRED [QUANTITY <MODde N: bits]** <SUBJ [N: procesador <MODde PN: Intel <ATTR A: primer <DET D: el]]

- Circunstancial.

(320) # (null) 10 0-45|¿**Cuánto** cuesta una barra de pan en Polonia?
whQ [V: costar third sing present **<QCIRC QUANTITY** <SUBJ [N: barra <MODde N: pan <DET D: una] <CIRCen PN: Polonia]

(321) # (null) 11 0-31|¿**Cada cuánto** se ve un cometa?
whQ [V: ver third sing present <OBJ [N: cometa <DET D: un] <SUBJ IMP **<QCIRCcada QUANTITY**]

(322) # (null) 13 0-54|¿**Con cuántos años** ganó su primer Óscar Al Pacino?
whQ [V: ganar third sing past <OBJ [PN: Óscar <MOD A: primer <DET D: su] <SUBJ PN: Al Pacino **<QCIRCcon [QUANTITY <MODde N: años]**]

- Complemento preposicional.

(323) # (null) 14 0-68|¿**De cuántos asuntos** trató la última reunión de Rajoy con Obama?
whQ [V: tratar third sing past **<QPCde [QUANTITY <MODde N: asuntos]** <SUBJ [N: reunión <MODde PN: Rajoy <ATTR A: última <DET D: la] <CIRCcon PN: Obama]

- Modificador.

(324) # (null) 15 0-71|¿**Dentro de cuánto tiempo** pasará cerca de la Tierra el cometa Halley?
whQ [V: pasar third sing future <SUBJ [N: cometa <MOD PN: Halley <DET D: el] <**QCIRC [X: dentro <MODde [QUANTITY <MODde N: tiempo]]** <CIRC [X: cerca <MODde [N: tierra <DET D: la]]]

QUANTITY +PERSON

- Objeto directo.

(325) # (null) 16 0-68|¿**A cuántos votantes** reunió Mas en su último mitin de campaña?
whQ [V: reunir third sing past **<QOBJ [QUANTITY <MODde N: votantes]** <SUBJ PN: Mas <CIRCen [N: mitin <MODde N: campaña <ATTR A: último <DET D: su]]

QUANTITY +MONEY

- Objeto directo.

(326) # (null) 17 0-22|¿**Cuánto** cobra Villa?
whQ [V: cobrar third sing present **<QOBJ QUANTITY +MONEY** <SUBJ PN: Villa]

```
QUANTITY +MEASURE
```

- Objeto directo.

(327) `# (null) 18 0-35|¿`**`Cuánto`** `mide un campo de fútbol?`
`whQ [V: medir third sing present` **`<QOBJ QUANTITY +MEASURE`** `<SUBJ [N: campo <MODde N: fútbol <DET D: un]]`

g) CUÁL

Hemos visto que *cuál* es el único interrogativo vacío de contenido semántico propio y que su significado se construye a través de la entidad a la que se apunta en la pregunta. Esa entidad no está restringida semánticamente y puede ser de cualquier tipo: animada, no animada, concreta, abstracta, etc.

También hemos visto que en un sistema de BR son esperables tres tipos de construcciones para las preguntas con *cuál*:

1) *Cuál* como determinante acompañando a un sustantivo: el sustantivo constituye el conjunto, la clase o tipo del referente al que apunta el interrogativo.
   (328) *¿Cuál* ciudad *tiene más población en el mundo?*
   (329) *¿Cuál* río *es el más caudaloso del mundo?*
   (330) *¿Dentro de cuál* museo *se guarda La Gioconda?*

2) *Cuál* como pronombre modificado por una frase preposicional con una frase nominal en su interior: la frase nominal determina el conjunto, la clase o tipo del referente al que apunta el interrogativo.
   (331) *¿Cuál de* los Beatles *tocaba la batería?*
   (332) *¿Cuál* de los países europeos *es el más afectado por la crisis?*

3) *Cuál* como pronombre en una estructura copulativa ecuativa en la que el referente al que apunta el interrogativo es directamente el constituyente de la ecuativa que funciona o bien como sujeto o bien como predicativo.

   (333) *¿Cuál fue* la causa del tsunami que asoló Malasia?

Ya hemos visto que *cuál* funcionando como determinante no aparece en la representación (lo mismo ocurre con *qué*; cf. *infra*):

(334)`# (null) 19 0-49|¿`**`Cuál ciudad`** `tiene más población en el mundo?`
`whQ [V: tener third sing present <OBJ [N: población <QUANT Q: más]` **`<QSUBJ N: ciudad`** `<CIRCen [N: mundo <DET D: el]]`

Esta decisión de representación obedece al hecho de que *cuál* funcionando como determinante no aporta ningún significado a la pregunta.

Cuando *cuál* funciona como núcleo de la frase interrogativa, tiene los siguientes valores en SpQA:

- Cuando apunta a una entidad genérica: ENTITY.

(335) # (null) 20 0-42|¿**Cuál de los Beatles** tocaba la batería?
whQ [V: tocar third sing imperfect <OBJ [N: batería <DET D: la] **<QSUBJ [ENTITY <MODde [PN: Beatles <DET D: los]]**]

Utilizamos ENTITY cuando no es posible determinar por la información gramatical más detalles sobre el carácter de la entidad por la que se pregunta. En los casos en los que hay una frase preposicional que especifica semánticamente la entidad (como en el ejemplo), la recogemos en la representación.

Este valor también funciona en las estructuras copulativas en las que se nos pide la identificación de la entidad presente en la pregunta:

(336) # (null) 21 0-43|¿**Cuál** es la profesión de Gianni Versace?
whQ [V: ser third sing present <PRED [N: profesión <MODde PN: Gianni Versace <DET D: la] **<QSUBJ ENTITY**]

(337) # (null) 22 0-46|¿**Cuáles** son los ingredientes de la sangría?
whQ [V: ser third plu present <PRED [N: ingredientes <MODde [N: sangría <DET D: la] <DET D: los] **<QSUBJ ENTITY**]

- En las preguntas copulativas en las que el sujeto es una entidad nombrada:
  (338) ¿*Cuál es* Barack Obama*?*

  el valor de la frase interrogativa es DESCRIPTION, ya que la identificación no consiste en aportar el nombre de la entidad (porque este ya nos es dado en la pregunta) sino una descripción de la misma:

(338a) # (null) 23 0-24|¿**Cuál** es Barack Obama?
whQ [V: ser third sing present **<QPRED DESCRIPTION** <SUBJ PN: Barack Obama]

- Cuando se pregunta por una entidad que se corresponde con una persona: PERSON. Es el caso de la frase interrogativa funcionando como objeto directo de persona, es decir: objeto directo con la preposición *a*:

(339) `# (null) 7 0-55|¿A cuál de los cantantes españoles imitaba Jesulín?`
`whQ [V: imitar third sing imperfect <QOBJ` **`[PERSON <MODde [N: españoles <ATTR A: cantantes <DET D: los]]`** `<SUBJ PN: Jesulín]`

- Cuando se pregunta por una entidad en la construcción:
  de + *cuál* + *ser* + sujeto

  la representación es `ENTITY +PROPERTY`.

(340) `# (null) 8 0-42|¿De cuál de los Rolling es Satisfaction?`
`whQ [V: ser third sing present` **`<QPREDde [ENTITY +PROPERTY <MODde [PN: Rolling <DET D: los]]`** `<SUBJ PN: Satisfaction]`

Como ocurre con otros interrogativos en esta construcción, la entidad por la que se pregunta es además la propietaria/autora del sujeto de la pregunta.

- Cuando se pregunta por el modo o manera de hacer algo (paráfrasis de *cómo*): `MANNER`.
  Este valor se ciñe a la siguiente construcción léxico sintáctica:
  de + *cuál* + *modalNoun*
  Como veremos, *qué* tiene el mismo valor en esta misma construcción (cf. *infra*).

(341) `# (null) 26 0-42|¿`**`De cuál forma`** `se prepara la salsa rosa?`
`whQ [V: preparar third sing present <OBJ [N: salsa <ATTR A: rosa <DET D: la] <SUBJ IMP` **`<QCIRC MANNER`**`]`

En esta construcción hemos optado por no representar ni el sustantivo al que determina *cuál* ni la preposición. Nuevamente hemos optado por una representación más semántica que sintáctica al considerar que toda la construcción *de* + *cuál* + *modalNoun* es equivalente al valor semántico *MANNER*. De este modo, la representación se equipara con la que tendríamos si la pregunta se construyera con *cómo*:

(342) `# (null) 27 0-33|¿`**`Cómo`** `se prepara la salsa rosa?`
`whQ [V: preparar third sing present <OBJ [N: salsa <ATTR A: rosa <DET D: la] <SUBJ IMP` **`<QCIRC MANNER`**`]`

- Cuando se pregunta por la causa de algo (paráfrasis de *por qué*): `CAUSE`.
  Este valor se ciñe a la siguiente construcción léxico sintáctica:
  por + *cuál* + *causeNoun*
  Como veremos, *qué* tiene el mismo valor en esta misma construcción (cf. *infra*).

(343) `# (null) 28 0-48|¿`**`Por cuál motivo`** `se retiró Esperanza Aguirre?`
`whQ [V: retirar third sing past <REF se third GENDER sing <SUBJ PN: Esperanza Aguirre` **`<QCIRC CAUSE`**`]`

Como en el caso anterior, no recogemos en la representación el valor de la preposición y el *causeNoun*. De esta forma, la representación es paralela a la de la misma pregunta con *por qué*:

(344) `# (null) 29 0-40|¿`**`Por qué`** `se retiró Esperanza Aguirre?`
`whQ [V: retirar third sing past <REF se third GENDER sing <SUBJ PN: Esperanza Aguirre` **`<QCIRC CAUSE`**`]`

- Cuando se pregunta por una ubicación física (paráfrasis de *dónde*): la representación en este caso es algo más compleja. La construcción que está debajo de este significado es siempre:
  preposición locativa + *cuál* + *locNoun*
  (345) ¿En cuál país *se produce más café?*

  Como veremos, *qué* tiene el mismo valor en esta misma construcción (cf. *infra*).

Vimos en el capítulo 2 (cf. sección 2.3.2.2) que hay varias preposiciones con valor locativo, en concreto: *ante, bajo, sobre, de, desde, a, hacia, hasta, para, por, en, entre, tras.* De esta lista, solo hemos implementado en esta construcción las siguientes preposiciones: *bajo, sobre, desde, hasta, hacia, en, entre.*

Pese a que *ante* tiene, aisladamente, un significado puramente locativo, la hemos descartado porque creemos que su uso está más ligado a construcciones que no son en realidad locativas, como:

(346) *¿Ante cual país cedió Dinamarca en las negociaciones de su tratado?*

Además, *ante* no está documentada en nuestros corpus.

Por otro lado, hemos descartado: *a*, *de*, *para, por* y *tras* porque en esta construcción pueden presentar un valor no locativo que nos parece muy probable:

(347) *¿A cuál país eligió?*
(348) *¿De cuál país depende?*
(349) *¿Para cuál país luchó Byron?*
(350) *¿Por cuál causa luchó el Che?*
(351) *¿Tras cuál país votó Francia en Eurovisión?*

Lo anterior también podría aplicarse para la preposición *en*, en ejemplos como:

(352) *¿En cuál país recayeron las miradas?*

Sin embargo, un análisis detallado de las construcciones *en* + *qué* + fn en el corpus Clef reveló que los valores más comunes eran el locativo y el temporal. Es por esta razón por la que, en el caso de la combinación *en* + *qué/cuál*[145] + fn locativa, consideramos como prioritario el valor locativo y no otros posibles como el de (352).

En cuanto a la representación, en este caso, a diferencia de los anteriores, tanto el valor de la preposición como el del sustantivo es relevante, ya que ambos aportan contenido semántico que precisa al de 'localización'. Nos interesa, por tanto, representar la función sintáctica, el valor de 'localización', la preposición y el sustantivo locativo. Para recoger toda esa información, hemos optado por la siguiente representación:

(353) `# (null) 30 0-39|¿`**`En cuál país`** `se produce más café?`
`whQ [V: producir third sing present <SUBJ IMP` **`<QCIRC_LOCATIONen N: país`** `<OBJ [N: café <QUANT Q: más]]`

Como se ve en el ejemplo, hemos optado por incorporar la información semántica al identificador de función sintáctica, separando ambas informaciones con un guión.

Hemos utilizado esta misma estructura para el siguiente valor:

- Cuando se pregunta por una ubicación temporal (paráfrasis de *cuándo*): `CIRC_TIMEprep.` La construcción que está debajo de este significado es siempre:
  preposición con valor temporal + *cuál* + *tempNoun*

(354) `# (null) 31 0-58|¿`**`En cuál año`** `ganó Bush las elecciones por primera vez?`
`whQ [V: ganar third sing past <OBJ [N: elecciones <DET D: las] <SUBJ PN: Bush` **`<QCIRC_TIMEen N: año`** `<CIRC X: por primera vez]`

En este caso, hemos tenido en cuenta todas las preposiciones con valor temporal (cf. cap. 2, sección 2.3.2.2): *a, de, desde, durante, en, entre, hacia, hasta, para, por, tras*.

Como veremos, *qué* tiene el mismo valor en esta misma construcción (cf. *infra*).

145 Aunque los datos de Clef corresponden a *qué* y no a *cuál*, consideramos estos dos interrogativos como equivalentes en este tipo de construcciones.

EJEMPLOS DE ANÁLISIS

ENTITY

- Sujeto.

(355) # (null) 32 0-33|¿**Cuál avión** es el más grande?
whQ [V: ser third sing present <PRED [A: grande <QUANT X:más <DET D: el] **<QSUBJ N: avión**]

(356) # (null) 33 0-70|¿**Cuál de los participantes de Master Chef** ganó el concurso de 2013?
whQ [V: ganar third sing past <OBJ [N: concurso <DATEde 2013 <DET D: el] **<QSUBJ [ENTITY <MODde [N: participantes <MODde PN: Master Chef** <DET D: los]]]

- Objeto directo.

(357) # (null) 34 0-47|¿**Cuál modelo famoso de avión** fabrica Boeing?
whQ [V: fabricar third sing present **<QOBJ [N: modelo <ATTR A: famoso <MODde N: avión]** <SUBJ PN: Boeing]

(358) # (null) 35 0-52|¿**Cuál de los coches de F1** conduce Fernando Alonso?
whQ [V: conducir third sing present <**QOBJ [ENTITY <MODde [N: coches <MODde PN: F1 <DET D: los]]** <SUBJ PN: Fernando Alonso]

- Objeto indirecto.

(359) # (null) 36 0-47|¿**A cuál de los bancos** le debe más dinero IU?
whQ [V: deber third sing present <IOBJ le third GENDER sing **<QIOBJ [ENTITY <MODde [N: bancos <DET D: los]]** <OBJ [N: dinero <QUANT Q: más] <SUBJ PN: IU]

- Complemento preposicional.

(360) # (null) 37 0-70|¿**En cuál de los Rolling Stones** se convirtió Michael Phillip Jagger?
whQ [V: convertir third sing past <REF se third GENDER sing **<QPCen [ENTITY <MODde [PN: Rolling Stones <DET D: los]]** <SUBJ PN: Michael Phillip Jagger]

- Circunstancial.

(361) # (null) 38 0-35|¿**De cuál país** viene el aguacate?
whQ [V: venir third sing present <SUBJ [N: aguacate <DET D: el] **<QCIRCde N: país**]

- Modificador.

(362) # (null) 39 0-48|¿**Después de cuál período** viene el Cretácico?
whQ [V: venir third sing present <SUBJ [N: cretácico <DET D: el] <**QCIRC [X: después <MODde N: período]**]

DESCRIPTION (copulativas)

- Predicativo.

(363) # (null) 40 0-55| ¿**Cuál de los candidatos demócratas** es Barack Obama?
whQ [V: ser third sing present **<QPRED [DESCRIPTION <MODde [N: candidatos <ATTR A: demócratas <DET D: los]]** <SUBJ PN: Barack Obama]

ENTITY +PROPERTY (copulativas)

- Predicativo.

(364) # (null) 41 0-39|¿**De cuál de los Beatles** es Let It Be?
whQ [V: ser third sing present **<QPREDde [ENTITY +PROPERTY <MODde [PN: Beatles <DET D: los]]** <SUBJ PN: Let It Be]

(365) # (null) 42 0-44|¿**De cuál escritor español** es La Colmena?
whQ [V: ser third sing present **<QPREDde [N: escritor <ATTR A: español]** <SUBJ PN: La Colmena]

MANNER

- Circunstancial.

(366) # (null) 43 0-42|¿**De cuál manera** se hace un nudo Windsor?
whQ [V: hacer third sing present <REF se third GENDER sing <SUBJ [N: undo <MOD PN: Windsor <DET D: un] **<QCIRC MANNER**]

CAUSE

- Circunstancial.

(367) # (null) 44 0-37|¿**Por cuál razón** flotan los barcos?
whQ [V: flotar third plu present <SUBJ [N: barcos <DET D: los] **<QCIRC CAUSE**]

CIRC_LOCATION

- Circunstancial.

(368) # (null) 45 0-36| ¿**Desde cuál lugar** partió Colón?
whQ [V: partir third sing past <SUBJ PN: Colón **<QCIRC_LOCATIONdesde N: lugar**]

CIRC_TIME

- Circunstancial.

(369) # (null) 3 0-46|¿**En cuál año** fue la Batalla de Aljubarrota?
whQ [V: ser third sing past <SUBJ [PN: Batalla de Aljubarrota <DET D: la] **<QCIRC_TIMEen N: año**]

h) *QUÉ*

Hemos visto que *qué* tiene también varios valores semánticos: 'entidad', 'tipo o clase', 'cantidad' y 'definición' (en estructuras copulativas). Ya hemos explicado que en general los valores de 'entidad' y 'tipo o clase' no se pueden diferenciar.

En la formalización distinguimos en primer lugar entre los casos en los que *qué* funciona como determinante de aquellos en los que funciona como pronombre.

Si funciona como determinante el sustantivo al que determina constituye el núcleo semántico de la frase interrogativa (y, por lo tanto, de la incógnita). Como con *cuál*, en estos casos solo se muestra el sustantivo en la representación:

(370) # (null) 47 0-43|¿**Qué jugador del Madrid** marca más goles?
whQ [V: marcar third sing present <OBJ [N: goles <QUANT Q: más] <QSUBJ **[N: jugador <MODde [PN: Madrid <DET D: el]]**]

Esto es así excepto en dos construcciones concretas con un valor semántico fijo que ya hemos tratado al analizar *cuál*:

- *de* + *qué* + *modalNoun*: el sustantivo y la preposición se omiten y se muestra el valor MANNER.

(371) # (null) 48 0-51| ¿**De qué forma** quiere la UE acabar con la crisis?
whQ [V: querer third sing present <OBJ [V:acabar <CIRCcon [N: crisis <DET D: la]] <SUBJ [PN: UE <DET D: la] **<QCIRC MANNER**]

- *por* + *cuál* + *causeNoun*: el sustantivo y la preposición también se omiten y se muestra el valor CAUSE.

(372) # (null) 50 0-54|¿**Por qué motivo** ha subido los impuestos el gobierno?
whQ [V: subir third sing present_perfect <OBJ [A:impuestos <DET los] <SUBJ [N: gobierno <DET D: el] **<QCIRC CAUSE**]

Se distinguen dos valores semánticos más para *qué* funcionando como determinante:

- con la construcción:
  preposición locativa (*bajo, sobre, desde, hasta, hacia, en, entre*)[146] + *cuál* + *locNoun*
  el valor es CIRC_LOCATIONprep.

