

# Interfaces de Lenguaje Natural para Consultar Bases de Datos en Español

por Marco Antonio Aguirre Lam y Rodolfo A. Pazos Rangel

## La Importancia de las Interfaces de Lenguaje Natural para Bases de Datos

Desde tiempos remotos, el ser humano ha tenido en mente la idea de imitar los procesos del pensamiento de manera artificial. Con la creación de la computadora electrónica, el sueño del hombre ha sido la posibilidad de establecer comunicación con ésta a través del lenguaje natural y obtener información mediante dicha interacción.

En las últimas décadas la información ha jugado un papel importante en nuestra vida cotidiana, la mayoría de las personas solicitan información antes de tomar una decisión importante. Actualmente, las fuentes más grandes de información se encuentran almacenadas en bases de datos (BDs). Las BDs contienen una colección de datos relacionados entre sí, los cuales son estructurados para modelar la información que se encuentra en el mundo real.

Para que un usuario obtenga la información de una BD, es necesario que se formule una consulta de tal manera que la computadora la interprete y genere la respuesta correcta (usualmente se utiliza un lenguaje de consulta a BDs, tal como SQL, por sus siglas en inglés, Structured Query Language). Desafortunadamente no cualquier usuario es capaz de escribir tales consultas, especialmente aquéllos que carecen de conocimientos computacionales. Únicamente profesionales de la computación pueden formular ese tipo de consultas, lo cual es costoso y tiene limitaciones de tiempo.

La manera normal en la que las personas solicitan información es mediante preguntas en lenguaje natural, pero las computadoras no comprenden directamente este tipo de lenguaje. Debido a esto surgen las Interfaces de Lenguaje Natural para Bases de Datos (ILNBDs), las cuales proveen una alternativa de solución a los problemas que ocurren en sistemas para obtener información, ya que permiten a los usuarios acceder a la información almacenada en BDs a través de una consulta en lenguaje natural (ver Figura 1).

Existen otro tipo de sistemas que suelen ser confundidos con las ILNBDs y son conocidos como “Sistemas de Pregunta-Respuesta” (Q&A, por sus siglas en inglés, Question and Answering). A diferencia de éstos, las ILNBDs, sólo pueden devolver un resultado, el cual se supone que tiene que ser el correcto. Los sistemas Q&A retornan un conjunto de resultados (regularmente elegi-

dos mediante reconocimiento de palabras clave), de los cuales el usuario necesita elegir el que considera correcto. Ejemplos de este tipo de sistemas son los buscadores en Internet, tales como Google, Yahoo, etc.

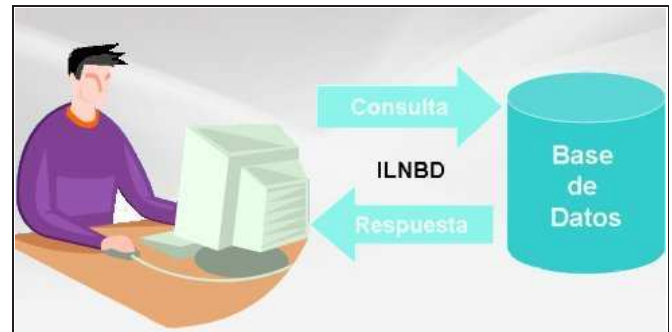


Figura 1. Flujo en una ILNBD: consulta en LN y resultado de la consulta

Los principales beneficiarios de las ILNBDs son de dos tipos: los ejecutivos de empresas que necesitan formular consultas a las BDs de sus empresas de maneras que no están previstas en los sistemas de información corporativos, y usuarios de Internet que desearían consultar BDs con más flexibilidad que la disponible actualmente en las páginas de Web. Para estos usuarios, las consultas mal comprendidas por las ILNBDs pueden ocasionar que éstas entreguen a los usuarios resultados incorrectos o confusos, lo cual es potencialmente peligroso, ya que el usuario puede depender de estos resultados para tomar decisiones importantes. De aquí la importancia de desarrollar ILNBDs que tengan una alta confiabilidad.

La necesidad de ILNBDs se ha incrementado últimamente, debido a que, con el crecimiento de las tecnologías, muchos usuarios solicitan acceder a la información (contenida en BDs) desde diferentes tipos de sistemas, tales como: computadoras, celulares, PDAs (por sus siglas en inglés, Personal Digital Assistants), etc. Incluso, el magnate de la computación, Bill Gates, afirmó: “las interfaces de lenguaje natural dominarán el desarrollo de las computadoras personales en la nueva década” [2][3].

## Limitaciones

Las interfaces de lenguaje natural para acceder a BDs han sido utilizadas desde finales de los 60s y principios de los 70s. Aunque se han desarrollado una gran canti-

dad de ILNBDs desde entonces, las limitaciones y deficiencias que poseen son de llamar la atención. Algunas ILNBDs comerciales ilustran esta situación; por ejemplo, LanguageAccess fue descontinuado por IBM; English Query (desarrollado por Microsoft) fue incluido por última vez en SQL Server ver. 8.0, liberado en el año

2000; English Wizard (desarrollado por Linguistic Technology Corporation) fue descontinuado hace algunos años; y finalmente, uno de los más exitosos, Access English Language Front End (ELF) desarrollado por ELF Software Co., aunque continúa su distribución la compañía ya no existe.

## El desempeño de una ILNBD esta ligado con la configuración de dominio, entre mejor configurado se encuentre, mayor es el entendimiento de las consultas por parte de la interfaz.

La mayoría de los usuarios de ILNBDs consideran que los sistemas existentes son bastante robustos y precisos, pero esta evaluación es relativa; las ILNBDs que han obtenido buenos resultados (alrededor del 95 % de consultas correctamente traducidas) son de dominio específico —es decir, aquéllas que únicamente pueden consultar una sola BD— mientras que las independientes de dominio —es decir, aquéllas que pueden usarse para consultar diferentes BDs— continúan teniendo deficiencias en el proceso de traducción (con una tasa de éxito entre el 80-90 %).

Con respecto a la independencia de dominio, un factor que ha limitado el uso de las ILNBDs ha sido la dificultad para configurar dichos sistemas a una BD en particular, a menudo los administradores de BDs necesitan altos niveles de experiencia y llevan a cabo esta tarea empleando un largo tiempo de configuración [5].

Las ILNBDs que ofrecen independencia de dominio necesitan un proceso inicial de configuración antes de que los usuarios finales puedan introducir sus consultas. La configuración consiste en proporcionar al sistema las palabras y conceptos necesarios para el dominio (llamado también diccionario de datos o de dominio), el cual está relacionado con la información almacenada en la BD.

Aunque algunas ILNBDs cuentan con mecanismos sencillos de configuración, desafortunadamente las configuraciones iniciales solamente tienen un buen desempeño con BDs pequeñas y su desempeño es insatisfactorio para BDs más grandes. En muchos casos se puede modificar la configuración inicial para mejorar el desempeño, pero esta tarea suele ser compleja, tediosa y requiere una profunda comprensión del funcionamiento de la interfaz.

Algunos de los principales problemas encontrados en el proceso de traducción están relacionados con:

1. el uso de palabras o sintagmas de diferentes categorías sintácticas para referirse a tablas y columnas de la BD (tales como sustantivos, verbos, adjetivos, y preposiciones),
2. la elipsis semántica, que ocurre cuando se omiten

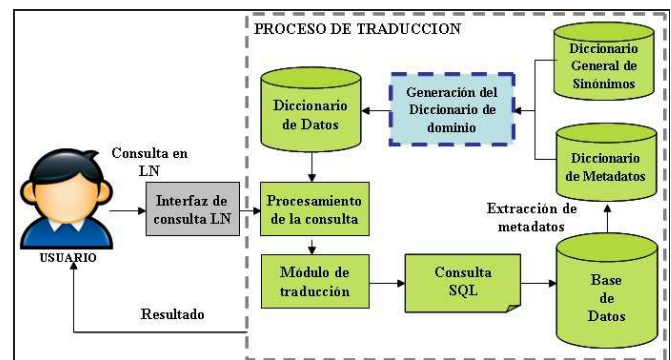
palabras que son necesarias para el claro entendimiento de la consulta,

3. la capacidad de cobertura de SQL, como el uso de varias tablas de la BD y el uso de funciones de agregación, y
4. otro tipo de problemas que involucran errores humanos, tales como, información no existente en la BD, palabras que indican valores imprecisos, etc.