(373) # (null) 51 0-31|¿**En qué nación** nació Lenin?
whQ [V: nacer third sing past <SUBJ PN: Lenin **<QCIRC_LOCATIONen N: nación**]

- con las construcciones:
  preposición temporal (*a, de, desde, durante, en, entre, hacia, hasta, para, por, tras*) + *qué* + *tempNoun*
  y
  qué + *año*
  el valor semántico es QCIRC_TIME, con o sin preposición:

(374) # (null) 52 0-51|¿**En qué año** finalizó la Guerra Civil española?
whQ [V: finalizar third sing past <SUBJ [PN: Guerra Civil <ATTR A: española <DET D: la] **<QCIRC_TIMEen N: año**]

(375) # (null) 53 0-46|¿**Qué año** se convirtió Alaska en un estado?
whQ [V: convertir third sing past <REF se third GENDER sing <SUBJ PN: Alaska **<QCIRC_TIME N: año** <PCen [N: estado <DET D: un]]

146 Cf. *supra* (p. 87) para los criterios de selección de las preposiciones locativas.

Cuando funciona como pronombre *qué* puede tener tres valores en SpQA:

- Cuando no le sigue un verbo que implique cuantificación ni forma parte de cierta construcción copulativa (cf. *infra*): ENTITY.

(376) # (null) 54 0-40|¿**Qué** buscaban Jasón y los Argonautas?
whQ [V: buscar third plu imperfect **<QOBJ ENTITY** <SUBJ PN: Jasón | [N: argonautas <DET D: los]]

El valor ENTITY también se aplica a la siguiente construcción copulativa en la que la frase interrogativa funciona como predicativo: *de* + *qué* + *ser* + sujeto.

(377)# (null) 55 0-35|¿**De qué** es el helado de nocciola?
whQ [V: ser third sing present **<QPREDde ENTITY** <SUBJ [N: helado <MODde UNKNOWN N: nocciola? <DET D: el]]

- Cuando al pronombre le sigue un verbo que implique cuantificación en la construcción:
  qué + *VerbQuant* + entidad
  establecemos un valor general cuantitativo, QUANTITY más una especificación de qué tipo de valor se está cuantificando. En cuanto a este valor, para *qué*, a diferencia de para *cuánto*, solo contemplamos la posibilidad de +MONEY:

(378)# (null) 56 0-34|¿**Qué** costó el Túnel del Canal?
whQ [V: costar third sing past **<QOBJ QUANTITY +MONEY** <SUBJ [PN: Túnel del Canal <DET D: el]]

En cuanto a la otra especificación posible con *VerbQuant*, +MEASURE, en el caso de *qué* la desechamos porque existe ambigüedad en casos como:

(379) ¿Qué *mide la escala de Mohs?*

En (379) se pregunta por qué tipo de entidad es la que se mide con la escala de Mohs y no por una magnitud como ocurre, por ejemplo, en (380):

(380) ¿Qué *mide el Everest?*

Como (379) y (380) son estructuralmente iguales, la única forma de distinguirlas es manejando conocimiento del mundo. Por esta razón, en SpQA analizamos estos casos con el valor general ENTITY, que puede ser tanto una entidad (379) como una cantidad (380).

(381) # (null) 57 0-30|¿**Qué** mide la escala de Mohs?
whQ [V: medir third sing present **<QOBJ ENTITY** <SUBJ [N: escala <MODde PN: Mohs <DET D: la]]

(382) # (null) 58 0-23|¿**Qué** mide el Everest?
whQ [V: medir third sing present **<QOBJ ENTITY** <SUBJ [PN: Everest <DET D: el]]

- en la estructura copulativa:
  qué + *ser* + entidad nombrada funcionando como sujeto
  el valor es DEFINITION.

(383)# (null) 59 0-25|¿**Qué** es la Mamba Negra?
whQ [V: ser third sing present **<QPRED DEFINITION** <SUBJ [PN: Mamba Negra <DET D: la]]

EJEMPLOS DE ANÁLISIS

ENTITY (entidad, cantidad, tipo o clase)

- Sujeto.

(384) # (null) 60 0-26|¿**Qué** pone huevos azules?
whQ [V: poner third sing present <OBJ [N: huevos <ATTR A: azules] **<QSUBJ ENTITY**]

(385) # (null) 9 0-49|¿Qué de todo el escándalo Bárcenas es ilegal?
whQ [V: ser third sing present <PRED A: ilegal **<QSUBJ [ENTITY <MODde [N: escándalo <MOD PN: Bárcenas <DET D: todo | D: el]]**]

- Objeto directo.

(386) # (null) 0 0-38|¿**Qué** inventó Alexander Graham Bell?
whQ [V: inventar third sing past **<QOBJ ENTITY** <SUBJ PN: Alexander Graham Bell]

QUANTITY +MONEY

- Objeto directo.

(387) # (null) 1 0-31|¿**Qué** costará el nuevo museo?
whQ [V: costar third sing future **<QOBJ QUANTITY +MONEY** <SUBJ [N: museo <ATTR A: nuevo <DET D: el]]

DEFINITION

- Predicativo.

(388) # (null) 2 0-17|¿**Qué** es la ONU?
whQ [V: ser third sing present **<QPRED DEFINITION** <SUBJ [PN: ONU <DET D: la]]

CIRC_TIME

- Circunstancial.

(389) # (null) 3 0-52|¿**En qué día** se inició la Segunda Guerra Mundial?
whQ [V: iniciar third sing past <REF se third GENDER sing <SUBJ [PN: Guerra Mundial <MOD A: segunda <DET D: la] **<QCIRC_TIMEen N: día**]

(390) # (null) 4 0-50|¿**Qué día** se terminó la Primera Guerra Mundial?
whQ [V: terminar third sing past <REF se third GENDER sing <SUBJ [PN: Guerra Mundial <MOD A: primera <DET D: la] **<QCIRC_TIME N: día**]

(391) # (null) 5 0-68|¿**Qué año** ganó por primera vez las elecciones José María Aznar?
whQ [V: ganar third sing past <OBJ [N: elecciones <DET D: las] <SUBJ PN: José María Aznar **<QCIRC_TIME N: año** <CIRC X: por primera vez]

(392) # (null) 6 0-32|¿**Qué mes** nació Justin Bieber?
whQ [V: nacer third sing past <SUBJ PN: Justin Bieber **<QCIRC_TIME N: mes**]

CIRC_LOCATION

- Circunstancial.

(393) # (null) 7 0-38|¿**Hacia qué lugar** partió Marco Polo?
whQ [V: partir third sing past <SUBJ PN: Marco Polo **<QCIRC_LOCATIONhacia N: lugar**]

MANNER

- Circunstancial.

(394) # (null) 8 0-36|¿**De qué manera** se suicidó Cobain?
whQ [V: suicidar third sing past <REF se third GENDER sing <SUBJ PN: Cobain **<QCIRC MANNER**]

CAUSE

- Circunstancial.

(395) # (null) 9 0-34|¿**Por qué causa** el cielo es azul?
whQ [V: ser third sing present <SUBJ [N: cielo <DET D: el] <PRED A: azul **<QCIRC CAUSE**]

4.4.3.4.3 Las preguntas con disyunción

Como hemos visto, esta categoría incluye en SpQA las disyuntivas propiamente dichas y las preguntas parciales con disyunción.

En el capítulo 2 (cf. sección 2.2.4.1.3) se explicó que en algunos casos no es posible distinguir (por escrito) entre una pregunta total con disyunción y una disyuntiva:

(396) *¿Bebió agua o vino?*
(396a) TOTAL: Bebió agua o vino + incógnita [¿Sí? / ¿No?]
(396b) DISYUNTIVA: Bebió + incógnita [¿Agua? / ¿Vino?]

Para poder diferenciar ambas interpretaciones es necesario manejar datos relativos a la entonación o acceder directamente al hablante mediante interacción. Por esta razón, siempre que hay disyunción en una estructura con forma de total, en SpQA esta se analiza como una disyuntiva.

En el caso de preguntas con disyunción la diferencia entre totales y parciales se diluye un poco. La diferencia semántica entre ambos tipos de preguntas es que, en el caso de las parciales con disyunción, se especifica el tipo semántico de los elementos incluidos en la disyunción mediante la información contenida en la frase interrogativa (*qué* = 'entidad no animada'). Por lo demás, ambos tipos de preguntas son equivalentes semánticamente:

(397a) *¿Qué compró: agua o vino?* > *compró* + incógnita = *qué* = *agua/vino*
(397b) *¿Compró agua o vino?* > *compró* + incógnita = *agua/vino*

Por esta razón, en SpQA utilizamos la misma representación para las dos estructuras.

4.4.3.4.3.1 Representación

Como hemos adelantado, en la representación de SpQA disyuntivas y parciales con disyunción se analizan de la misma manera. En dicha representación, ambas estructuras se identifican como `disjQ`. Por otra parte, el ámbito de la variable interrogativa afecta a la estructura disyuntiva y se marca, como con las parciales, con una `Q` añadida al indicador de función sintáctica:

(398) `# (null) 1 0-47|¿Qué inventó Edison, la bombilla o la radio?`
`disjQ [V: inventar third sing past` **`<QOBJ [DISJUNCTION <DISJ [N: bombilla <DET D: la] <DISJ [N: radio <DET D: la]]`** `<SUBJ PN: Edison]`

(399) `# (null) 2 0-41|¿Edison inventó la bombilla o la radio?`
`disjQ [V: inventar third sing past` **`<QOBJ [DISJUNCTION <DISJ [N: bombilla <DET D: la] <DISJ [N: radio <DET D: la]]`** `<SUBJ PN: Edison]`

Como se puede ver, en (398) y (399) el nodo del constituyente tiene el valor de `disjunction`; ese nodo, a su vez, tiene como valores posibles los distintos miembros de la disyunción:

```
[DISJUNCTION,DISJ,N: bombilla]
[DISJUNCTION,DISJ,N: radio]
[N: bombilla, DET, D: la]
[N: radio, DET, D: la]
[V: inventar third sing past, QOBJ, DISJUNCTION]
[V: inventar third sing past, SUBJ, PN: Edison]
```

Contrariamente a lo que hacemos en el caso de las parciales, para las parciales con disyunción no aprovechamos el valor semántico especificado en el interrogativo cuando este constituye el único elemento de la frase interrogativa. Esta decisión se debe a que, en este caso, el valor impreciso del interrogativo se concreta a través de la disyunción, por lo que no vemos útil la representación semántica del interrogativo para el sistema de BR.

Cuando el interrogativo funciona como determinante, la representación es la siguiente:

(400) # (null) 0 0-70|¿**Qué animal** tiene más músculos: **la oruga, el hombre o el elefante**?
disjQ [V: tener third sing present <OBJ [N: músculos <QUANT Q: más] <QSUBJ **[DISJUNCTION <DISJ [N: oruga <DET D: la] <DISJ [N: hombre <DET D: el] <DISJ [N: elefante <DET D: el]]**]

TRIPLETES
[DISJUNCTION, DISJ, N: elefante]
[DISJUNCTION, DISJ, N: hombre]
[DISJUNCTION, DISJ, N: oruga]
[N: elefante, DET, D: el]
[N: hombre, DET, D: el]
[N: músculos, QUANT, Q: más]
[N: oruga, DET, D: la]
[V: tener third sing present, OBJ, N: músculos]
[V: tener third sing present, QSUBJ, DISJUNCTION]

Como se ve en el ejemplo, en este caso la representación no mantiene el sustantivo de la frase interrogativa, *animal*. En su lugar, la *head* del nodo interrogativo es DISJUNCTION. Nos hemos decantado por esta interpretación porque, por razones técnicas relativas a la arquitectura de la *transduction* en la gramática, si elegíamos el sustantivo como head, habría que establecer algún tipo de relación sintáctica entre este y la estructura disyuntiva. Teniendo en cuenta que la incógnita se reduce a uno de los elementos de la disyuntiva y que el sustantivo solo aporta información relativa al tipo o clase de esos elementos, consideramos que no vale la pena introducir una nueva relación sintáctica para la disyunción (con su consiguiente triplete de dependencias) en la representación[147].

4.5 Conclusiones generales del capítulo

En el presente capítulo se han descrito el sistema de representación y la gramática formal de SpQA.

El sistema de representación utilizado en SpQA consiste en un grafo dependencial que condensa, de forma simple y compacta, información relativa a distintos niveles lingüísticos: léxico, sintáctico y semántico. El grafo permite la extracción de tripletes de dependencias y ha sido utilizado en otras gramáticas escritas en AGFL y diseñadas para la Recuperación de Información.

La gramática formal sobre la que se construye SpQA está escrita en el formaliso AGFL. Al estar diseñada para el análisis de preguntas, los módulos generales relativos a los tipos de frases o cláusulas no son tan completos como lo serían en una gramática formal de propósito general. La gramática tiene un diseño modular, donde una serie de módulos se encargan de las clases de palabras, otra serie de módulos de los tipos de frases y un tercer grupo de módulos de los tipos de cláusulas. La principal aportación de

147 En caso de cambiar de idea, el cambio en la *transduction* sería simple.

este trabajo es el módulo de las oraciones interrogativas. En dicho módulo se describen tres tipos de preguntas, cada una con una representación diferente en el grafo dependencial: preguntas totales (*ynQ*), preguntas parciales (*whQ*) y preguntas disyuntivas (*disjQ*). En el módulo de las disyuntivas, disyuntivas y parciales con disyunción reciben la misma representación. Las reglas que se encargan de la descripción de cada tipo de pregunta recogen la información teórica (capítulo 2) y de corpus (capítulo 3) susceptible de formalización. Para las preguntas parciales, se desarrollan una serie de valores semánticos para la incógnita de la frase interrogativa construidos a partir de información léxico sintáctica. Estos valores semánticos permiten lidiar con algunos casos de paráfrasis en la frase interrogativa.

A partir de esta gramática formal, el grafo dependencial que genera SpQA recoge la siguiente información relativa a tres niveles:

- **Léxico:** etiquetado de la clase de palabra para cada uno de los nodos del grafo.
- **Sintáctico:** identificación y etiquetado de las dependencias sintácticas de la pregunta, a nivel de la frase y de la oración, posibilidad de identificación y etiquetado de constituyentes oracionales y normalizaciones sintácticas.
- **Semántico:** identificación de la variable interrogativa, asignación de un valor semántico a la frase interrogativa en las parciales con tratamiento limitado de la paráfrasis e identificación de entidades nombradas y estructuras temporales.

Capítulo 5

Evaluación de SpQA

En las páginas que siguen presentaremos la evaluación de SpQA.

En primer lugar, describiremos el método de evaluación y los objetivos que esta persigue. A continuación, mostraremos los resultados obtenidos por el sistema en dicho marco de evaluación. Finalmente presentaremos un análisis de los principales errores del sistema y las conclusiones generales de la evaluación.

5.1 Método de evaluación

El método escogido se propone evaluar la eficacia de SpQA en el tratamiento de aquellos parámetros considerados como objetivos de nuestro sistema en el análisis de preguntas:

- Reconocimiento de una estructura lingüística como pregunta (punto de partida del análisis).

- Análisis de la variable interrogativa contenida en la pregunta: determinación del tipo de variable (pregunta total, parcial o disyuntiva, cf. capítulo 2, sección 2.2.1) y de su valor semántico.

- Análisis sintáctico de la pregunta: identificación de dependencias y de sus funciones (e, indirectamente, de los constituyentes oracionales y sus funciones).

- Reconocimiento específico de entidades nombradas (NE) y fechas.

La evaluación es de tipo intrínseco, pues la eficacia del sistema se determina respecto a un *gold standard* preestablecido y construido de acuerdo a qué es correcto en el modelo de SpQA (qué funciones sintácticas se contemplan, qué valores semánticos, etc.).

La elección de una evaluación de tipo intrínseco responde al hecho de que nos interesa medir el comportamiento del *parser* respecto a los parámetros anteriores, para los cuales ha sido diseñado.

Somos conscientes de que además de esta evaluación intrínseca lo deseable sería, dado que SpQA es un *parser* orientado a BR, una evaluación extrínseca en la que se midiera la eficacia del analizador integrado en uno de estos sistemas. Lamentablemente, este tipo de evaluación está más allá de nuestras posibilidades técnicas en el momento presente. Por esta razón, este tipo de evaluación constituye el siguiente objetivo que nos planteamos para SpQA en un futuro (cf. conclusiones finales). Por lo tanto, la evaluación extrínseca quedará fuera de este trabajo.

5.1.1 Variables evaluadas

Las variables concretas que hemos tenido en cuenta en la evaluación de SpQA responden a los parámetros arriba señalados y son las siguientes:

1) Reconocimiento de estructuras lingüísticas como preguntas por parte del sistema: se evalúa la efectividad del *parser* reconociendo las estructuras de un corpus como preguntas.

2) Reconocimiento del tipo de variable interrogativa presente en la pregunta: se evalúa la capacidad del *parser* para reconocer una pregunta como parcial (`whQ`), total (`ynQ`) o disyuntiva (`disjQ`) y, por lo tanto, para determinar la variable interrogativa de la pregunta (toda la oración en las totales; la frase interrogativa en las parciales; la disyunción en las disyuntivas).

3) Análisis sintáctico global de la pregunta y análisis sintáctico semántico del constituyente interrogativo: se evalúa el análisis sintáctico y semántico de las preguntas respecto a los siguientes parámetros:

3.1) Nivel de constituyentes oracionales[148]:

- **Límite de constituyente** (LC): se mide la eficacia del *parser* en el reconocimiento de los límites de cada uno de los constituyentes sintácticos presentes en la oración. Se tienen en cuenta solo los constituyentes a nivel oracional.

- **Núcleo del constituyente** (NC): se mide la eficacia del *parser* en el reconocimiento del núcleo sintáctico de cada uno de los constituyentes oracionales de la pregunta.
  Consideramos que esta variable puede tener sentido en aquellos casos en los que la LC no es correcta. A través de ella intentamos determinar los casos en los que el *parser* reconoce al menos la parte más importante (desde el punto de vista sintáctico) de un constituyente.

148 Como hemos visto, los límites de los constituyentes oracionales se marcan en el grafo de dependencias mediatne corchetes (cf. capítulo 4, sección 4.3).

- **Función sintáctica del constituyente completo** (FCC): evalúa la eficacia del *parser* en la asignación de una función sintáctica a un constituyente cuyos límites están correctamente reconocidos (LC correcto). Las funciones sintácticas (oracionales) tenidas en cuenta son aquellas contempladas en SpQA: SUBJ, OBJ, IOBJ, PRED, PC y CIRC.

- **Función sintáctica respecto al núcleo del constituyente** (FNC): evalúa la eficacia del *parser* en la asignación de una función sintáctica a un constituyente cuyo núcleo está correctamente asignado.

3.2) Nivel de dependencias sintácticas:

- **Reconocimiento de dependencias** (RD): se mide la eficacia del *parser* en el reconocimiento de las dependencias contempladas en SpQA. Por lo tanto, se evalúan tanto las relaciones a nivel oracional como de la frase.
  Computan como errores tanto los casos en los que se reconoce mal el nodo de la dependencia[149] como las dependencias mal asignadas[150].

- **Función sintáctica de la dependencia** (FD): se evalúa la eficacia del *parser* en la asignación de funciones para cada una de las dependencias de la oración (tanto a nivel oracional como de la frase).
  Se tienen en cuenta aquellas dependencias cuyo nodo ha sido correctamente delimitado (cf. nota 140), aunque la *head* de la dependencia no esté correctamente asignada[151] (cf. infra, ejemplo (5)).

3.3) Nivel semántico:

- **Valor semántico del constituyente interrogativo** (SEM): en las parciales, se evalúa la eficiencia del *parser* en la asignación de un valor semántico de entre los contemplados en SpQA (cf. capítulo 4, sección 4.4.3.4.2.2.2) al constituyente interrogativo.

- **Reconocimiento específico de entidades nombradas** (nodo `PN`) **y de fechas** (relator `DATE`): se evalúa la eficacia del *parser* en el reconocimiento de estos dos tipos de estructuras.

149 Por ejemplo, en las entidades nombradas como *Día de la Bestia* (en la frase nominal *la película el Día de la Bestia*), que no se reconocen como unidad.

150 *Head* mal asignada.

151 De esta manera, medimos exclusivamente que la función que le corresponde al nodo se asigne correctamente. De esta manera, no computan como errores casos como los que veremos en los que un modificador en una frase nominal compleja es asignado a la *head* incorrecta (cf. ejemplo (5)).

5.1.2 La evaluación de *parsers*

En los últimos años, el desarrollo y aplicación de diferentes modelos de *parsing*[152] ha ocupado un lugar muy importante en el campo del Procesamiento del Lenguaje Natural. Este hecho ha provocado a su vez el interés por el desarrollo de diversos métodos de evaluación de *parsers* (Black et al., 1991; Lin, 1995; 1998; Carroll et al., 1996; Carroll, Briscoe, y Sanfilippo, 1998; Carroll, Minnen, y Briscoe, 1999). En este contexto, se ha revelado como especialmente compleja la comparación de herramientas pertenecientes a diferentes modelos teóricos (Musillo, y Sima'an, 2002; Gaizauskas, y Wilks, 1998).

Entre los diferentes métodos de evaluación de analizadores sintácticos, las medidas del esquema *PARSEVAL* (Black et al., 1991) se han convertido en las más utilizadas (Gaizauskas, y Wilks, 1998, p. 2). El esquema *PARSEVAL* utiliza un corpus anotado[153] como *gold standard*. De esta manera, compara un análisis candidato (salida del *parser*) con un análisis de referencia en el *gold standard* y ofrece valores para tres medidas: *crossing brackets*, *precision* y *recall*. La primera medida cuenta el número de constituyentes correctos e incorrectos (en términos de límites de constituyente) que se producen. *Precision* es la proporción (o porcentaje) de constituyentes en la salida que aparecen en el *gold standard* (número de constituyentes correctos dividido por número de constituyentes en la salida del *parser*). *Recall* es la proporción de constituyentes en el *gold standard* que aparecen en la salida (número de constituyentes correctos en la salida del *parser* dividido por total de constituyentes correctos en el *gold standard*). En su versión más simple, estas medidas se aplican solo al reconocimiento de constituyentes; en su versión más estricta, las medidas se aplican también a la asignación de funciones a cada constituyente.

En la evaluación de SpQA aplicamos el esquema de *PARSEVAL* a las siguientes variables:

- Reconocimiento de preguntas.
- Reconocimiento de la variable interrogativa (tipo de pregunta).
- Análisis sintáctico global y análisis sintáctico semántico de la frase interrogativa.

Además de la *precision* y el *recall* entendidos como acabamos de explicar, se mide también el *F-score*[154].

Al reconocimiento específico de entidades y fechas aplicamos un esquema de evaluación más simple: tenemos en cuenta cuántas estructuras reconoce SpQA como correctas respecto al total.

5.1.3 Corpus de preguntas

Para llevar a cabo nuestra evaluación, utilizamos un corpus de preguntas creado *ad hoc*.

152 Cf. Musillo, y Sima'an, (2002, p. 1) para una selección de estos modelos.
153 Para el inglés, el más usado, es el Penn Treebank (Marcus et al., 1993).
154 http://en.wikipedia.org/wiki/F1_score

El corpus contiene 170 preguntas, de las cuales:

- 100 son interrogativas parciales.
- 50 son interrogativas totales.
- 20 son interrogativas disyuntivas (10 disyuntivas propiamente dichas y 10 parciales con disyunción).

El número total de preguntas no es muy elevado debido al tipo de análisis manual que necesitamos para nuestra evaluación (el cual detallaremos en la sección siguiente). En cuanto al reparto por tipo de preguntas, hemos dado prioridad a las parciales porque gran parte del trabajo de esta tesis gira en torno al análisis de la frase interrogativa.

Las preguntas se han extraído aleatoriamente de dos de nuestros corpus Clef y Wiki[155]. Un 90% corresponde a CLEF y un 10% a WIKI[156].

5.1.4 El *gold standard*

A partir del análisis manual de este corpus de 170 preguntas se elaboró un *Gold Standard* (GS). El GS consiste en las 170 preguntas con la siguiente anotación:

- Tipo de pregunta: parcial, total o disyuntiva.
- Análisis sintáctico de cada pregunta: dependencias presentes en las preguntas con sus funciones; constituyentes oracionales y sus funciones; núcleo sintáctico de los constituyentes oracionales.
- Análisis semántico de la frase interrogativa: valor semántico del constituyente interrogativo según los parámetros de SpQA (cf. capítulo 4, sección 4.4.3.4.2.2.2).
- Anotación de entidades nombradas (en una selección de 100 preguntas) y fechas (todas las del corpus).

5.2. Resultados de la evaluación

Construido el GS, se analizó el corpus de 170 preguntas con SpQA y se procedió a la evaluación de las variables enumeradas más arriba, comparando la salida del *parser* con el GS.

A continuación recogemos los resultados obtenidos.

Para aquellas variables a las que se aplica el método *PARSEVAL*, ofrecemos los resultados correspondientes a *precision*, *recall* y *F-score*.

5.2.1 Reconocimiento como preguntas

La siguiente tabla recoge los resultados obtenidos:

155 Las preguntas de Wiki se han sometido a una corrección manual.

156 Se han seleccionado pocas preguntas de Wiki porque exigían una corrección manual para ser utilizadas en la evaluación.

Precision	0.99
Recall	0.99
F-score	0.99

Tabla 27: Reconocimiento como preguntas.

De las 170 preguntas del corpus SpQA falla únicamente en el reconocimiento de una de las totales, que analiza como si fueran dos (totales) en lugar de una.

5.2.2 Reconocimiento de la variable interrogativa

En la clasificación de las preguntas como parciales (`whQ`), totales (`ynQ`) o disyuntivas (`disjQ`), estos son los resultados:

	Parciales	**Totales**	**Disyuntivas**	**Total**
Precision	1	0.96	1	0.99
Recall	1	0.98	1	0.99
F-score	1	0.97	1	0.99

Tabla 28: Reconocimiento de la variable interrogativa.

SpQA clasifica y, por lo tanto, determina correctamente la variable interrogativa de 169 de las 170 preguntas del corpus. El error viene de la misma pregunta que causaba error en la variable anterior (recordemos que el *parser* reconoce dos totales cuando en el GS hay solo una).

5.2.3 Análisis sintáctico global

Seguimos el siguiente orden en la exposición: resultados de las parciales, las totales, las disyuntivas y datos globales. Recordemos que se evalúan las siguientes variables:

- **Constituyentes oracionales**: límite de constituyente oracional (LC), núcleo de constituyente oracional (NC), función de constituyente oracional correcto (FCC), función de constituyente con el núcleo sintáctico correctamente delimitado (FNC).
- **Dependencias**: reconocimiento de dependencias (RD) y función de las dependencias (FD).

5.2.3.1 Parciales

5.2.3.1.1 Constituyentes oracionales

En la siguiente tabla sintetizamos los resultados para todas las variables.

	LC	NC	FCC	FNC
Precision	0.94	0.97	0.88	0.89
Recall	0.96	0.99	0.9	0.91
F-score	0.95	0.98	0.89	0.90

Tabla 29: Análisis de constituyentes oracionales en las preguntas parciales.

Los resultados del análisis de las parciales son en general buenos, siempre por encima del 0.85 en todas las medidas.

El *recall* es, para todos los valores, más alto que la *precision*, siempre con valores por encima de 0.9. Veremos que la superioridad del *recall* frente a la *precision* será una constante para todas las variables y todos los tipos de preguntas.

Los resultados son mejores en el reconocimiento de constituyentes que en el análisis de funciones, donde la *precision* alcanza su valor más bajo (0.878). Esto también será una constante para todos los tipos de preguntas.

En el reconocimiento de constituyentes, los resultados están en casi todos los casos por encima del 0.95. El reconocimiento del núcleo del constituyente mejora claramente los resultados del reconocimiento del constituyente completo.

Los resultados más bajos corresponden a la asignación de funciones al constituyente correctamente identificado (FCC), aunque ninguno de los valores está por debajo de 0.87. Otra vez, los datos relativos al núcleo (FNC) mejoran los del constituyente completo.

5.2.3.1.2 Dependencias

	RD	FD
Precision	0.94	0.91
Recall	0.95	0.92
F-score	0.95	0.91

Tabla 30: Análisis de dependencias en las preguntas parciales.

Los resultados son buenos, con un *F-score* por encima de 0.9 para los dos valores evaluados.

Como en el caso de los constituyentes a nivel oracional, el *recall* es superior a la *precision,* y los valores en la asignación de funciones inferiores a los valores en el reconocimiento de de dependencias. Estos dos puntos serán, también para el análisis en dependencias, constantes para todos los tipos de preguntas.

Comparando los resultados con los de las variables correspondientes a los constituyentes oracionales, se observa que el reconocimiento de dependencias tiene un *F-score* ligeramente inferior al reconocimiento de los límites del constituyente oracional. Sin embargo, los resultados en la asignación de funciones son superiores en el caso de las dependencias. En general, los valores para reconocimiento de unidades y asignación de función están más próximos entre sí en el caso de las dependencias que en el caso de los constituyentes a nivel oracional, donde existe un mayor desequilibrio.

5.2.3.2 Totales

5.2.3.2.1 Constituyentes oracionales

	LC	NC	FCC	FNC
Precision	0.84	0.87	0.8	0.82
Recall	0.91	0.94	0.86	0.89
F-score	0.87	0.91	0.83	0.85

Tabla 31: Análisis de constituyentes en las preguntas totales.

Los resultados generales son inferiores a los de las parciales, aunque ningún valor se sitúa por debajo de 0.8.

El análisis de los resultados es paralelo al anterior: *recall* superior a *precision,* valores más altos en el núcleo del constituyente tanto en identificación como en asignación de función, valores más bajos para la asignación de función al constituyente completo respecto a la identificación del límite de constituyentes.

5.2.3.2.2 Dependencias

	RD	FD
Precision	0.91	0.88
Recall	0.94	0.90
F-score	0.92	0.89

Tabla 32: Análisis de dependencias en las preguntas totales.

Los resultados también son buenos en general, aunque inferiores a los del análisis por dependencias en las parciales (como también ocurría en las variables de la sección anterior).

Comparando con los resultados del análisis por constituyentes oracionales, los resultados son superiores para las dos variables medidas (como también ocurría con las parciales).

5.2.3.3 Disyuntivas

5.2.3.3.1 Constituyentes oracionales

	LC	NC	FCC	FNC
Precision	0.96	0.98	0.86	0.87
Recall	0.97	1	0.87	0.88
F-score	0.96	0.99	0.86	0.88

Tabla 33: Análisis de constituyentes en las preguntas disyuntivas.

Los resultados de las disyuntivas son globalmente superiores a los de las totales. Respecto a a las parciales, son mejores en el reconocimiento de constituyentes, pero peores en la asignación de función. Ningún valor se sitúa por debajo del 0.85.

Llama la atención, como decíamos, un cierto desequilibrio entre reconocimiento de constituyentes, con todos los valores por encima del 0.95, y asignación de funciones, cuyo valor más alto no llega a 0.9.

5.2.3.3.2 Dependencias

	RD	FD
Precision	0.94	0.91
Recall	0.96	0.93
F-score	0.95	0.92

Tabla 34: Análisis de dependencias en las preguntas disyuntivas.

Los resultados para las dos variables están por encima del 0.9 en todas las medidas.

Como en el caso de las parciales, existe un mayor equilibrio entre los resultados de las dos variables que en el análisis por constituyentes.

5.2.3.4 Global

En esta sección analizamos los resultados correspondientes a todas las preguntas del corpus.

5.2.3.4.1 Constituyentes oracionales

	LC	NC	FCC	FNC
Precision	0.91	0.94	0.85	0.87
Recall	0.95	0.98	0.89	0.90
F-score	0.93	0.96	0.87	0.89

Tabla 35: Análisis de constituyentes: resultados globales.

Ningún valor se encuentra por debajo de 0.85. El *F-score* en el reconocimiento de constituyentes se sitúa por encima de 0.9 para las dos variables y en la asignación de funciones por encima de 0.86.

5.2.3.4.2 Dependencias

	RD	FD
Precision	0.93	0.90
Recall	0.95	0.92
F-score	0.94	0.90

Tabla 36: Análisis de dependencias: resultados globales.

Todos los valores se sitúan otra vez por encima de 0.9.

Los resultados son superiores al del análisis por constituyentes. Además, como ocurría en el caso de las parciales y las disyuntivas, los valores para las dos variables medidas son más equilibrados que los de las variables en el análisis por constituyentes oracionales.

5.2.4 Análisis sintáctico semántico de la frase interrogativa

En esta sección, a las variables sintácticas se suma la evaluación en la asignación de un valor semántico de entre los posibles en SpQA. Los resultados concretos para el constituyente interrogativo son los siguientes.

5.2.4.1 Constituyentes oracionales y valor semántico de la frase interrogativa

	LC	NC	FCC	FNC	SEM
Precision	1	1	0.9	0.9	1
Recall	1	1	0.9	0.9	1
F-score	1	1	0.9	0.9	1

Tabla 37: Análisis de constituyentes y análisis semántico de la frase interrogativa.

En lo que respecta al constituyente interrogativo en solitario, los resultados son mejores que los anteriores.

En el plano sintáctico, solo hay errores en la asignación de funciones (10 preguntas de 100).

En la asignación de los valores semánticos de SpQA, el *parser* obtiene una eficacia del 100%.

5.2.4.2 Dependencias

	RD	FD
Precision	1	0.92
Recall	1	0.92
F-score	1	0.92

Tabla 38: Análisis de dependencias de la frase interrogativa.

En el caso de la frase interrogativa, los resultados del análisis en dependencias son más similares a los del análisis en constituyentes oracionales. Como ocurría en el apartado anterior, no hay errores en el reconocimiento y asignación de dependencias en la frase interrogativa. Los errores se limitan a la asignación de funciones a esas dependencias, y son los mismos que los del apartado anterior (es decir: son errores en la asignación de las funciones de las dependencias a nivel oracional y no de la frase).

5.2.5 Reconocimiento de entidades nombradas y fechas

Entidades nombradas

El análisis se ha realizado sobre un corpus de entidades nombradas extraídas de 100 preguntas. El total de entidades del corpus asciende a 60.

Fechas

Debido al escaso número de fechas en las preguntas del corpus, se han tenido en cuenta todos los casos que, aún así, suman solo un total de 7.

Resultados

	TOTAL	SPQA
NER (selección 100 preguntas)	60	51 - 85%
FECHAS (total en el corpus)	7	6 – 85.71%

Tabla 39: Reconocimiento de entidades nombradas y fechas.

Como vemos, los resultados son similares para las dos variables: en torno al 85% de eficacia.

5.3 Análisis de errores

En esta sección llevamos a cabo un sucinto análisis[157] de los principales errores de SpQA en el plano sintáctico.

De las variables evaluadas, nos ocupamos de los errores en los siguientes ámbitos:

- reconocimiento de unidades: reconocimiento de dependencias (RD) y constituyentes oracionales (LC);
- etiquetado de unidades: etiquetado de la función de dependencias (FD) y constituyentes oracionales (FCC).

5.3.1 Reconocimiento de unidades

En esta sección nos ocupamos de los principales errores de SpQA en el reconocimiento de unidades sintácticas. Para ello, clasificaremos en grupos los errores del *parser* en el reconocimiento y asignación de dependencias, indicando, cuando sea

157 Se analizan los errores que presentan más de un caso.

pertinente, qué errores influyen en el reconocimiento correcto de constituyentes a nivel oracional (LC). Mostraremos, además, ejemplos para cada tipo de error.

Los errores que comete SpQA en el reconocimiento de unidades afectan a 48 dependencias de las 1012 que hay en el corpus. Los errores tienen la siguiente distribución por tipo de pregunta:

	Parciales	Totales	Disyuntivas	Total
Errores RD	26	15	7	48
Total dependencias	581	246	185	1.012

Tabla 40: Distribución de errores en el reconocimiento de dependencias por tipo de pregunta.

Los principales errores que comete SpQA en el reconocimiento y asignación de dependencias son los siguientes:

1) Error en el reconocimiento de entidades nombradas (NE): afecta a pocos casos (10.41%[158]). En nuestra evaluación computamos como errores todos los casos en los que la NE no es reconocida correctamente como PN. Esto incluye tanto los casos en los que no se reconocen correctamente los límites de la entidad nombrada (1), como aquellos en los que estos sí se reconocen, pero SpQA no etiqueta el constituyente como PN (2).

(1) # (null) 1 0-44|¿En qué país está **El Lago de los Cisnes**?
whQ [V: estar third sing present **<SUBJ PN: El Lago** <QCIRC_LOCATIONen N: país **<CIRCde [N: cisnes <DET D: los]**]

(2) # (null) 2 0-59|¿Cuántos miembros tuvo el partido nazi durante la **Segunda Guerra Mundial**?
whQ [V: tener third sing past <QOBJ [QUANTITY <MODde N: miembros] <CIRCdurante **[PN: Guerra Mundial <MOD A: segunda <DET D: la]** <SUBJ [N: partido <MOD A: nazi <DET D: el]]

2) Dependencias mal asignadas: este es el error más común que se da en el análisis de SpQA (el 68.75% de los errores son de este tipo). Podemos subclasificar los errores en la asignación de dependencias en los siguientes subtipos:

2.1) Dependencia a nivel de la frase que se asigna al nivel de la oración: es el error más común (35.41%). Dos son los errores que más se repiten:

158 El tanto por ciento se calcula respecto al total de errores.

- frase preposicional que funciona como MOD que es reconocida como CC (3);
- unidad que funciona como aposición y que es reconocida como un constituyente a nivel oracional (4).

(3) # (null) 32 0-81|¿Cuántos miembros de la escolta murieron en **el atentado contra el juez Falcone**?
whQ [V: morir third plu past <CIRCen [N: atentado <DET D: el] <QSUBJ [QUANTITY <MODde N: miembros <MODde [N: escolta <DET D: la]] **<CIRCcontra [N: juez <MOD PN: Falcone <DET D: el]**]

(4) # (null) 58 0-67|¿Qué significa **la palabra francesa "brut"** en una botella de vino?
whQ [V: significar third sing present **<OBJ UNKNOWN N: "brut" <CIRC [N: palabra <ATTR A: francesa <DET D: la]** <QSUBJ ENTITY <CIRCen [N: botella <MODde N: vino <DET D: una]]

Como se puede ver, ambos errores corresponden al ámbito de la modificación dentro de la frase nominal.

Estos errores provocan fallos en el reconocimiento de los límites de los constituyentes oracionales (LC) constituidos por NP.

2.2) *Head* mal asignada dentro de una frase nominal compleja: es el segundo tipo de error más común (14.58%).

(5) # (null) 88 0-47|¿Qué son **los hexagonales del panal de abejas**?
whQ [V: ser third plu present <QPRED DEFINITION **<SUBJ [N: hexagonales <MODde [N: panal <DET D: el] <MODde N: abejas <DET D: los]**]

TRIPLETES
[N: hexagonales, DET, D: los]
[N: hexagonales, MODde, N: abejas]
[N: hexagonales, MODde, N: panal]
[N: panal, DET, el]
[V: ser third plu present, QPRED, DEFINITION]
[V: ser third plu present, SUBJ, N: hexagonales]

En el ejemplo, el <MODde abejas se asigna a la *head hexagonales* en lugar de a *panal*.