La investigación sobre ILNBDs continúa siendo un problema de investigación abierto por el simple hecho de que ninguna interfaz ha logrado realizar el proceso de traducción totalmente correcto.

### Antecedentes

Los primeros sistemas de consulta en lenguaje natural fueron desarrollados en los 60s, y principalmente eran interfaces de lenguaje natural para sistemas expertos, diseñados para funcionar en dominios específicos. Sólo podían ser usados en una BD particular, por lo cual era difícil modificarlos para consultar diferentes BDs.



**Figura 2. Arquitectura de la primera versión de la ILNBD desarrollada**

El auge del desarrollo de ILNBDs se produjo en los años 80s. A pesar de la implementación de muchos sistemas en ese entonces, poco a poco fue disminuyendo su uso debido a varios problemas que afectan su desempeño.

Existen muchas ILNBDs desarrolladas en las últimas dos décadas, la mayoría son para idiomas extranjeros, sobre todo inglés. Para conocer más detalles acerca del estado del arte de ILNBDs, se recomienda consultar [6].

En idioma español existen muy pocos trabajos en ILNBDs, por lo cual nos encontramos desarrollando una ILNBD que forma parte de un proyecto denominado “Interfaz de Lenguaje Natural Español hacia Bases de Datos para Usuarios de Internet”, el cual comenzó a desarrollarse en el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) desde el año 2001 y se ha continuado en el Instituto Tecnológico de Ciudad Madero (ITCM) desde el año 2002 hasta la fecha. Dicho proyecto tiene como fin implementar una interfaz que permita a usuarios casuales e inexpertos consultar BDs mediante expresiones en lenguaje español no acotado.

Una característica importante de esta ILNBD es su independencia de dominio; es decir, su capacidad para interactuar con BDs de diferentes dominios. En la siguiente sección se describen las aportaciones que se han realizado en el desarrollo de nuestra ILNBD.

## Aportaciones

Como se mencionó anteriormente, este trabajo forma parte de un proyecto iniciado en el año 2001, desde entonces han sido implementadas dos versiones de la ILNBD, y se encuentra en proceso el desarrollo de una nueva versión. A continuación se mencionarán las contribuciones realizadas.

La primera versión de la interfaz fue desarrollada por CENIDET-ITCM [7]. Utiliza un método para la traducción de consultas en español a SQL, lo cual permite a los usuarios realizar consultas a una BD sin necesidad de configuraciones tediosas. El manejo de la independencia del dominio es la característica principal de esta interfaz, y para lograrlo utiliza un preprocesador que genera automáticamente un diccionario de dominio y una técnica de traducción que utiliza sustantivos (que hacen referencia a tablas o columnas de la BD), y preposiciones y conjunciones, ya que estas últimas mantienen el significado en cualquier contexto y no necesitan ser configuradas para un dominio particular.

En dicho trabajo se construye el diccionario de dominio (o diccionario de datos) para la interfaz de forma automática a partir de: 1) un diccionario de sinónimos (que sirve al etiquetador gramatical para relacionar sinónimos de las palabras de la consulta), 2) un etiquetador gramatical (que etiqueta las palabras de la consulta), y 3) un diccionario de metadatos (el cual contiene información estructural de la BD).

El módulo de traducción consta de tres fases:

1. La separación de las frases Select y Where, que consiste en la división de la consulta en las partes que correspondan a los requerimientos y a las con-

diciones de búsqueda. Las preposiciones y conjunciones son representadas como operaciones usando la teoría de conjuntos.

2. La identificación de tablas y columnas, en la cual se pueden relacionar las tablas y las columnas con los sustantivos que se encuentran en la consulta.
3. La construcción del grafo relacional, en donde se establecen las relaciones entre las tablas que fueron mencionadas en la consulta.

Después de este procesamiento, se realiza la conversión de la información encontrada a una consulta SQL, para ser ejecutada y enviar los resultados al usuario. En la Figura 2, se muestra la arquitectura simplificada de la primera versión de la ILNBD desarrollada.