Estos errores afectan al interior de la frase nominal y, por tanto, no provocan problemas en el reconocimiento de los límites de los constituyentes oracionales (pero sí en su estructura interna).

2.3) Dependencia al nivel de la oración asignada al nivel de la frase: es el tercer tipo de error más común dentro de este grupo (10.41%). Estos casos suelen estar generados por otros errores de análisis en las oraciones. No hay ningún patrón que se repita dentro de este grupo.

(6) # (null) 1 0-26|¿Es la araña un insecto?
ynQ [V: ser third sing present <PRED [N: araña **<MOD [N: insecto <DET D: un]** <DET D: la]]

Este error afecta al reconocimiento de constituyentes oracionales desempeñados por NP.

2.4) Error en la asignación de la dependencia al verbo en casos con verbos subordinados: la dependencia se asigna al verbo equivocado. Hay pocos casos (6.25%).

(7) # (null) 2 0-71|¿Existen en España especies de caza que mudan anualmente **los cuernos**?
ynQ [V: existir third plu present <OBJ [N: especies <MODde N: caza >SUBJ [V: mudar third plu present <CIRC X: anualmente]] <CIRCen PN: España **<SUBJ [N: cuernos <DET D: los]**]

TRIPLETES
[N: cuernos, DET, D: los]
[N: especies, MODde, N: caza]
[V: existir third plu present, CIRCen, PN: España]
[V: existir third plu present, OBJ, N: especies]
[V: existir third plu present, SUBJ, N: cuernos]
[V: mudar third plu present, CIRC, X: anualmente]
[V: mudar third plu present, SUBJ, N: especies]

En el ejemplo, el *cuernos* es dependiente del verbo principal, *existir*, en lugar de del verbo subordinado, *mudar*. Este error, además, provoca que los límites del constituyente a nivel oracional *especies de caza que mudan anualmente los cuernos*, no se reconozcan, de forma que el constituyente es reconocido solo parcialmente (*especies de caza que mudan anualmente*).

3) Errores en el reconocimiento de la dependencia en estructuras disyuntivas y coordinadas: hay pocos casos (8.33%).

(8) # (null) 94 0-83|¿Qué factores influyen para que la población viva o no en determinadas regiones?
whQ [V: influir third plu present <OBJ [N: para <MOD X: que <MOD [N: población **<MOD disjunction <DISJ A: viva <DISJ X: no]** <DET D: la]] <QSUBJ N: factores <CIRCen [N: regiones <ATTR A: determinadas]]

Afecta al reconocimiento de constituyentes oracionales. En concreto, afecta en tres casos a los límites de verbos en coordinación o disyunción (como en el ejemplo).

Reconocimiento del NC y reconocimiento del LC

Como explicábamos más arriba, una de las variables evaluadas es el reconocimiento correcto de, al menos, el núcleo del constituyente sintáctico (NC).

Si observamos la variable NC en relación a los casos erróneos de LC, podemos hacernos una idea de si la variable NC puede aportar algo o no al análisis de preguntas.

Revisados los casos de errores en LC con NC correcto, nuestra conclusión es que, en general, la correcta asignación del núcleo del constituyente no aporta información útil cuando los límites del constituyente han sido mal asignados. Esto se debe a que, en la mayoría de los casos, la única identificación del núcleo del constituyente no es suficiente para determinar la entidad a la que apunta ese constituyente. Veamos un ejemplo:

(9) `# (null) 28 0-60|¿Cuál fue` **`el nombre en clave del Desembarco de Normandía`**`?`
`whQ [V: ser third sing past` **`<PRED [N: nombre <DET D: el]`** `<QSUBJ ENTITY` **`<CIRCen [N: clave <MODde [PN: Desembarco de Normandía <DET D: el]]`**`]`

En el ejemplo, SpQA reconoce como `PRED` el constituyente *el nombre* en lugar de *el nombre en clave del desembarco de Normandía.* El resto de la información, *en clave del desembarco de Normandía*, es analizada como `CIRC`. De esta manera, la interpretación que da SpQA de la pregunta es totalmente errónea, al considerar *clave del Desembarco de Normandía* una entidad.

5.3.2 Asignación de funciones

En este apartado nos ocupamos de los errores en la asignación de una etiqueta de función sintáctica a cada dependencia. Como en el caso anterior, clasificamos los errores por grupos y señalamos los errores que afectan al etiquetado de constituyentes oracionales.

Por tipos de pregunta, este es el reparto de errores en la asignación de funciones sintácticas a las dependencias:

	Parciales	Totales	Disyuntivas	Total
Errores FD	47	24	13	84
Total dependencias	581	246	185	1.012

Tabla 41: Distribución de errores en la asignación de funciones sintácticas a las dependencias.

Se pueden distinguir los siguientes tipos de errores:

1) Error en el reconocimiento de entidades nombradas (NE): son los mismos casos que en la sección anterior (5.95%).

Este error afecta al etiquetado de constituyentes oracionales pese a que, en algunos casos incluidos como errores, SpQA asigna una función adecuada al fragmento de NE:

(10) # (null) 13 0-59|¿Se necesita Internet para instalar el **Age of Empires III**?
ynQ [V: necesitar third sing present <REF se third GENDER sing <PCpara [V: instalar **<OBJ [PN: Age <DET D: el] <PRED [UNKNOWN N: of <MOD PN: Empires III]]** <SUBJ PN: Internet]

2) Errores en funciones aisladas: es el error más común (66.66% del total). Estos errores se pueden subclasificar en los siguientes tipos:

2.1) Confusión entre dos funciones a nivel oracional: (39.28%).

Los errores son variados y no hay ningún patrón de error que se repita con mayor frecuencia.

Afecta a la etiquetación de constituyentes a nivel oracional.

(11) # (null) 24 0-39|¿Metallica vendrá otra vez **a México**?
ynQ [V: venir third sing future <CIRC [N: vez <DET D: otra] **<IOBJ PN: México** <SUBJ PN: Metallica]

Por funciones, se da el siguiente reparto de errores (nos referimos a número de casos en los que una función no ha sido correctamente asignada):

- CC = 13.09%[159]
- SUBJ = 7.14%
- OBJ = 4.76%
- PRED = 4.76%
- VBO = 1.19%

2.2) Asignación de una función al nivel de la oración cuando la dependencia funciona a nivel de la frase (19.04%). Como veíamos en la sección anterior (cf. *supra*), hay dos grandes tipos de errores:

159 Respecto al CC, en 3 de estos 11 casos la función asignada es el PC, que podría ser considerada una función sintáctica equiparable al CC.

- frase preposicional que funciona como MOD que es reconocida como CC (3);
- unidad que funciona como aposición y que es reconocida como SUBJ u OBJ (4).

2.3) Asignación de una función al nivel de la frase cuando la dependencia funciona a nivel de la oración: menos común (5.92% de los casos). Casos variados y sin un patrón claro (6).

3) Errores en pares de funciones: se incluyen en este grupo errores que se producen entre un par de funciones en una misma pregunta (21.42%). Hay dos tipos de errores:

- confusión SUBJ/OBJ (14.28%);

(12) `# (null) 14 0-45|¿Qué película ganó el Oso de Oro de 1988?`
`whQ [V: ganar third sing past` **`<QOBJ N: película <SUBJ [PN: Oso de Oro <DATEde 1988 <DET D: el]`**`]`

- confusión SUBJ/PRED (7.14%).

(13) `# (null) 38 0-47|¿Cuál es la extensión de la Selva Lacandona?`
`whQ [V: ser third sing present` **`<PRED [N: extensión <MODde [PN: Selva Lacandona <DET D: la] <DET D: la] <QSUBJ ENTITY`**`]`

Esta confusión se da siempre en las preguntas parciales, entre el interrogativo y un constituyente oracional.

5.3.3 Errores en el constituyente interrogativo

Como hemos visto, los errores en el constituyente interrogativo se limitan a la asignación de funciones, donde 10 de las 100 preguntas contienen errores.

Estos 10 errores se clasifican de la siguiente manera:

- 5 casos en los que hay confusión QOBJ/QSUBJ;
- 4 casos en los que hay confusión QPRED/QSUBJ;
- 1 caso en el que un QIOBJ se confunde con QPC.

5.3.4 Conclusiones sobre el análisis de errores

En relación al reconocimiento de unidades, podemos concluir que SpQA necesita mejoras, principalmente, en dos aspectos:

- Modificación en las NP: reconocimiento de PP como modificadores (y no como *CIRC*) y de las aposiciones; correcta asignación de dependencias en los modificadores en frases nominales complejas.

- Reconocimiento de estructuras coordinadas/disyuntivas (en lo que respecta a verbos, fundamentalmente).

Por otro, si valoramos los errores de análisis sintáctico en un contexto de análisis de la pregunta en BR, los errores más graves nos parecen aquellos relacionados con el reconocimiento del verbo y de las NP. A nivel oracional, las PP y XP suelen funcionar como *CIRC*, pero las NP son generalmente las entidades sobre las que recae el mayor peso informativo de las preguntas. Es por eso que queda claro que se debe prestar especial atención a los errores de modificación, tanto a los que afectan al reconocimiento de las NP en su totalidad, como a los que afectan a la estructura interna de esas NP. Asimismo, el verbo es la entidad que muestra generalmente qué relación mantienen estas entidades entre sí en el marco determinado por los *CIRC*. Es por eso por lo que el reconocimiento de NP y de verbos es crucial en SpQA. A este respecto, es interesante apuntar que, en algunos casos en los que los límites de la NP no se reconocen, al menos estas sí se reconocen parcialmente (cf. ejemplo (7)), de manera que es viable algún tipo de búsqueda de información a partir de este segmento parcial de información.

En lo que se refiere a la asignación de funciones, observamos ciertos patrones de error.

En los errores de función aislados, los más comunes son aquellos en los que se asigna una función a nivel oracional equivocada. En este grupo, la función que más problemas presenta es el sujeto, seguido del complemento circunstancial. El segundo tipo de error más común del *parser* es justamente el contrario: el que consiste en asignar un valor oracional a un constituyente al nivel de la frase. Aquí destacan dos funciones como las que más problemas provocan, con diferencia: las PP que funcionan como *MOD* en una NP y son identificadas como *CC*; las aposiciones que funcionan como *MOD* y son identificadas como *SUBJ* u *OBJ*.

En las confusiones entre pares de funciones en una misma oración, tenemos que dos son los pares que se confunden de forma sistemática: *SUBJ/OBJ* y *SUBJ/PRED*. Esta confusión se da siempre entre la frase interrogativa y otro constituyente oracional.

5.4. Conclusiones generales del capítulo

En el presente capítulo hemos abordado la evaluación de SpQA.

El método de evaluación persigue medir la capacidad de SpQA en el tratamiento de aquellos parámetros considerados como objetivos de nuestro sistema en el análisis de preguntas. Por esta razón, la evaluación es de tipo intrínseco y utiliza las medidas del sistema *PARSEVAL* (excepto para el reconocimiento de entidades nombradas y fechas), determinando *precision*, *recall* y *F-score* a partir de un *gold standard* construido manualmente.

Los resultados muestran que el sistema presenta una eficacia aceptable en todas las variables evaluadas. El *recall* es siempre superior a la *precision*, al igual que el reconocimiento de constituyentes a la asignación de función. Ningún resultado está por debajo de 0.85 en todas las variables evaluadas. Los resultados para el constituyente interrogativo en concreto, son superiores a los resultados globales: SpQA solo falla en la asignación de función en en un 10% de los casos. El análisis semántico tiene una

eficacia del 100%.

Se ha procedido también a un análisis sucinto de los principales errores del sistema.

En lo que respecta al reconocimiento de constituyentes, se destaca como crucial solucionar los problemas en el reconocimiento de NP (incluyendo en ellas las entidades nombradas) y verbos. Para las primeras, parece claro que son necesarias mejoras en el análisis de la modificación. Respecto al reconocimiento de verbos, se necesitan mejoras en el análisis de estructuras coordinadas/disyuntivas.

En relación a la asignación de funciones, se ha demostrado que los errores están más localizados, tanto a nivel global como para el constituyente interrogativo. Las funciones que más se confunden con otras a nivel oracional son el *SUBJ* y el *CC*. Las PP y las aposiciones presentan problemas para ser reconocidas como *MOD* dentro de una NP. Finalmente, en las preguntas parciales se da la confusión *SUBJ/OBJ* y *SUBJ/PRED*, siempre entre frase interrogativa y otro constituyente oracional.

Conclusiones

1. SpQA y BR: resultados

En la Búsqueda de Respuestas (BR), un usuario plantea una pregunta en lenguaje natural a un sistema automático que trata de buscar una respuesta en una o varias fuentes de conocimiento, extraerla y presentarla. La BR es una tarea compleja que implica la puesta en marcha de distintos procesos interdependientes. En esta tarea, se distinguen tres pasos básicos: análisis de la pregunta, búsqueda de la respuesta en la base de conocimiento, selección, extracción y presentación de la respuesta al usuario.

En el proceso de BR, el manejo de un análisis lingüístico profundo que trate de acercarse a la semántica del texto parece ser el planteamiento ideal. No obstante, aplicar este tipo de análisis a cada uno de los pasos citados puede resultar muy complejo, ya que en la búsqueda de la respuesta se procesan grandes cantidades de texto. Sin embargo, aplicar este tipo de procesamiento al análisis de la pregunta es no solo viable (las preguntas son estructuras cortas y, desde el punto de vista estructural, hasta cierto punto simples) sino deseable (especialmente por la especial relación semántica que mantienen las preguntas con sus respuestas). El análisis de la pregunta es además un paso fundamental en el proceso de BR ya que supone, en la mayoría de los sistemas, el punto de partida del que dependen todos los pasos posteriores. Si falla la interpretación que el sistema da de la pregunta, lo más probable es que la respuesta recuperada por este no sea la adecuada. Es por ello por lo que manejar una interpretación lo más cercana posible al significado de la pregunta construida a partir de un análisis lingüístico profundo constituye una tarea deseable en BR.

Dentro de ese análisis lingüístico profundo de la pregunta, el análisis sintáctico es una parte importante ya que aporta la información estructural, fundamental para la correcta interpretación de la pregunta. En la mayoría de los sistemas de BR que manejan un análisis semántico, este se construye además sobre un análisis sintáctico previo.

La mayoría de los sistemas de BR utilizan *parsers* de tipo general para realizar la tarea de análisis sintáctico, pese a que se han demostrado las limitaciones de este tipo de analizadores en dominios específicos (Gildea, 2001; McClosky, et al., 2006; Foster, 2010) y, concretamente, en el análisis de preguntas (Hermjakob, 2001; Petrov et al., 2010 para el inglés o Gayo, 2011a, y Gayo, 2011b, para el español). Por esta razón se ha señalado la necesidad de utilizar *parsers* ligeros y específicos para el

análisis de preguntas en el área de BR (Fliedner, 2007; Katz y Lin, 2003).

Partiendo de las premisas anteriores, en este trabajo se defiende una BR motivada lingüísticamente y se plantea SpQA: un analizador diseñado para BR en español y construido a partir de información lingüística (de tipo teórico y de uso real) sobre el funcionamiento de las preguntas.

La profundización en la BR y, en concreto, en las necesidades del análisis de la pregunta, constituye el punto de partida de SpQA (capítulo 1).

El estudio lingüístico sobre el que se asienta SpQA (capítulos 2 y 3) permite la identificación de los rasgos (gramaticales y semánticos) pertinentes para obtener el significado de la pregunta. En el trabajo teórico se definen las preguntas como oraciones interrogativas directas que se utilizan para demandar información, y se presentan sus principales características (gramaticales y semánticas) de funcionamiento. El trabajo en corpus (capítulo 3), posibilita concretar y al mismo tiempo ampliar la información proporcionada por la teoría (capítulo 2). Concreta porque permite determinar, hasta donde es posible, el uso real de los rasgos lingüísticos señalados por los estudios teóricos. Amplía porque proporciona información sobre el funcionamiento de las preguntas no contemplada por la teoría (otras ordenaciones de argumentos posibles, significados nuevos para la frase interrogativa en las parciales, incidencia de la negación, etc.).

Toda esta información se sintetiza en la gramática formal escrita en AGFL a partir de la que se genera SpQA (capítulo 4). El *parser* tiene en cuenta las necesidades de la BR en el análisis de preguntas y proporciona un análisis sintáctico y semántico de estas lo más exhaustivo posible. Este análisis se construye a partir de la información léxica y sintáctica presente en la pregunta (sin utilizar medios externos) y se sintetiza en un grafo de dependencias. El grafo condensa en una representación simple información de tres niveles:

1) **Léxico**: etiquetado de la clase de palabra para cada una de las palabras representadas en el grafo[160].

2) **Sintáctico.**
   a) Identificación y etiquetado de las dependencias sintácticas de la pregunta (a nivel de la frase y de la oración).
   b) Identificación (mediante corchetes; cf. capítulo 4), de los límites de los constituyentes oracionales y de su función sintáctica.
   c) Normalizaciones sintácticas («despasivización», marcado de estructuras impersonales, normalización del formato de las fechas, etc.).

3) **Semántico.**
   a) Identificación, a partir de la estructura sintáctica, de la variable interrogativa o incógnita presente en la pregunta (diferenciación entre preguntas totales, parciales y disyuntivas).
   b) Especificación de valores semánticos concretos de la variable en las parciales. Esta especificación del valor semántico se construye a partir de la

160 En el capítulo 4 vimos que no todas las palabras presentes en la pregunta tienen representación en el grafo.

información léxica (presente en el interrogativo, el verbo, los sustantivos y las preposiciones) y de la información sintáctica (un interrogativo en determinada construcción se asocia a un determinado valor, por ejemplo, *quién* + verbo *ser* + entidad nombrada = 'descripción').

c) Identificación de entidades nombradas, estructuras cuantificativas y estructuras temporales.
d) Posibilidad de establecer la diferenciación entre preguntas abiertas y preguntas informativas a partir del análisis semántico del interrogativo[161].
e) Tratamiento (limitado) de la paráfrasis de la frase interrogativa: SpQA analiza como equivalentes ciertas estructuras sintácticas de la frase interrogativa que tienen el mismo significado (por ejemplo: *qué modo/manera = cómo*; cf. capítulo 4, sección 4.4.3.4.2.2.2).

El grafo de SpQA permite además la extracción de tripletes de dependencias.

Sumado a lo anterior, las características del formalismo AGFL posibilitan añadir, eliminar o modificar de forma sencilla la información representada en el grafo mediante cambios en la *transduction* de la gramática formal a partir de la que se genera SpQA.

La evaluación de SpQA (capítulo 5), finalmente, demuestra que lo que se propone con el *parser* se consigue de modo aceptable, tanto a nivel sintáctico como semántico (especialmente a este último).

El análisis de errores muestra, en el nivel sintáctico, la necesidad de mejoras en el tratamiento de la frase nominal, especialmente, en lo concerniente a la modificación. Los problemas en la frase nominal afectan a su vez a la identificación y etiquetado de constituyentes oracionales. También son necesarias mejoras en la identificación de estructuras copulativas y disyuntivas, especialmente, en constituyentes verbales.

En lo referente al etiquetado de funciones sintácticas, el principal problema afecta al par sujeto/objeto directo. Este error podría subsanarse en muchos casos con el manejo de cierta información semántica, fundamentalmente: información sobre la construcción semántica del verbo más cierta información semántica sobre las entidades que funcionan como sujeto y objeto. Veamos un ejemplo:

¿Qué película ganó el Oso de Oro de 1989?

Para la correcta identificación de las funciones sujeto/objeto, el sistema podría valerse de la siguiente información semántica:

1. Valencia semántica del verbo:

ganar = alguien/algo gana algo (= un premio, una carrera, etc.)

161 Este aspecto no está integrado en el *parser*, pero sí se trata en el capítulo 2, donde cada valor semántico se asocia con una pregunta, bien de tipo abierto, bien de tipo informativo (cf. capítulo 2, sección 2.3.2.1.1). Esta información podría utilizarse en el módulo de análisis de la pregunta combinada con la información proporcionada por SpQA sobre el valor semántico del constituyente interrogativo.

SUJETO = entidad que gana
OBJETO DIRECTO = entidad ganada

2. Información semántica sobre las entidades en la pregunta:
Oso de Oro = galardón

Manejando estos dos datos, el sistema podría determinar que el sujeto de la oración es *película* y que el objeto es *Oso de Oro de 1989*.

Si manejar este tipo de información semántica no fuese posible, una estrategia más sencilla podría consistir en utilizar una asignación de roles contraria a la ofrecida por SpQA si el sistema de BR no encuentra ninguna respuesta que encaje con el análisis ofrecido por el *parser*. Siguiendo con el ejemplo, SpQA ofrecería un análisis erróneo en el que el sujeto es *Oso de Oro de 1989* y el objeto directo *película*. Con el mecanismo que planteamos, si, a partir de este análisis, el sistema no encontrase ninguna respuesta, se volvería al análisis de la pregunta y se sugeriría una asignación de roles opuesta: sujeto para *película* y objeto para *Oso de Oro de 1989*.

2. Trabajo futuro

En las secciones siguientes recogemos las vías de desarrollo futuro que se prevén para SpQA.

2.1 Utilización y evaluación de SpQA en un sistema de BR

En primer lugar, el principal aspecto pendiente de SpQA es su integración efectiva en un sistema de BR. Su utilización en un sistema real, capaz de manejar el análisis que el *parser* proporciona, debería arrojar nueva luz tanto sobre sus ventajas como sobre sus carencias y sus necesidades de mejora. Por lo tanto, este aspecto, que no ha sido posible tratar en el marco de este trabajo, es el primer paso natural que se plantea para su desarrollo futuro.

2.2 Carencias en el modelo de SpQA: fórmulas alternativas a las preguntas para demandar información

En los tres corpus con los que se ha trabajado, se ha comprobado la hipótesis de que las preguntas tal y como se entienden en este trabajo constituyen la principal forma utilizada para demandar información. Paralelamente, se han documentado estructuras alternativas para demandar información en dos de nuestros corpus: en Trivial y, especialmente, en Wiki. En algunos casos, esas formas se asemejan a las preguntas que maneja SpQA. Se trata de fórmulas que se parecen estructuralmente a parciales, totales o disyuntivas, pero que presentan ciertas divergencias (gramaticales o semánticas). Estas fórmulas se documentan sobre todo en Wiki, aunque su peso en el corpus, comparado con el de las preguntas «canónicas» o el de ciertas peticiones de información que veremos más abajo (cf. *infra*), es menor.

A continuación recogemos tres ejemplos de este tipo de «preguntas alternativas»:

(1) *¿Para ubicar la ciudad de rana en la imagen satelital se nesesitaría?*[162]
(2) *¿El motin que organizaron los segidores de iturbide para ponerlo como emperador qué ocurrio?*
(3) *¿Dentro del sitema religioso de los curacazgos a las huacas mayores se les denominaba como pacarinas mientras que a las huacasa personales les denominaban como?*

En (1) tenemos una estructura con forma de total en la que, sin embargo, la variable está más cercana a la de una pregunta parcial, ya que se pregunta por un elemento que cumple las características especificadas en la pregunta y no por una afirmación o negación de lo que se dice. De hecho, la siguiente parcial sería equivalente:

(1a) *¿Qué se nesesitaría para ubicar la ciudad de rana en la imagen satelital?*

En (2) tenemos un tipo de pregunta que constituye, estructuralmente, una especie de híbrido entre totales y parciales Semánticamente, sin embargo, se trata de una parcial, ya que la variable interrogativa es la que le corresponde a las parciales (y no *sí/no*). Podría interpretarse que en esta estructura se presenta, en primer lugar, el tema del que trata la «pregunta» (*el motín que organizaron los seguidores de Iturbide para ponerlo como emperado*r), seguido de la parcial que se refiere a ese tema (*qué ocurrió*). Esta fórmula sería equivalente a la parcial, más esperable:

(1) *¿Qué ocurrió en el motín que organizaron los seguidores de Iturbide para ponerlo como emperador?*

O a una pregunta parcial con el tema desgajado:

(2) *En el motín que organizaron los seguidores de Iturbide para ponerlo como emperador, ¿qué ocurrió?*

El ejemplo (3) podría considerarse una variante de (2)[163], pero con el interrogativo en posición final absoluta. La colocación del interrogativo al final es, como hemos visto, bastante particular (cf. capítulo 2). Este tipo de estructuras no son muy abundantes en Wiki, y destaca su uso con el interrogativo *cómo*.

Junto a estas fórmulas similares a las preguntas, en Wiki se documenta también un tipo de petición de información que consiste en una frase nominal interrogada:

(1) *¿10 animales foráneos?*
(2) *¿Bestimenta de honduras?*

162 Los tres ejemplos son de Wiki y, como hasta ahora, mantenemos la ortografía original.
163 No entraremos aquí a valorar si se trata o no de estructuras equivalentes.

(3) *¿Biografia corta de cristobal colon?*
(4) *¿El tiempo cronologico de la odisea?*
(5) *¿El valor del kilolitro?*
(6) *¿El vicepresidente del peru?*

Este tipo de peticiones de información parecen ser muy abundantes en el corpus[164]. Algunos casos parecen simplificaciones de parciales:

(2a) *¿Cuál es la vestimenta de Honduras?*
(5a) *¿Cuál es el valor del kilolitro?*
(6a) *¿Quién es el vicepresidente de Perú?*

Como vemos, en los tres casos se trata de estructuras copulativas en las que se nos pide una entidad (2a) o el «valor» de una entidad (5a, 6a).

En los otros casos, sin embargo, la paráfrasis como parcial es más forzada. Esas peticiones de información podrían parafrasearse como:

(1a) *Quiero el nombre de 10 animales foráneos.*
(3a) *Quiero una biografía corta de Cristóbal Colón.*
(4a) *Quiero información sobre el tiempo cronológico en la Odisea.*

Si nos atenemos al significado, lo que tenemos en realidad en todos los casos es una especie de identificación, es decir, lo que se demanda es «algo» que sea equivalente a esa frase nominal, de ahí que las parciales equivalentes se construyan como copulativas con *cuál*. Como ocurría en las preguntas con *cuál* (cf. capítulo 4), el grado de especificación de esa «identificación» o equivalencia varía: en (1), (2), (5) y (6) se trata de valores concretos (el nombre de diez animales foráneos en (1)[165], un tipo de vestimenta en (2), etc.) del mismo tipo que los propios de las preguntas informativas, mientras que en (3) y (4) se trata de valores más complejos, próximos a los propios de las preguntas abiertas.

Incorporación de estas estructuras a SpQA

Consideramos necesario un estudio de la incidencia de estos tipos de estructuras en corpus para contemplar su integración en SpQA.

Sabemos que las «preguntas alternativas» que veíamos al principio de la sección son escasas, sin embargo, las peticiones de información con forma de frase nominal sí son muy abundantes en Wiki. Con el fin de incorporarlas a SpQA, sería necesario determinar cómo de comunes son y si aparecen en otros corpus de preguntas reales de usuarios.

En caso de implementarlas, estas estructuras se interpretarían semánticamente de forma similar a las preguntas copulativas con *cuál,* considerando que lo que se

164 No tenemos datos exactos, pero un análisis superficial así lo indica.
165 Aunque el autor de la pregunta no ha especificado foráneos respecto a dónde.

pide es un valor equivalente a la frase nominal interrogada. El grado de abstracción de ese valor se acercaría más o menos la estructura a las preguntas abiertas o a las informativas.

2.3 Incorporación de análisis semántico

Otro de los desarrollos futuros deseables para SpQA es la incorporación de un análisis de tipo semántico.

Manejando un lexicón de verbos con valencias semánticas no sería complicado añadir información sobre los roles semánticos al grafo de dependencias. Por ejemplo, utilizando el tipo de información sobre valencia semántica de los verbos codificada en la base de datos ADESSE[166] (García-Miguel, Vaamonde, y Domínguez, 2010).

ADESSE contiene para cada verbo, entre otras cosas, una clasificación de tipo semántico (1) así como la valencia sintáctico semántica (2). Por ejemplo, para el verbo *ganar*[167], ADESSE nos da la siguiente información:

(1) Tipo semántico
GANAR – tipo semántico – ADQUISICIÓN

(2) Esquema valencial sintáctico semántico[168]

Voz	**Argumentos semánticos y Funciones sintácticas**	**N.º_de ejemplos**
GANARact	A1:POS-F=SUBJ ; A2:POS=ODIR	108

A partir del análisis sintáctico de SpQA (1) y manejando el lexicón con la información de ADESSE, sería posible enriquecer la representación del grafo con roles semánticos (2):

```
(1) # (null) 12 0-50|¿Cuántos Óscar ganó La Guerra de las
Galaxias?
whQ [V: ganar third sing past <QOBJ [QUANTITY <MODde PN:
Óscar] <SUBJ PN: La Guerra de las Galaxias]
```

```
(2) ¿Cuántos Oscar ganó La guerra de las Galaxias?
whQ [V_ADQUISICIÓN: ganar third sing past <QOBJ_A2 [QUANTITY
<MODde PN: Óscar] <SUBJ_A1 PN: La Guerra de las Galaxias]
```

De este modo, mediante un mecanismo sencillo, se añadiría una representación de tipo semántico a la que ya existe en SpQA.

166 http://adesse.uvigo.es/
167 En su primera acepción en ADESSE, cf.: http://adesse.uvigo.es/data/verbos.php?sense=1912
168 Reproducimos la representación de ADESSE, cf.: http://adesse.uvigo.es/data/verbos.php?sense=1912

Conclusions

1. SpQA and QA: results

In Question Answering (QA), a user poses a question in natural language to an automatic system, and this system tries to find an answer in a knowledge source, extract it and give it to the user. QA is a complex task that encompasses different processes working interdependently, since the user makes the question until the system retrieves the answer. These processes are structured in three steps: question analysis; analysis of the knowledge source and selection of information pieces that can contain the answer; selection, extraction and generation of the answer.

In QA, the ideal approach seems to be one that uses representations of texts with deep linguistic knowledge, that are close to the semantics of texts. However, applying this kind of analysis in every step of QA process can be very complex because of the amount of text that must be processed searching the question. Applying this processing in question analysis is, nevertheless, viable (questions are short linguistic structures) and desirable (because of the semantic relation between questions and answers). Furthermore, question analysis is a crucial step in QA because all the subsequent processes rely on it.

One crucial aspect of the linguistic analysis of questions is syntactic analysis, because it gives the structural information, which is basic for a right interpretation of questions. Besides, most QA systems that make semantic analysis build this analysis on a previous syntactic representation.

Most QA systems use general parsers for syntactic analysis, despite the fact that some studies have shown that general parsers accuracy drops out of domain (Gildea, 2001; McClosky, et al., 2006; Foster, 2010) . This fact has been shown particularly in question analysis (cf. Hermjakob, 2001; Petrov et al., 2010 for English or Gayo, 2011a, y Gayo, 2011b, for Spanish). For this reason, some researchers suggest the use of specific light parsers designed for question analysis in QA (Fliedner, 2007; Katz & Lin, 2003).

Taking into account these factors, this work defends linguistically motivated QA and proposes SpQA: a parser designed for QA in Spanish and built from linguistic information (theoretical and of real use) about questions.

Description of QA task and, in particular, the description about the needs of question analysis, is the starting point of SpQA (chapter 1).

The linguistic research about questions (chapters 2 and 3) allows the identification of relevant features (grammatical and semantic) to reach the meaning of a question. In the theoretical study (chapter 2), questions are defined as interrogative clauses used to demand information, and their main grammatical and semantic features are described. Corpus work (chapter 3), makes possible the realization and, at the same time, increase of the information obtained in chapter 2. Corpus work increases this information because it provides data about questions not considered in theoretical descriptions (new orders of arguments, new meanings of interrogatives, etc.).

All these aspects and linguistic information are synthesized in the formal grammar (written in AGFL formalism) that generates SpQA. The parser takes into account the needs of question analysis in QA and provides an exhaustive syntactic and semantic analysis of questions. This analysis is built from lexical and syntactic information in the question (external tools as Wordnet are not used) and it is synthesized in a dependency graph. The graph condenses in a simple representation information of three linguistic levels:

1) Lexical: part of speech tagging.
2) Syntactic:
   a) Identification and labeling of syntactic dependencies.
   b) Identification and labeling of syntactic constituents.
   c) Syntactic normalizations (depasivization, normalization of date format, etc.).
3) Semantic.
   a) Question classification (yes/no question; WH-question; disjunctive question).
   b) Identification of the semantic value of the interrogative phrase in WH-questions. It allows a (limited) treatment of paraphrase in the interrogative phrase.
   c) Named Entity, quantities and dates recognition.
   d) Classification in open or informative questions from the analysis of the interrogative phrase.

Furthermore, the graph allows the extraction of dependency triples.

In addition, AGFL makes possible the addition, deletion or change of the information represented in the graph through changes in the transduction specified in the grammar.

Finally, the intrinsic evaluation of SpQA (chapter 5), shows that the goal of the parser is reasonably reached, as much at the syntactic level as at the semantic level (especially in this one).

Error analysis shows the necessity of improvements in the nominal phrase, especially, in the treatment of modification. Errors in the nominal phrase affect also the identification and labeling of clause constituents. Improvements in the identification of copulative and disjunctive structures are also needed, especially, in verbal constituents.

Concerning syntactic labeling, the main problem is the pair subject/object. This error could be corrected with the use of specific semantic information, basically: information about the semantic valence of the verb plus certain information about the entities that function as subject and object. For example:

¿Qué película ganó el Oso de Oro de 1989?
(Which movie did win the Gold Bear of 1989?)

To get the correct identification of subject and object, the system could use this semantic information:

1. Semantic valence of the verb:

ganar (to win) = someone/something wins something/someone (= an award, a race, etc.)
SUBJECT = entity that wins;
OBJECT= entity that is won.

2. Semantic information about entities in the question:

Golden Bear = award.

Handling these data, the system could determine that the subject is *película* (film) and the object *Oso de Oro de 1989* (Golden Bear of 1989).

If using this kind of information is not possible, a simpler strategy could be using a syntactic labeling opposed to that offered by SpQA if the QA system does not find any answer that fits with the syntactic analysis of the parser. In the previous example, SpQA would give a wrong analysis. With the mechanism that we suggest, if, using this wrong analysis, the system does not find an answer, it would return to this analysis changing the roles of the labeling: subject for *película* (film) and object for *Oso de Oro de 1989* (Golden Bear of 1989).

2. Future work

In the next sections we present future work for SpQA.

2.1 Integration and evaluation of SpQA in a QA system

First of all, the main pendent aspect of SpQA is its effective integration in a QA system. This integration in a real QA system, capable of handling the analysis that the parser offers, should give more data about advantages as well as needs of improvement of SpQA. Therefore, this step, that has not been possible in this research, is the first natural and necessary step for future work.

2.2 Lacks of SpQA model: alternative ways to demand information

In our three corpora, we have proved that questions as are described in this dissertation are the main way to ask. At the same time, alternative structures to demand information have been documented in two of our corpora: in Trivial and, especially, in Wiki. In some cases, these structures are similar to SpQA questions. They look like WH-questions, yes/no questions or disjunctive questions, but they present grammatical and semantic differences. These structures are documented most in Wiki, nevertheless, their importance in the corpus, compared with "canonical" questions or other structures (cf. *infra*), is smaller.

Next, we present three examples of this kind of "alternative structures":

(1) *¿Para ubicar la ciudad de rana en la imagen satelital se nesesitaría?*[169]
(To find rana city in the image of the satellite would it need?)
(2) *¿El motin que organizaron los segidores de iturbide para ponerlo como emperador qué ocurrio?*
(The rebellion organised by iturbide followers to proclaim him emperor what did it happen?)
(3) *¿Dentro del sitema religioso de los curacazgos a las huacas mayores se les denominaba como pacarinas mientras que a las huacas personales les denominaban como?*
(In curacazgos religious system, head huacas were named pacarinas while personal huacas were named how?)

In (1) we have a clause with the form of a yes/no question. Its interrogative variable, however, is closer to those of WH-questions, considering that the question asks for an element that agrees with the features specified in the question and not for yes/no. In fact, the next WH-question is equivalent to (1):

(1a) *¿Qué se nesesitaría para ubicar la ciudad de rana en la imagen satelital?*
(What would be needed to find the city of rana in the image of the satellite?)

In (2) we have a clause that is, structurally, a kind of hybrid between yes/no questions and WH-questions. Semantically, however, it is a WH-question, since the interrogative variable is the same as in WH-questions (and not yes/no). It is possible that, in this clause, the topic is presented first (*The rebellion organised by iturbide followers to proclaim him emperor*), followed by the WH-question that refers to this topic (*what did it happen?*). This formula is equivalent to the WH-question:

(1) *¿Qué ocurrió en el motín que organizaron los seguidores de Iturbide para ponerlo como emperador?*
(What did it happen in the rebellion organised by Iturbide followers to proclaim him emperor?)

169 All these examples come from Wiki. We keep the original spelling as we have done along the dissertation.

Or a WH-question where the topic splits off from the question:

(2) *En el motín que organizaron los seguidores de Iturbide para ponerlo como emperador, ¿qué ocurrió?*

(In the rebellion organised by Iturbide followers to proclaim him emperor, what did it happen?)

Example (3) could be considered a variant of (2)[170], but with the interrogative word in final position. The placement of the interrogative at the end of the question is very strange, as we have seen (cf. chapter 2). These types of structures are not very common in Wiki, and in most of the cases the interrogative word used is *cómo* (How).

Together with these formulas similar to questions, Wiki also contains a type of demand for information that consists of an interrogative nominal phrase:

(1) *¿10 animales foráneos?*
(10 foreign animals?)
(2) *¿Bestimenta de honduras?*
(Outfit of Honduras?)
(3) *¿Biografia corta de cristobal colon?*
(Short biography of Cristobal Colon?)
(4) *¿El tiempo cronologico de la odisea?*
(Chronological time in the Odissey?)
(5) *¿El valor del kilolitro?*
(Value of the kiloliter?

This type of information requests seems to be very common in the corpus[171]. Some cases look like simplifications of WH-questions:

(2a) *¿Cuál es la vestimenta de Honduras?*
(Which is the outfit of Honduras?)
(5a) *¿Cuál es el valor del kilolitro?*
(Which is the value of kiloliter)
(6a) *¿Quién es el vicepresidente de Perú?*
(Who is the vice-president of Perú?)

We can see that in these three cases we have copulative structures that ask for an entity (2a) or the value of an entity (5a, 6a).

The rest of the cases do not accept the interpretation as WH-questions. These information requests could be paraphrased as follows:

(1a) *Quiero el nombre de 10 animales foráneos.*
(I demand for the name of 10 foreign animals.)

170 We will not discuss here if these are equivalent structures or not.
171 We do not handle exact numbers, but a superficial analysis suggests this idea.