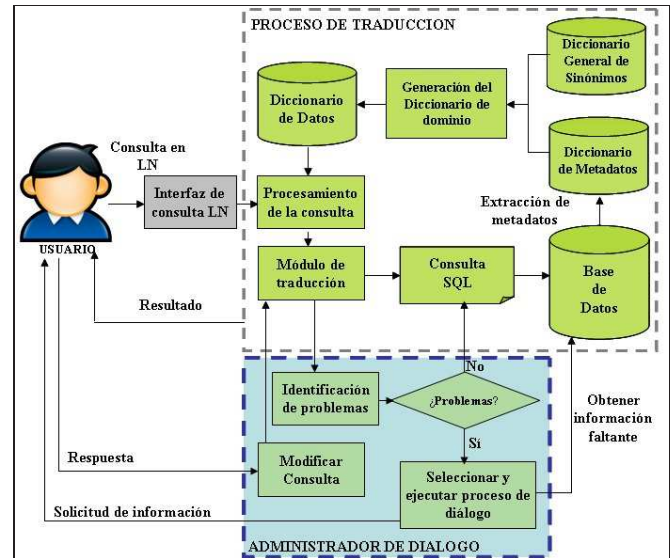


Figura 3. Arquitectura de la ILNBD con administrador de diálogo

A pesar del buen desempeño que tiene esta interfaz, no permite al usuario un diálogo de aclaración cuando la consulta no puede ser traducida (es decir, si la consulta tiene problemas de elipsis semántica), lo cual limita su capacidad para obtener un mayor porcentaje de aciertos en las consultas traducidas.

Posteriormente se continuó con la mejora de este trabajo en [8], en el cual la propuesta principal fue implementar procesos de diálogo independientes del dominio para un administrador de diálogo para la ILNBD. Para lograr que los diálogos tuvieran las características anteriormente mencionadas, se formalizó una tipificación de problemas en consultas que permite abarcar la mayoría de los casos, de esta manera fueron implementados los procesos de diálogo correspondientes a los problemas tipificados.

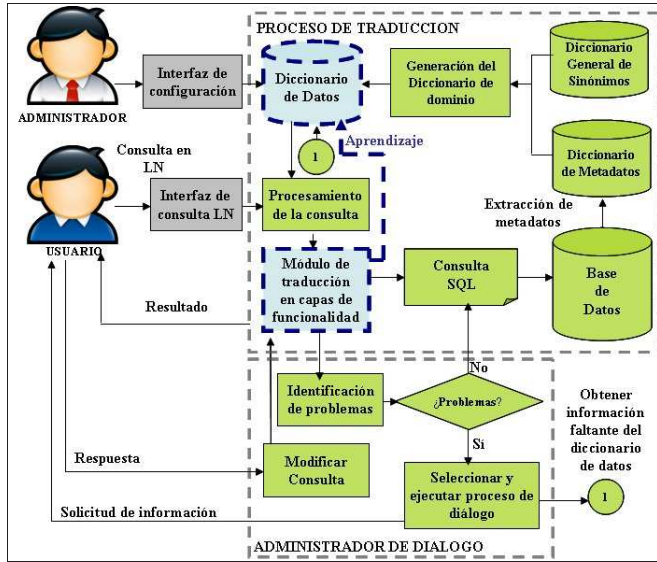


Figura 4. Arquitectura propuesta de la nueva ILNBD

En esta versión de la ILNBD, el proceso de traducción se comunica con el administrador de diálogo, el cual identifica el tipo de problema que tiene la consulta, basándose en la tipificación de problemas descrita en [8]. Si existe algún problema identificado, el administrador selecciona y ejecuta un proceso de diálogo que solicita información al usuario. Con la información recibida, modifica los datos de la consulta y la envía nuevamente al inicio del proceso de traducción. En caso de no existir ningún problema con la consulta, el proceso continúa de manera normal, y en caso contrario se inicia un nuevo proceso de diálogo. En la Figura 3, se muestra la arquitectura de la ILNBD con administrador de diálogo.

La tipificación realizada en esta versión abarca problemas de elipsis semántica; sin embargo, al realizar otro análisis en diversos corpus de consultas, además de encontrar que la mayoría de los problemas son de elipsis, fueron encontrados otros tipos de problemas, tales como el uso de palabras o sintagmas de diferentes categorías sintácticas para referirse a las tablas o columnas de la BD, la capacidad de cobertura de SQL, valores imprecisos (que son palabras que describen valores que pertenecen a un rango como tarde, pesado, etc.), entre otros, los cuales se describen en [9].