(3a) *Quiero una biografía corta de Cristóbal Colón.*
(I demand a short biography of Cristobal Colón.)
(4a) *Quiero información sobre el tiempo cronológico en la Odisea.*
(I demand information about chronological time in the Odissey.)

Concerning only the meaning, we have in all these cases an identification, that is, what is demanded is "something" that must be equivalent to the nominal phrase. For this reason, the equivalent WH-questions are copulatives with *cuál* (which) as interrogative word. In fact, the specification of this identification varies, as it happens in WH-questions with *cuál*: (1), (2), (5) and (6) ask for concrete entities (the name of 10 foreign animals in (1), a type of clothes in (2), etc.), of the same type that the entities of informative questions, while in (3) and (4), the values are more complex, close to those of open questions.

Integration of these structures in SpQA

We think that a corpus study of these structures is necessary to consider their integration in SpQA.

We know that the "alternative questions" presented first are very scarce. However, the information requests with the form of nominal phrases are very common in Wiki. For their incorporation in SpQA, we need to know first how common they are and if they appear in other corpora of real questions.

In case they are implemented, these structures would be interpreted semantically as WH-questions with *cuál*, considering that they ask for a value equivalent to the value of interrogative phrases in these WH-questions. The degree of abstraction of this value would get more or less close these structures to open or informative questions.

2.3 Integration of semantic analysis

Another desirable future step for SpQA is the integration of semantic information to the dependency graph.

Using a lexicon of verbs with semantic valences, it would be possible to add data about semantic roles. For example, using the type of information encoded in the database ADESSE[172] (García-Miguel, Vaamonde, & Domínguez, 2010).

ADESSE contains, for every verb, a semantic classification (1) and its syntactic and semantic valence (2). For example, for verb *ganar*[173], ADESSE gives these data:

(1) Semantic type:
GANAR – ADQUISICIÓN
(TO WIN - ADQUISITION)

172 http://adesse.uvigo.es/
173 In its first aception in ADESSE: http://adesse.uvigo.es/data/verbos.php?sense=1912

(2) Syntactic and semantic scheme[174]

Voice in ADESSE	Semantic Roles and Syntactic Functions	Number of examples
GANARact	A1:POS-F=SUBJ ; A2:POS=ODIR	108

Using this lexicon with the information of ADESSE, it would be possible to enrich the SpQA dependency graph (1) with semantic roles (2):

(1) ¿Cuántos Oscar ganó La Guerra de las Galaxias?
[V: ganar third sing <SUBJ PN: La Guerra de las Galaxias <OBJ QUANT PN: Oscar]

How many Oscar did Star Wars win?
[V: win third sing <SUBJ PN: Star Wars <OBJ QUANT PN: Oscar]

(2) ¿Cuántos Oscar ganó La guerra de las Galaxias?
[V_**ADQUISICIÓN**: ganar third sing <SUBJ_**A1** PN: La Guerra de las Galaxias <OBJ_**A2** QUANT PN: Oscar]

How many Oscar did Star Wars win?
[V_**ADQUISITION**: win third sing <SUBJ_**A1** PN: Star Wars <OBJ_**A2** QUANT PN: Oscar]

In this way, through a simple mechanism, a semantic representation would be added to the current analysis of SpQA.

174 We reproduce ADESSE representation: http://adesse.uvigo.es/data/verbos.php?sense=1912

Bibliografía

Álvarez, C. et al. (1998). AVALON, una gramática formal basada en corpus. *Procesamiento Del Lenguaje Natural,* (23), 132-139.

Abney, S. (1991). Parsing by Chunks. In *Principle-based Parsing* (pp. 257-278). Netherlands: Springer.

Alarcos Llorach, E. (1994). *Gramática de la lengua Española*. Madrid: Espasa Calpe.

Androutsopoulos, I., Ritchie, G. D., & Thanisch, P. (1995). Natural Language Interfaces to Databases - An Introduction. *Natural Language Engineering, 1*(709), 50.

Athanasiadou, A. (1991). The Discourse Function of Questions. *Pragmatics, 1* (1), 107-122.

Athanasiadou, A. (1994). The Pragmatics of Answers. *Pragmatics, 4*(4), 561-574.

Attardi, G., Cisternino, A., Formica, F., Simi, M., Tommasi, A. & Zavattari, C. (2002). PiQASso: Pisa Question Answering System. *Word Journal of the International Linguistic Association,* 633-641.

Bäuerle, R. (1979). Questions and Answers. In R. Bäuerle (Ed.), *Semantics from Different Points of View*. Berlin: Springer Verlag.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Boston: Addison Wesley.

Baeza-Yates, R., & Raghavan, P. (2010). Next Generation Search. *Search, 5950*, 11-23.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet Project. In C. Boitet, & P. Whitelock (Eds.*), Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (pp. 86-90). Association for Computational Linguistics.

Barwise, J. & Perry, J. (1983). *Situations and attitudes.* MIT Press.

Belnap, N. D. (1966). Questions, Answers, and Presuppositions. *Journal of Philosophy, 63*(20), 609-611.

Belnap, N. D., & Steel, T. B. (1976). *The Logic of Questions and Answers.* New Haven: Yale University Press.

Belnap, N. D. (1983). Approaches to the Semantics of Questions in Natural Language. In Bäuerle, R., Schwarze C. y von Stechow, A. (Ed.), *Meaning, Use and Interpretation of Language*. Berlín: Walter de Gruyter.

Bernardi, R., & Kirschner, M. (2010). From Artificial Questions to Real User Interaction Logs: Real Challenges for Interactive Question Answering Systems. In *LREC 2010 Workshop on Web Logs and Question Answering (WLQA 2010).* Malta: LREC.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, 284* (5), 34-43.

Bernstein, A., Kaufmann, E., Kaiser, C., & Kiefer, C. (2006). Ginseng: A Guided Input Natural Language Search Engine for Querying Ontologies. In *Jena User Conference*, Bristol, UK, Mayo.

Black, E. et al. (1991). A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. In *Proceedings of the DARPA Speech and Natural Language Workshop* (pp. 306-311). Association for Computational Linguistics.

Bolinger, D. (1978). Yes/No Questions are not Alternative Questions. In H. Hiz (Ed.), *Questions* (pp. 87-106). Dordrecht: Foris.

Bosque, I. (1980). *Sobre la negación*. Madrid: Cátedra.

Bosque, I. (1984). La selección de las palabras interrogativas. *Verba,* (XI), 245-273.

Bouma, G., Mur, J., Noord, G. V., & Groningen, R. (2005). Reasoning over Dependency Relations for QA. In *Knowledge and Reasoning for Answering Questions (KRAQ'05), IJCAI Workshop* (pp. 15-21).

Bouma, G. (2006). Linguistic Knowledge and Question Answering. *Knowledge Creation Diffusion Utilization, 46*(3), 2-3.

Buchholz, S., & Daelemans, W. (2001). Complex Answers: A Case Study using a WWW Question Answering System. *Journal for Natural Language Engineering, 7*(4), 301-323.

Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S., Israel, D., Jacquemin, C., Lin, C.-Y., Maiorano, S., Miller, G., Moldovan, D., Ogden, B., Prager, J., Riloff, E., Singhal, A., Shrihari, R., Strzalkowski, T., Voorhees, E. & Weischedel, R. (2001). Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). NIST. Disponible en: http://www-nlpir.nist.gov/projects/duc/roadmapping.html

Cao, Y., Liu, F., Simpson, P., Antieau, L. a. B.,Andrew, Cimino, J. J., Ely, J., & Yu, H. (2011). AskHERMES: An Online Question Answering System for Complex Clinical Questions. *Journal of Biomedical Informatics, 44*(2), 277-288.

Carroll, J. et al. (1996). *Sparkle Work Package 1: Specification of Phrasal Parsing.* Disponible en: http://www.ilc.pi.cnr.it/sparkle/wp1-prefinal/ wp1- prefinal.html

Carroll, J., Briscoe, T., & Sanfilippo, A. (1998). Parser Evaluation: A Survey and a New Proposal. In *Proceedings, First International Conference on Language Resources and Evaluation* (pp. 447-454).

Carroll, J., Minnen, G., & Briscoe, T. (1999). Corpus Annotation for Parser Evaluation. In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora (LINC)* (pp. 35-41).

Carroll, J. (2005). The Oxford Handbook of Computational Linguistics. In R. Mitkov (Ed.), (pp. 233-248). OUP Oxford.

Carvalho, G., de Matos, D. M. & Rocio, V. (2010). Improving IdSay: A Characterization of Strengths and Weaknesses in Question Answering Systems for Portuguese. In T. A. S. Pardo, A. Branco, A. Klautau, R. Vieira & V. L. S. de Lima (Eds.), *PROPOR* (pp. 1-10). Springer.

Chalendar, G. D. et al. (2002). The Question Answering System QALC at LIMSI: Experiments in using Web and WordNet. In *Proceedings of the Eleventh Text REtrieval Conference (TREC)* (pp. 407-416). National Institute of Standards and Technology (NIST).

Chang, N., Narayanan, S., & Petruck, M. R. L. (2002). Putting Frames in Perspective. In *Proceedings of the 19th International Conference on Computational Linguistics* (pp. 1-7).

Cimiano, P., & Haase, P. (2007). Porting Natural Language Interfaces between Domains -- an Experimental User Study with the ORAKEL System. *Human Factors,* 180.

Clarke, C. L. A., Cormack, G. V., Kemkes, G., Laszlo, M., Lynam, T. R., Terra, E. L., & Tilker, P. L. (2002). Statistical Selection of Exact Answers (MultiText Experiments for TREC 2002). In *Proceedings of the TREC 2002 Conference* (pp. 823-831). National Institute of Standards and Technology (NIST).

Clarke, C. L. A., Cormack, G. V., Lynam, T.R., & Terra, E. L. (2006). Question Answering by Passage Selection. In T. Strzalkowski, & S. Harabagiu (Eds.), *Advances in Open Domain Question Answering.* Netherlands: Springer.

Clifton, T., & Teahan, W. (2005). Bangor at TREC 2004: Question Answering Track. In *Proceedings of the 13th TREC* (p. 782). National Institute of Standards and Technology (NIST).

Contreras, H. (1999). Relaciones entre las construcciones interrogativas, exclamativas y relativas. In I. Bosque, & V. Demonte (Eds.), *Gramática descriptiva de la lengua española* (pp. 1931-1963). Madrid: Espasa Calpe.

Cui, H., Sun, R., Li, K., Kan, M., & Chua, T. (2005). Question Answering Passage Retrieval using Dependency Relations. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR 05* (pp. 400-407). ACM Press.

Díaz, D. T. (2009). *Sistemas de clasificación de preguntas basados en corpus para la búsqueda de respuestas*. Tesis doctoral, Universitat d'Alacant.

Damljanovic, D., Agatonovic, M., & Cunningham, H. (2011). FREyA: An Interactive Way of Querying Linked Data using Natural Language. In *Proceedings of 1st Workshop on Question Answering Over Linked Data QALD1 Collocated with the 8th Extended Semantic Web Conference ESWC* (pp. 10-23).

Dang, H. T., Lin, J. J. & Kelly, D. (2006). Overview of the TREC 2006 Question Answering Track 99. In E. M. Voorhees & L. P. Buckland (Eds.), *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006*. National Institute of Standards and Technology (NIST).

De Boni, M. (2004). *Relevance in Open Domain Question Answering: Theoretical Framework and Applications*. Tesis doctoral, Department of Computer Science, University of York.

Delpech, E., & Saint-Dizier, P. (2008). Investigating the Structure of Procedural Texts for Answering How-to Questions. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (2008)* (pp. 46-51).

Diller, A. M. (1984). *La pragmatique des questions et des réponses*. Tubinga: Gunter Narr.

Drubig, H. B. (2003). Towards a Typology of Focus and Focus Constructions. *Linguistics, 41*(1), 1-50.

Dumitrescu, D. (1992). Preguntas con multiconstituyentes interrogativos en español. *Hispania, 1*(75), 164-170.

Engelmore, R. et al. (1993). *Knowledge-based Systems in Japan.* International Technology Research Institute.

Erbach, G. (2004). Evaluating Human Question Answering Performance under Time Constraints. Disponible en: http://www.mcgreg.net/pub/human-qa/EvaluatingHumanQA.pdf

Escandell, M. V. (1988). *La interrogación en español: Semántica y pragmática.* Madrid: Editorial de la Universidad Complutense.

Escandell, M. V. (1993). *Introducción a la pragmática*. Barcelona: Ariel.

Escandell, M. V. (1999). Los enunciados interrogativos. Aspectos semánticos y pragmáticos. In I. Bosque y V. Demonte (Eds.), *Gramática Descriptiva de la Lengua Española* (pp. 3929-3991). Madrid: Real Academia Española.

Fava, E. (1995). Tipi di frasi principali. Il tipo interrogativo. In L. Renzi, G. Salvi & A. Cardinaletti (Eds.), *Grande grammatica italiana di consultazione (vol. III)* (pp. 70-127). Bolonia: Il Mulino.

Fazzinga, B., & Lukasiewicz, T. (2010). Semantic Search on the Web. *Semantic Web, 1*(1-2), 89-96.

Fellbaum, C. (Ed.) (1998). *Wordnet, an Electronic Lexical Database*. MIT Press.

Fernández Ramírez, S. (1951). *Gramática española, 2. Los sonidos.* Madrid: Arco Libros.

Ferrández, O., Izquierdo, R., Ferrández, S., & Vicedo, J. L. (2009). Addressing Ontology-based Question Answering with Collections of User Queries. In *Information Processing & Management, 45*(2), 175-188.

Ferrucci, D. et al. (2009). Towards the Open Advancement of Question Answering Systems. *Science,* 24789.

Ferrucci, D. et al. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine, 31*(3), 59-79.

Fleischman, M., Hovy, E. & Echihabi, A. (2003). Offline Strategies for Online Question Answering: Answering Questions before they are asked. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* (pp. 1-7). Morristown, NJ, USA: Association for Computational Linguistics.

Fliedner, G. (2007). *Linguistically Informed Question Answering*. Tesis doctoral, German Research Center for Artificial Intelligence, Saarland University.

Foster, J. (2010). Cba to check the Spelling Investigating Parser Performance on Discussion Forum Posts. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics,* Los Angeles, California (pp. 381-384).

Freed, A. F. (1994). The Form and Function of Questions in Informal Dyadic Conversation. *Journal of Pragmatics,* (21), 621-644.

Freitas, A., De Oliveira, J.G., O'Riain, S., Curry, E., & Pereira Da Silva, J.C. (2011). Treo: Combining Entity-Search, Spreading Activation and Semantic Relatedness for Querying Linked Data. In *Proceedings of the 1st Workshop on Question Answering Over Linked Data* (QALD-1).

Gaizauskas, R., Hepple, M. & Huyck, C. (1998). A Scheme for Comparative Evaluation of Diverse Parsing Systems. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain (pp. 143-149).

Gaizauskas, R., & Wilks, Y. (1998). Information Extraction: Beyond Document Retrieval. *Journal of Documentation, 54*(1), 70-105.

García-Miguel, J. M., Vaamonde, G. & Domínguez, F. G. (2010). ADESSE, a Database with Syntactic and Semantic Annotation of a Corpus of Spanish. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (Eds.), *LREC*. European Language Resources Association.

García Riverón, R. (1996). *Aspectos de la entonación hispánica (I: Metodología; II análisis acústico de muestras del español de Cuba)*. Cáceres: Universidad de Extremadura.

Gayo, I. (2010). *Una propuesta de formalización para cláusulas interrogativas directas parciales basada en datos de corpus*. Trabajo de Investigación Tutelado, Universidade de Santiago de Compostela.

Gayo, I. (2011a). Análisis de preguntas para Búsqueda de respuestas: Evaluación de tres parsers del español. In *Actas Del XXVII Congreso de La Sociedad Española para el Procesamiento del Lenguaje Natural* (pp. 419-426). Huelva: SEPLN.

Gayo, I. (2011b). Question parsing for QA in Spanish. In *Proceedings of the Student Research Workshop Associated with the 8th International Conference on Recent Advances in Natural Language Processing* (pp. 73-78). Hissar, Bulgaria: RANLP.

Gildea, D. (2001). Corpus Variation and Parser Performance. In L. Lee & D. Harman (Eds.), *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing* (pp. 167-202). Stroudsburg: Association for Computational Linguistics.

Ginzburg, J. (1995). Resolving questions, II. *Linguistics and Philosophy, 6*(18), 567-609.

Ginzburg, J. (1996). Interrogatives: Questions, Facts and Dialogue. In *The Handbook of Contemporary Semantic Theory* (pp. 385-422). Cambridge: Blackwell.

Glinos, D. G., & Gomez, F. (2006). Syntax-based Concept Extraction for Question Answering using SEMEX. *Artificial Intelligence,* (1), 789-790.

Gochet, P. (2011). Possible World Semantics. In M. Sbisà J. O. Östman & J. Verschueren (Eds.), *Philosophical perspectives for pragmatics* (pp. 244-252). John Benjamins Publishing Company.

Goodall, G. (2004). On the Syntax and Processing of WH-questions in Spanish. In *Proceedings of the 23rd West Coast Conference on Formal Linguistics* (pp. 101-114). California: Davis.

Goody, E. N. (1978). *Questions and politeness: Strategies in Social Interaction.* Cambridge University Press.

Green, B. F., & Laboratory, L. (1961). *Baseball: An Automatic Question-Answerer.* Massachusetts Institute of Technology, Lincoln Laboratory.

Groenendijk, J. & Stokhof, M. (1984). On the Semantics of Questions and the Pragmatics of Answers. In *Varieties of Formal Semantics. Proceedings of the fourth Amsterdam Colloquium, September 1982* (pp. 143-170). Dordrecht: Foris.

Groenendijk, J. & Stokhof, M. (1989). Type-Shifting Rules and the Semantics of Interrogatives. In G. Chierchia, B. Partee & R. Turner (Eds.). In *Properties, Types and Meanings. Vol. 2: Semantic Issues* (pp. 21-68). Dordrecht: Kluwer.

Groenendijk, J. & Stokhof, M. (1992). A Note on Interrogatives and Adverbs of Quantification. In C. Barker & D. Dowty (Eds.), *ALT II: Proceedings of the Second Conference on Semantics and Linguistic Theory, 1992* (pp. 99-124). Columbus: The Ohio State University.

Groenendijk, J. & Stokhof, M. (1997). Questions. *Handbook of Logic and Language* (pp. 1055-1124). Amsterdam: Elsevier Science.

Gunning, D. et al. (2010). Project Halo Update --- Progress toward Digital Aristotle. *AI Magazine, 31*(3), 33-58.

Hamblin, C. L. (1958). Questions. *Australasian Journal of Philosophy, 3*(36), 159-168.

Hamblin, C. (1973). Questions in Montague English. *Foundations of Language,* (10), 41-53.

Harabagiu, S. M., Moldovan, D. I., Pasca, M., Mihalcea, R., Surdeanu, M., Bunescu, R. C., Girju, R., Rus, V. & Morarescu, P. (2000). FALCON: Boosting Knowledge for Answer Engines. In *Proceedings of the Ninth Text REtrieval Conference (TREC 9)* (pp. 479-488). National Institute of Standards and Technology (NIST).

Harabagiu, S. M., Bunescu, R. C. & Maiorano, S. J. (2001). Text and Knowledge Mining for Coreference Resolution. In *NAACL* (pp. 55-62). Morristown, NJ: Association for Computational Linguistics.

Harabagiu, S. M., Moldovan, D. I., Clark, C., Bowden, M., Hickl, A., & Wang, P. (2005). Employing two Question Answering Systems in TREC 2005. In *TREC, Special Publication* (pp. 500-266). National Institute of Standards and Technology (NIST).

Harabagiu, S. (2006). Questions and Intentions. In T. Strzalkowski & S. Harabagiu (Eds.), *Advances in Open Domain Question Answering* (pp. 99-147). Netherlands: Springer.

Hartrumpf, S. (2008). Semantic Decomposition for Question Answering. In M. Ghallab, C. D. Spyropoulos, N. Fakotakis & N. M. Avouris (Eds.), *ECAI* (pp. 313-317). IOS Press.

Hartrumpf, S., Glöckner, I. & Leveling, J. (2008). Efficient Question Answering with Question Decomposition and Multiple Answer Streams. In C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. J. F. Jones, M. Kurimo, T. Mandl, A. Peñas & V. Petras (Eds.), *CLEF* (pp. 421-428). Springer.

Hausser, R. & D. Zaefferer (1979). Questions and Answers in a Context-Dependent Montague Grammar. In Guenthner, F. & Schmidt, S.J. (Ed.), *Formal Semantics and Pragmatics for Natural Languages* (pp. 339-358). Dordrecht: Reidel.

Hendler, J. (2010). Web 3.0: The Dawn of Semantic Search. *Computer, 43*(1), 77-80.

Hermjakob, U. (2001). Parsing and Question Classification for Question Answering. In *Proceedings of the ACL Workshop on Open-Domain Question Answering* (pp. 1-6). Association for Computational Linguistics.

Hickl, A., Wang, P., Lehmann, J., & Harabagiu, S. (2006). FERRET: Interactive Question-Answering for Real-World Environments. *Computational Linguistics,* (July), 25-28.

Hirschman, L., & Gaizauskas, R. (2001). Natural Language Question Answering: The View from here. *Natural Language Engineering, 7*(4), 275-300.

Hong, M. (1995). *The Semantics and Pragmatics of Questions and Alternatives.* Tesis doctoral, University of Texas.

Hovy, E., Gerber, L., Hermjakob, U., Junk, M. & Lin, C. (2000). Question Answering in Webclopedia. In *Proceedings of the 9th Text Retrieval Conference (TREC).* National Institute of Standards and Technology (NIST).

Hovy, E., Gerber, L., Hermjakob, U., Lin, C.-Y. & Ravichandran, D. (2001). Toward Semantics-based Answer Pinpointing. In *Proceedings of the First International Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '01*, March 18 - 21, 2001, San Diego, CA, USA (pp. 1—7). Morristown, NJ, US: Association for Computational Linguistics.

Hovy, E., Hermjakob, U., & Lin, C. (2001). The Use of External Knowledge in Factoid QA. In *Proceedings of the Tenth Text REtrieval Conference, TREC* (pp. 644-652). National Institute of Standards and Technology (NIST).

Hovy, E., Hermjakob, U., & Ravichandran, D. (2002). A Question/Answer Typology with Surface Text Patterns. In *Proceedings of the Second International Conference on Human Language Technology Research (LREC)* (pp. 247-251). San Diego, California: LREC.

Huddleston, R. (1994). The Contrast between Interrogatives and Questions. *Journal of Linguistics,* (30), 411-439.

Hudson, R. A. (1975). The Meaning of Questions. *Language, 51*(1), 1-31.

Ittycheriah, A., Franz, M., Zhu, W.-J., Ratnaparkhi, A., & Mammone, R. J. (2002). IBM's Statistical Question Answering System-TREC-11. In *Proceedings of the Eleventh Text Retrieval Conference (TREC 2002)*. National Institute of Standards and Technology (NIST).

Jacques, F. (1981). L'interrogation: Force illocutoire et interaction verbale. *Langue Française,* (52), 70-79.

Kaisser, M. & Becker, T. (2004). Question Answering by Searching Large Corpora with Linguistic Methods. In *Proceedings of the 13th Text REtrieval Conference (TREC)* (pp. 500-261). National Institute of Standards and Technology (NIST).

Karttunen, L. (1977). Syntax and Semantics of Questions. *Linguistics and Philosophy,* (1), 3-44.

Katz, B. & Lin, J. (2003). Selectively using Relations to Improve Precision in Question Answering. In *Proceedings of the EACL Workshop on Natural Language Processing for Question Answering* (pp. 43-50).

Katz, B., Lin, J. J., Loreto, D., Hildebrandt, W., Bilotti, M. W., Felshin, S., Fernandes, A., Marton, G. & Mora, F. (2003). Integrating Web-based and Corpus-based Techniques for Question Answering. *Artificial Intelligence, 405*(November), 296.

Katz, B., Bilotti, M. W., Felshin, S., Fernandes, A., Hildebrandt, W., Katzir, R., Lin, J. J., Loreto, D., Marton, G., Mora, F. & Uzuner, Ö. (2004). Answering Multiple Questions on a Topic from Heterogeneous Resources. In E. M. Voorhees & L. P. Buckland (Eds.), *Proceedings of the 13th Annual Text REtrieval Conference (TREC 2004).* National Institute of Standards and Technology (NIST).

Katz, B., Borchardt, G., & Felshin, S. (2005). Syntactic and Semantic Decomposition Strategies for Question Answering from Multiple Resources. In *Proceedings of the AAAI 2005 Workshop on Inference for Textual Question Answering,* July 2005, Pittsburgh, PA, (pp. 35-41).

Kaufmann, E., Bernstein, A., & Zumstein, R. (2006). Querix: A Natural Language Interface to Query Ontologies based on Clarification Dialogs. *Machine Learning,* (November), 5-6.

Kaufmann, E., Bernstein, A., & Fischer, L. (2007). NLP-reduce: A 'naïve' but Domain-independent Natural Language Interface for Querying Ontologies. In *4th European Semantic Web Conference ESWC 2007* (pp. 1-2).

Kaufmann, S. (2009). *Questions and Inquisitive Semantics.* No publicado.

Kawahara, D., Kaji, N., & Kurohashi, S. (2002). Question and Answering System based on Predicate-argument Matching. In *Proceedings of NTCIR-3.*

Keenan, E. L. & Hull, R. D. (1973). The Logical Presuppositions of Questions and Answers. In J. S. Petöfi & D. Franchk (Eds.), *Präsuppositionen in philosophie und linguistik* (pp. 441-466). Frankfurt am Main: Athenäum.

Kelly, D., Kantor, P. B., Morse, E. L., Scholtz, J., & Sun, Y. (2006). User-centered Evaluation of Interactive Question Answering Systems. *Computational Linguistics,* (June), 49-56.

Kiefer, F. (1980). Yes-no Questions as WH-questions. In J. R. Searle, F. Kiefer & M. Bierwisch (Eds.), *Speech acts theory and pragmatics*. Dordrecht: Reidel.

Klein, D., & Manning, C. D. (2003). Fast Exact Inference with a Factored Model for Natural Language Parsing. *Science, 15*(2002), 3-10.

Koster, C.H.A., Seutter, M., & Seibert, O. (2007). Parsing the Medline Corpus. In *Proceedings RANLP 2007* (pp. 325-329).

Koster, C.H.A., Seutter, M. & Seibert, O. (2008). *Manual for the AGFL system version 2.8*. Nijmegen.

Koster, C. H. A. (2011). Text mining for IP DUPIRA an Aboutness-based Dependency Parser for Dutch. No publicado.

Krifka, M. (2001). For a Structured Meaning Account of Questions and Answers, (revised version). In C. Féry and W. Sternefeld (Eds.), *Audiatur* (pp. 287-319). Berlin: Akademie.

Kupiec, J. (1993). MURAX: A Robust Linguistic Approach for Question Answering using an On-line Encyclopedia. In *Proceedings of SIGIR 1993* (pp. 181-190). New York: ACM.

Kwok, C., Etzioni, O., & Weld, D. S. (2001). Scaling Question Answering to the Web. *ACM Transactions on Information Systems, 19*(3), 242-262.

Lavenus, K., Grivolla, J., Gillard, L. & Bellot, P. (2004). Question-Answer Matching: Two Complementary Methods. In C. Fluhr, G. Grefenstette & W. B. Croft (Eds.), *RIAO* (pp. 244-259). CID.

Lehnert, W. (1977). Human and Computational Question Answering. *Cognitive Science, 1*(1), 47-73.

Lehnert, W. G. (1978). *The Process of Question Answering.* Lawrence Erlbaum Associates.

Lehnert, W. G. (1980). Question Answering in Natural Language Processing. In L. Bolc (Ed.), *Natural Language Question Answering Systems* (pp. 9-71). Múnich: Carl Hanser Verlag.

Lenat, D. B., Witbrock, M. J., Baxter, D., Blackstone, E., Deaton, C., Schneider, D., Scott, J. & Shepard, B. (2010). Harnessing Cyc to Answer Clinical Researchers' Ad Hoc Queries. *AI Magazine*, 31, 13-32.

Leveling, J. (2010). A Comparative Analysis: QA Evaluation Questions versus Real-world Queries. In *LREC 2010 Workshop on Web Logs and Question Answering (WLQA 2010)* (pp. 16-22). Malta: LREC.

Li, X. & Roth, D. (2006). Learning Question Classifiers: The Role of Semantic Information. *Natural Language Engineering*, 12, 229-249.

Lin, D. (1995). A Dependency-based Method for Evaluating Broad-Coverage Parsers. *IJCAI-95*, Montreal (pp. 1420-1425).

Lin, D. (1998). Dependency-based Evaluation of MINIPAR. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*.

Lin, D., & Pantel, P. (2001). Discovery of Inference Rules for Question Answering. *Natural Language Engineering, 7*(4), 343-360.

Lin, J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., & Karger, D. R. (2003). What makes a Good Answer? : The Role of Context in Question Answering. In *Proceedings of Interact 2003*, September (pp. 25-32).

Linckels, S. & Meinel, C. (2005). A Simple Solution for an Intelligent Librarian System. In N. Guimarães & P. T. Isaías (Eds.), *IADIS AC* (pp. 495-503). IADIS.

Litkowski, K. C. (1999). Question-answering Using Semantic Relation Triples. In E.Voorhees & D. Harman, (Red.), *Proceedings of the Eigth Text Retrieval Conference (TREC8)* (pp. 349–3569). Gaithersburg, Maryland: National Institute of Standards and Technology (NIST).

Litkowski, K. (2004). Senseval-3 Task: Automatic Labeling of Semantic Roles. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text* (pp. 9-12).

López, V., Uren, V., Motta, E., & Pasin, M. (2007). AquaLog: An ontology-driven Question Answering System for Organizational Semantic Intranets. *Web Semantics Science Services and Agents on the World Wide Web, 5*(2), 72-105.

López, V., Sabou, M., Uren, V., & Motta, E. (2009). Cross-ontology Question Answering on the Semantic Web -an Initial Evaluation. In *Proceedings of the Fifth International Conference on Knowledge Capture,* Redondo Beach, California, USA (pp. 17-24).

López, V., Uren, V., Sabou, M., & Motta, E. (2011). Is Question Answering fit for the Semantic Web? : A Survey. *Semantic Web: Interoperability, Usability, Applicability, 2*(2), 125-155.

Lyons, J. (1977). *Semantics*. Londres: Cambrigde University Press.

Manzini, M. R. (1992). *Locality*. Massachusetts: Cambridge.

Matuszek, C., Cabral, J., Witbrock, M., & Deoliveira, J. (2006). An Introduction to the Syntax and Content of Cyc. In *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and its Applications to Knowledge Representation and Question Answering, 3864*(1447) (pp. 44-49).

Maybury, M. T. (2003). Toward a Question Answering Roadmap. *Text,* 3-14.

Maybury, M. T., & Pa, M. (2004). New Directions in Question Answering. In M. T. Maybury (Ed.), *New Directions in Question Answering* (pp. 383-386). Menlo Park, CA: AAAI Press and Cambridge, MA: The MIT Pres.

McClosky, D., Charniak, E. & Johnson, M. (2006). Reranking and Self-Training for Parser Adaptation. In N. Calzolari, C. Cardie & P. Isabelle (Eds.), In *Proceedings of COLING-ACL-06* (pp. 337–344). Sydney: The Association for Computer Linguistics.

Moldovan, D. I., & Rus, V. (2001). Logic Form Transformation of WordNet and its Applicability to Question Answering. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics,* Toulouse, France (pp. 402-409).

Moldovan, D., Harabagiu, S., Girju, R., Morarescu, P., Lacatusu, F., Novischi, A., Badulescu, A. & Bolohan, O. (2002). LCC Tools for Question Answering. In Voorhees & Buckland (Eds.), *Proceedings of the 11th Text REtrieval Conference (TREC-2002)* (pp. 388-397). Gaithersburg: National Institute of Standards and Technology (NIST).

Moldovan, D., Clark, C., Harabagiu, S., & Maiorano, S. (2003). COGEX: A Logic Prover for Question Answering. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1,* Edmonton, Canada (pp. 87-93).

Moldovan, D., Pasca, M., Harabagiu, S., & Surdeanu, M. (2003). Performance Issues and Error Analysis in an Open-domain Question Answering System. *ACM Transactions on Information Systems TOIS, 21*(2), 133-154.

Moldovan, D. I., Clark, C., & Bowden, M. (2007). Lymba's PowerAnswer 4 in TREC 2007. In *Text REtrieval Conference (2007)*. National Institute of Standards and Technology (NIST).

Mollá, D., & Gardiner, M. (2004). Answerfinder - Question Answering by combining Lexical, Syntactic and Semantic Information. *In Australasian Language Technology Workshop (ALTW) 2004* (pp. 9-16).

Moreda, P., Llorens, H., Saquete, E., & Palomar, M. (2011). Combining Semantic Information in Question Answering Systems. *Inf. Process. Manage., 47*(6), 870-885.

Musillo, G., & Sima'an, K. (2002). Towards comparing Parsers from Different Linguistic Frameworks: An Information Theoretic Approach. In *Proceedings of Workshop beyond PARSEVAL- Towards Improved Evaluation Measures for Parsing Systems at the LREC02*, Las Palmas Canary Islands (pp. 44-51).

Narayanan, S., & Harabagiu, S. (2004a). Answering Questions using Advanced Semantics and Probabilistic Inference. In S. Harabagiu, & F. Lacatusu (Eds.), *HLT-NAACL 2004: Workshop on Pragmatics of Question Answering* (pp. 10-16). Boston: Association for Computational Linguistics.

Narayanan, S., & Harabagiu, S. (2004b). Question Answering based on Semantic Structures. In *Proceedings of Coling 2004* (pp. 693-701). Geneva, Switzerland: COLING.

Neumann, G. & Xu, F. (2003). Mining Answers in German Web Pages. In *WI '03: Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence*. Washington, DC, USA: IEEE Computer Society.

Niles, I., & Pease, A. (2001). Towards a Standard Upper Ontology. In C. Welty & B. Smith (Eds.), *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)* (pp. 2-9). ACM.

Niremberg, S. & Raskin, V. (2004). *Ontological Semantics*. Cambridge, Massachusetts, London, England: MIT Press.

Nyberg, E., Durme, B. V., Huang, Y., & Kup, A. (2001). Towards Light Semantic Processing for Question Answering. *Sciences New York,* 54-61.

Oostdijk, N.H.J. (1998). The TOSCA parsing system reviewed. In J. Carroll (Ed.), *Proceedings of the LREC workshop The Evaluation of Parsing Systems* (pp. 63-69). Granada.

Oostdijk, N.H.J. (2003). Corpus Linguistics meets Language Technology: Deep syntactic parsing for question answering. In D. Archer, P. Rayson, A. Wilson & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 Conference. 28 March - 1 April 2003* (pp. 603-610). Lancaster.

Padró, L. et al. (2010). FreeLing 2.1: Five Years of Open-Source Language Processing Tools. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC'10)* (pp. 931–936). Malta: LREC.

Palmer, M., Gildea, D. & Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31, 71-106.

Pasca, M. A. (2003). Open-domain Question Answering from Large Text Collections. *Computational Linguistics, 29*(4), 665-667.

Petrov, S., Chang, P., Ringgaard, M., & Alshawi, H. (2010). Uptraining for Accurate Deterministic Question Parsing. *Computational Linguistics, 18*(4), 705-713.

Pomerantz, J. (2005). A Linguistic Analysis of Question Taxonomies. *Journal of the American Society for Information Science and Technology, 56*(7), 715-728.

Porto Dapena, J. A. (1997). *Relativos e interrogativos*. Madrid: Arco Libros.

Potts, C. (2006). How far can Pragmatic Mechanisms take us? *Theoretical Linguistics,* (32), 307-320.

Pradhan, S. S., Krugler, V., Bethard, S., Ward, W., Jurafsky, D., Martin, J. H., Blair-Goldensohn, S., Schlaikjer, A. H., Filatova, E., Duboué, P. A., Yu, H., Passonneau, R. J., Hatzivassiloglou, V., McKeown, K. & Illouz, G. (2002). Building a Foundation System for Producing Short Answers to Factual Questions. In *Proceedings of Text REtrieval Conference (TREC)*, Gaithersburg, MD, Nov. 19-22, 2002. National Institute of Standards and Technology (NIST).

Pradhan, S., Ward, W., Hacioglu, K., Martin, J. H., & Jurafsky, D. (2004). Shallow Semantic Parsing using Support Vector Machines. *Baseline, 2004*, 233-240.

Prager, J. M., Brown, E., Coden, A. & Radev, D. R. (2000). Question-answering by Predictive Annotation. In *Proceedings of SIGIR 2000*, Athens, Greece (pp. 184-191).

Punyakanok, V., Roth, D., & Yih, S. W. (2004). Mapping Dependencies Trees: An Application to Question Answering. *Text,* 1-10.

Qiu, Y. & Frei, H. P. (1993). Concept based Query Expansion. In *SIGIR '93: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 160-169), New York, NY, USA: ACM.

Quarteroni, S., Moschitti, A., Manandhar, S. & Basili, R. (2007). Advanced Structural Representations for Question Classification and Answer Re-ranking. In G. Amati, C. Carpineto & G. Romano (Eds.), *Advances in Information Retrieval --- Proceedings of the 29th European Conference on Information Retrieval (ECIR 2007), 2-5 April 2007, Rome, Italy* (pp. 234-245). Berlin—Heidelberg: Springer.

Quilis, A. (1985). Entonación dialectal hispánica. *Lea,* (7), 145-190.

Quilis, A. (1993). *Tratado de fonología y fonética españolas*. Madrid: Gredos.

Radev, D., Fan, W., Qi, H., Wu, H., & Grewal, A. (2005). Probabilistic Question Answering on the Web. *Journal of the American Society for Information Science and Technology, 56*(6), 571-583.

Ravichandran, D., & Hovy, E. (2002). Learning Surface Text Patterns for a Question Answering System. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics ACL 02, 2*(July) (pp. 41-47).

Real Academia Española (1973). *Esbozo de una nueva gramática de la lengua española*. Madrid: Espasa-Calpe.

Real Academia Española (2009). In RAE (Ed.), *Nueva gramática de la lengua española*. Madrid: Espasa-Calpe.

Real Academia Española (2011). In RAE (Ed.), *Diccionario de la lengua española* (22ª edición). En: http://www.rae.es/rae.html

Reich, I. (2003). *Frage, antwort und fokus*. Akademie Verlag.

Rinaldi, F., Dowdall, J., Kaljurand, K., Hess, M., & Mollá, D. (2003). Exploiting Paraphrases in a Question Answering System. In *Proceedings of the Second International Workshop on Paraphrasing, 16* (pp. 25-32).

Rizzi, L. (1990). *Relativized Minimality*. Massachusetts: Cambridge.

Rojo, G. (1978). Cláusulas y oraciones. *Verba,* (Anexo 14).

Rojo, G. & T. Jiménez (1989). *Fundamentos del análisis sintáctico funcional.* Santiago de Compostela: Lalia 2.

Rooth, M. (1992). A Theory of Focus Interpretation. *Natural Language Semantics, 1*(1), 75-116.

Roth, D., Cumby, C. M., Li, X., Morie, P., Nagarajan, R., Rizzolo, N., Small, K. & tau Yih, W. (2002). Question-Answering via Enhanced Understanding of Questions. In *Proceedings of The Eleventh Text Retrieval Conference (TREC 2002), NIST Special Publication.* National Institute of Standards and Technology (NIST).

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., & Scheffczyk, J. (2006). FrameNet II: Extended Theory and Practice. *Unpublished Manuscript, 1*, 119.

Ruthven, I., & Lalmas, M. (2003). A Survey on the Use of Relevance Feedback for Information Access Systems. *The Knowledge Engineering Review, 18*(2), 95-145.

Saint-Dizier, P., & Moens, M. (2011). Knowledge and Reasoning for Question Answering: Research Perspectives. *Inf. Process. Manage., 47*(6), 899-906.

Salloum, W. (2009). A Question Answering System based on Conceptual Graph Formalism. In *2009 Second International Symposium on Knowledge Acquisition and Modeling, 3*(3) (pp. 383-386).

Santalla, M. P. (2002). *A Formal Grammar of Spanish for Prhase-level Analysis applied to Information Retrieval*. Santiago: Servizo de Publicacións e Intercambio Científico.

Schank, R. C. (1972). Conceptual Dependency: A Theory of Natural Language Understanding. *Cognitive Psychology, 3*(4), 552-631.

Schank, R. C., Goldman, N., Reiger, C., & Riesbeck, C. K. (1973). MARGIE: Memory, Analysis, Response Generation, and Inference on English. In *Proceedings of the Third International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 255-261). Menlo Park, CA: SRI.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures* Lawrence. Erlbaum Associates.

Schlaefer, N., Ko, J., Betteridge, J., Pathak, M. A., Nyberg, E. & Sautter, G. (2007). Semantic Extensions of the Ephyra QA System for TREC 2007. In E. M. Voorhees & L. P. Buckland (Eds.), *TREC*. National Institute of Standards and Technology (NIST).

Selting, M. (1992). Prosody in Conversational Questions. *Journal of Pragmatics,* 315-345.

Simmons, R. F. (1965). Answering English Questions by Computer: A Survey. *Communications of the ACM, 8*(1), 53-70.

Sosa, J. M. (1991). *Fonética y fonología de la entonación del español hispanoamericano.* Tesis doctoral, Massachussets, University of Massachussets.

Soubbotin, M. M. (2001). Patterns of Potential Answer Expressions as Clues to the Right Answers. *In Proceedings of the Tenth Text REtrieval Conference TREC (pp.* 293-302). National Institute of Standards and Technology (NIST).

Srihari, R. K., Li, W., & Li, X. (2006). Question Answering Supported by Multiple Levels of Information Extraction. In T. Strzalkowski & S. Harabagiu (Eds.), *Advances in Open Domain Question Answering* (pp. 349-383). Netherlands: Springer.

Stahl, G. (1956). La lógica de las preguntas. *Anales de La Universidad De Chile,* (102), 71-75.

Stechow, A. (1991). Focusing and Backgrounding Operators. In W. Abraham (Ed.), *Discourse particles* (pp. 37-84). Amsterdam, Philadelphia: John Benjamins.

Strzalkowski, T. (2006). *The Future: Interactive, Collaborative Information Systems*. No publicado.

Sun, R., Jiang, J., Tan, Y. F., Cui, H., Chua, T.-S. & Kan, M.-Y. (2005). Using Syntactic and Semantic Relation Analysis in Question Answering. In E. M. Voorhees & L. P. Buckland (Eds.), *TREC*. National Institute of Standards and Technology (NIST).

Surdeanu, M., Ciaramita, M., & Zaragoza, H. (2008). Learning to Rank Answers on Large Online QA Collections. *Computational Linguistics,* (June), 719-727.

Sutcliffe, R. F. E, Kruschwitz, U. & Mandl,T. (2010). Web Logs and Question Answering. In *LREC 2010 Workshop on Web Logs and Question Answering (WLQA 2010)*. Malta: LREC.

Tablan, V., Damljanovic, D., & Bontcheva, K. (2008). A Natural Language Query Interface to Structured Information. *Interface, 5021*, 1-15.

Tannier, X., & Moriceau, V. (2009). Studying Syntactic Analysis in a QA System: FIDJI @ Respubliqa'09. In *Proceedings of the 10th Cross-Language Evaluation Forum Conference on Multilingual Information Access Evaluation: Text Retrieval Experiments,* Corfu, Greece, (pp. 237-244).

Tiedemann, J. (2005). Improving Passage Retrieval in Question Answering using NLP. *Lecture Notes in Computer Science, 3808*, 634-646.

Tomuro, N. (2004). Question Terminology and Representation for Question Type Classification. *Terminology, 10*(1), 153-168.

Torrego, E. (1984). On Inversion in Spanish and some of its Effects. *Linguistic Inquiry,* (15), 103-129.

TREC (2004). *TREC 2004 Judgments for TREC 2004 Factoid Questions*. National Institute of Standards and Technology (NIST).

Tunstall-Pedoe, W. (2010). True Knowledge: Open-domain Question Answering using Structured Knowledge and Inference. *AI Magazine, 31*(3), 80-92.

Unger, C. (2011). *Proceedings of 1st Workshop on Question Answering over Linked Data QALD1 Collocated with the 8th Extended Semantic Web Conference ESWC 2011*. Springer.

van Rooy, R. (2003). Questioning to resolve Decision Problems. *Linguistics and Philosophy, 26*(6), 727-763.

Vargas-Vera, M., & Motta, E. (2004). AQUA - Ontology-based Question Answering System. *Micai, 2972*, 468-477.

Verberne, S. (2010). *In search of the Why: Developing a System for Answering Why-Questions.* Tesis doctoral, Radboud University.

Voorhees, E., & Tice, D. M. (2000). The TREC-8 Question Answering Track Evaluation. In *TREC*. National Institute of Standards and Technology (NIST).

Voorhees, E. M. (2004). Overview of the TREC 2001 Question Answering Track. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)* (pp. 42-51). National Institute of Standards and Technology (NIST).

Voorhees, E. M., & Buckland, L. P. (Eds.) (2006). *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006*, Gaithersburg, Maryland, November 14-17, 2006. National Institute of Standards and Technology (NIST).

Wang, C., Xiong, M., Zhou, Q., & Yu, Y. (2007). PANTO: A Portable Natural Language Interface to Ontologies. In *4TH ESWC, INNSBRUCK* (pp. 473-487). Springer-Verlag.

Wonsever, D., Malcuori, M. & Aiala, R. (2012). Procesamiento de lenguaje natural en la universidad de la República. *Revista Digital Universitaria, 13*(5).

Woods, W., Kaplan, R., & Nash-Webber, B. (1972). The Lunar Sciences Natural Language Information System: Final report.

Wunderlich, D. (1981). Questions about Questions. In W. Klein & W. Levelt (Eds.), *Crossing the Boundaries in Linguistics* (pp. 131-158). Dordrecht: Reidel.

Zubizarreta, M. L. (1999). Las funciones informativas: Tema y foco. In I. Bosque, & V. Demonte (Eds.), *Gramática descriptiva de la lengua española* (pp. 4215-4244). Madrid: Espasa.

Apéndice 1

TRIVIAL
llamarse 10
denominar 2

CLEF
llamar 21
morir 5
convertir 2

WIKI

3100 llamar	119 formar	58 poner	36 comportar
1350 ser	118 alimentar	56 mediar	35 surgir
1327 conseguir	116 cambiar	56 leer	35 salir
1236 decir	115 funcionar	56 elaborar	34 producir
997 escribir	113 defender	55 ver	34 organizar
878 abreviar	104 ir	54 desplazar	33 terminar
723 hacer	102 medir	51 quitar	32 vencer
571 respirar	101 poder	50 dar	32 matar
282 reproducir	98 entrar	49 determinar	31 clonar
276 evolucionar	96 colaborar	48 usar	30 resolver
252 nacer	91ser	48 conservar	30 definir
231 clasificar	88 comer	47 borrar	30 afectar
228 sacar	78 influir	46 utilizar	29 evitar
198 tener	78 crear	46 representar	28 dormir
196 cuidar	74 jugar	45 intercambiar	27 registrar
179 capturar	71 incubar	44 originar	27 nombrar
174 llegar	69 subir	44 alterar	27 editar
173 encontrar	69 estudiar	43 estar	27 contribuir
167 pasar	65 realizar	42 aprender	26 dibujar
150 dividir	64 creer	40 saber	26 contar
149 sistemar	63 unir	40 ayudar	25 pronunciar
144 atrapar	63 comunicar	39 descargar	25 nutrir
133 favorecer	62 cortejar	39 abrir	
125 obtener	61 relacionar	38 ganar	
124 vestir	61 elegir	36 trabajar	
120 vivir	59 hablar	36 denominar	

Apéndice 2

Compendio de valores[175]

Valores no argumentales

Temporal

TIME (*cuándo*)

```
# (null) 10 0-42|¿Cuándo ocurrió la batalla de Iwo Jima?
whQ [V: ocurrir third sing past <SUBJ [N: batalla <MODde PN:
Iwo Jima <DET D: la] <QCIRC TIME]
```

CIRC_TIME+ tempNoun (*qué* + *año; en* + *qué/cuál* + *tempNoun*)

```
# (null) 11 0-57|¿En qué año California se convirtió en un
territorio?
whQ [V: convertir third sing past <REF se third GENDER sing
<SUBJ PN: California <QCIRC_TIMEen N: año <PCen [N: territorio
<DET D: un]]
```

```
# (null) 12 0-47|¿En cuál mes del año llueve más en Galicia?
whQ [V: llover third sing present <QCIRC_TIMEen [N: mes <MODde
[N: año <DET el]] <CIRC X: más <CIRCen PN: Galicia]
```

175 Distinguimos dos grupos: valores no argumentales y valores típicamente argumentales. En cada grupo, organizamos los valores por áreas de significado.

Locativo

LOCATION (*dónde*)

(null) 13 0-49|¿**Dónde** se encuentran las mejillas en el cuerpo?
whQ [V: encontrar third plu present <REF se third GENDER plu <SUBJ [N: mejillas <DET D: las] **<QCIRC LOCATION** <CIRCen [N: cuerpo <DET D: el]]

LOCATION +GEO (*dónde*)

(null) 14 0-32|¿**Dónde** nació Abraham Lincoln?
whQ [V: nacer third sing past <SUBJ PN: Abraham Lincoln **<QCIRC LOCATION +GEO**]

CIRC_LOCATION + locNoun (*en/hacia/desde* + *qué/cuál* + *locNoun*)

(null) 15 0-51|¿**En qué ciudad** está la universidad Southwestern?
whQ [V: estar third sing present <SUBJ [N: universidad <MOD PN: Southwestern <DET D: la] **<QCIRC_LOCATIONen N: ciudad**]

(null) 16 0-35|¿**Desde cuál lugar** partió Colón?
whQ [V: partir third sing past <SUBJ PN: Colón **<QCIRC_LOCATIONdesde N: lugar**]

Modo, manera

MANNER (*cómo; de* + *qué/cuál* + *modalNoun*)

```
# (null) 1 0-42|¿Cómo actúa la hormona del crecimiento?
whQ [V: actuar third sing present <SUBJ [N: hormona <MODde [N:
crecimiento <DET el] <DET D: la] <QCIRC MANNER]
```

```
# (null) 2 0-36|¿De qué forma murió Adolf Hitler?
whQ [V: morir third sing past <SUBJ PN: Adolf Hitler <QCIRC
MANNER]
```

```
# (null) 3 0-48| ¿De cuál manera puede ser tratada la alergia?
whQ [[V: ser third sing present <AUXposib V: poder] <PRED
A:tratada <QCIRC MANNER <SUBJ [N: alergia <DET D: la]]
```

Causa

CAUSE (*por qué; por* + *qué/cuál* + *causeNoun*)

```
# (null) 4 0-34|¿Por qué es útil una mariquita?
whQ [V: ser third sing present <SUBJ [N: mariquita <DET D:
una] <PRED A: útil <QCIRC CAUSE]
```

```
# (null) 5 0-49|¿Por qué razón no pueden volar las avestruces?
whQ [[V: volar third plu present <AUXposib V:poder] <NEG X: no
<SUBJ [N: avestruces <DET D: las] <QCIRC CAUSE]
```

```
# (null) 6 0-46| ¿Por cuál motivo se vuelve naranja la luna?
whQ [V: volver third sing present <REF se third GENDER sing
<PRED A: naranja <SUBJ [N: luna <DET D: la] <QCIRC CAUSE]
```

Nombre

NAME (*cómo* + *VerbDenom; de* + *qué/cuál* + *modalNoun* + *VerbDenom*)

```
# (null) 7 0-40|¿Cómo se llamaba el barco de Cousteau?
whQ [V: llamar third sing imperfect <REF se third GENDER sing
<SUBJ [N: barco <MODde PN: Cousteau <DET D: el] <QCIRC NAME]
```

```
# (null) 8 0-66|¿De qué forma se denomina la gente que hace
fuegos artificiales?
whQ [V: denominar third sing present <REF se third GENDER sing
<SUBJ [N: gente >SUBJ [V:hacer third sing present <OBJ [N:
fuegos <ATTR A: artificiales]] <DET D: la] <QCIRC NAME]
```

Valores típicamente argumentales

Entidad

ENTITY (*qué/cuál*)

```
# (null) 9 0-44| ¿Qué causó las inundaciones de Lynmouth?
whQ [V: causar third sing past <OBJ [N: inundaciones <MODde
PN: Lynmouth <DET D: las] <QSUBJ ENTITY]
```

(null) 10 0-44|¿**Cuál** era el juguete más popular en 1957?
whQ [V: ser third sing imperfect <PRED [N: juguete <ATTR [A: popular <QUANT X:más] <DET D: el] **<QSUBJ ENTITY** <DATEen 1957]

(null) 3 0-20|¿Qué vende Iberia?
whQ [V: vender third sing present **<QOBJ ENTITY** <SUBJ PN: Iberia]

ENTITY +PROPERTY (*de* + *cuál* + *ser* + *fn*)

(null) 11 0-77|¿**De cuál de los directores de Hollywood** es *La Guerra de las Galaxias*?
whQ [V: ser third sing present <QPREDde [ENTITY +PROPERTY <MODde [N: directores <MODde PN: Hollywood <DET D: los]] <SUBJ [PN: Guerra de las Galaxias <DET D: la]]

Persona

PERSON (*a* + *cuál* + verbo transitivo)

(null) 12 0-93| ¿**A cuál de los futbolistas del Real Madrid** se ve habitualmente en discotecas de la ciudad?
whQ [V: ver third sing present **<QOBJ [PERSON <MODde [N: futbolistas <MODde [PN: Real Madrid <DET el] <DET D: los]]** <SUBJ IMP <CIRC X: habitualmente <CIRCen [N: discotecas <MODde [N: ciudad <DET D: la]]]

NAME +PERSON (*quién*)

(null) 13 0-81|¿**Quiénes** eran los arquitectos que diseñaron el edificio Empire State Building?
whQ [V: ser third plu imperfect <PRED [N: arquitectos >SUBJ [V: diseñar third plu past <OBJ [[N: edificio <MOD PN: Empire State Building] <DET D: el]] <DET D: los] **<QSUBJ NAME +PERSON plu**]

NAME +PERSON +PROPERTY (*de* + *quién* + *ser*)

(null) 14 0-27|¿**De quién** es La Macarena?
whQ [V: ser third sing present **<QPREDde NAME +PERSON +PROPERTY** <SUBJ PN: La Macarena]

Cantidad

QUANTITY (*cuánto; cómo* + PP *(de +fadj/fadv)*+ *ser*)

(null) 15 0-50|¿**Cuánto** aumenta la población mundial cada año?
whQ [V: aumentar third sing present **<QOBJ QUANTITY** <SUBJ [N: población <ATTR A: mundial <DET D: la] <CIRC [N: año <DET D: cada]]

(null) 16 0-35| ¿**Cómo de alto** es el Golden Gate?
whQ [V: ser third sing present **<QPRED [QUANTITY <MOD de A: alto]** <SUBJ [PN: Golden Gate <DET D: el]]

QUANTITY +MONEY **(***qué* + *VerbQuant; a*+ *cómo***)**

```
# (null) 17 0-39| ¿Qué costó la Ciudad de la Cultura?
whQ [V: costar third sing past <QOBJ QUANTITY +MONEY <SUBJ
[PN: Ciudad de la Cultura <DET D: la]]
```

```
# (null) 18 0-26|¿Cuánto cobra Leo Messi?
whQ [V: cobrar third sing present <QOBJ QUANTITY +MONEY <SUBJ
PN: Leo Messi]
```

```
# (null) 19 0-73|¿A cómo vendieron sus acciones de Caixanova
los principales inversores?
whQ [V: vender third plu past <OBJ [N: inversores <ATTR A:
principales <DET D: los] <SUBJ [N: acciones <MODde PN:
Caixanova <DET D: sus] <QCIRCa QUANTITY +MONEY]
```

QUANTITY +MEASURE (*cuánto* + *VerbQuant*)

```
# (null) 20 0-31|¿Cuánto mide la Torre Eiffel?
whQ [V: medir third sing present <QOBJ QUANTITY +MEASURE <SUBJ
[PN: Torre Eiffel <DET D: la]]
```

QUANTITY +PERSON (*a* + *cuántos*+ *vbo transitivo*)

```
# (null) 1 0-39|¿A cuántos inocentes asesinó Manson?
whQ [V: asesinar third sing past <QOBJ [QUANTITY +PERSON
<MODde N: inocentes] <SUBJ PN: Manson]
```

Descripción

DESCRIPTION (*cómo+ ser; cuál + ser + NE*)

(null) 22 0-51|¿**Cómo** es en argot "un billete de cinco dólares"?
whQ [V: ser third sing present **<QPRED DESCRIPTION** <SUBJ [N: billete <MODde [N: dólares <QUANT Q: cinco] <DET D: un] <CIRCen N: argot]

(null) 23 0-67|¿**Cuál** es George Bush en la famosa foto del "Trío de las Azores"?
whQ [V: ser third sing present **<QPRED DESCRIPTION** <SUBJ PN: George Bush <CIRCen [N: foto <MODde [PN: " Trío de las Azores " <DET el] <ATTR A: famosa <DET D: la]]

DESCRIPTION +PERSON (*quién + ser + NE*)

(null) 24 0-25|¿**Quién** es Henry Butler?
whQ [V: ser third sing present **<QPRED [DESCRIPTION +PERSON]** <SUBJ PN: Henry Butler]

Definición

DEFINITION (*qué + ser*)

(null) 25 0-17|¿**Qué** es Airbus?
whQ [V: ser third sing present **<QPRED DEFINITION** <SUBJ PN: Airbus]

Un sistema de Búsqueda de Respuestas (BR) permite a un usuario realizar una pregunta en lenguaje natural y obtener automáticamente una respuesta correcta y concisa a esa pregunta. La BR es una tarea compleja que implica la puesta en marcha de diversos procesos interdependientes que se estructuran en tres fases o módulos: (1) análisis y comprensión de la pregunta; (2) análisis de la información de la fuente de conocimiento y selección de fragmentos susceptibles de contener la respuesta; (3) selección, extracción y generación de la respuesta.