Debido a lo anterior, actualmente nos encontramos desarrollando una nueva versión de la ILNBD. En esta versión se contempla la implementación de un nuevo diccionario de datos (a partir de un modelo de BD concebido por nosotros), el cual está basado en la información necesaria para resolver los problemas descritos en [9]. Además de lo anterior, se ha diseñado una arquitectura basada en capas funcionales, en donde cada capa soluciona los problemas que se van encontrando du-

rante el proceso de traducción. Adicionalmente, tanto el diccionario de datos, como la arquitectura en capas soportarán las estructuras para aprender de la información que se obtiene mediante la interacción con el usuario a través del administrador de diálogo.

El nuevo diccionario de datos contiene información sobre palabras de diferentes categorías sintácticas, información estructural de la BD y estadísticas de aprendizaje que hacen referencia a tablas, columnas, relaciones, funciones de agregación, valores alias, valores imprecisos, etc. En [9] se encuentran los detalles del modelo en el que está basado el diccionario de datos.

A diferencia de las versiones anteriores, en esta versión se hace notar la separación entre el usuario del sistema y el administrador. Además, la generación automática del diccionario de datos crea sólo información referente a la estructura de la BD y algunas descripciones, por lo cual el administrador del sistema puede modificar o agregar más información. Este proceso es independiente del proceso realizado por el usuario final cuando realiza las consultas.

Anteriormente la ILNBD tenía tres fases en su módulo de traducción. Al integrar el nuevo diccionario de datos, se crearon capas funcionales que soportan la información para solucionar problemas que en versiones anteriores no se contemplaban. En [6, 9] se describen los problemas encontrados en consultas.

Las capas funcionales diseñadas para la nueva versión de la ILNBD corresponden a:

1. El manejo de valores imprecisos, que son clarificados mediante el administrador de diálogo en caso de no encontrarse en el diccionario de datos.
2. La identificación de tablas y columnas, donde se relacionan las tablas y las columnas con las palabras de diversas categorías gramaticales que se encuentran en la consulta.
3. La identificación de las frases Select y Where, donde se separan los requerimientos de las condiciones de búsqueda.
4. El manejo de la elipsis semántica, donde se clarifican los problemas de elipsis a través del administrador de diálogo.
5. La determinación de reuniones implícitas, en donde se establecen las relaciones entre las tablas que fueron encontradas en la consulta.

Las versiones anteriores alcanzan un porcentaje de éxito en la traducción entre el 79 % - 89 %; con la inclusión del nuevo diccionario de datos (basado en el modelo que hemos desarrollado), la arquitectura basada en capas y el método de aprendizaje esperamos obtener una tasa de éxito hasta de un 95 %.



---

## Conclusiones

Las ILNBDs son herramientas de gran utilidad que facilitan la obtención de información contenida en BDs a través de consultas en LN. La exactitud de la información obtenida en estos sistemas es fundamental para la toma de decisiones; sin embargo, a pesar de su importancia y a la cantidad de sistemas desarrollados hasta la fecha, aún continúan existiendo muchas deficiencias. Los problemas encontrados son extremadamente complejos. Como ejemplo de las limitaciones encontradas, tenemos que incluso las ILNBDs comerciales han sido descontinuadas y el hecho de que ninguna se ha acercado al 100 % de consultas correctamente traducidas.

Para solucionar los problemas existentes hemos propuesto un nuevo diccionario de datos basado en un modelo de BDs que incluye información relevante para la correcta traducción de consultas en ILNBDs. Creemos que, para implementar un proceso de traducción exitoso, es necesaria una arquitectura basada en capas de funcionalidad (como las diseñadas en el modelo de comunicaciones OSI), donde cada capa funcional soluciona un problema y va refinando la información recibida. Con la inclusión de un método de aprendizaje, que obtenga información de la interacción con el usuario a través del administrador de diálogo, se espera alcanzar una tasa de éxito en la traducción cercana al 95 %. El aprendizaje obtenido será utilizado para modificar la información semántica contenida en el diccionario de datos implementado.

A pesar que desde hace varias décadas existe el desarrollo de este tipo de sistemas, el área de investigación sobre ILNBDs continúa abierta, por lo cual, es necesario identificar y evaluar cada uno de los tipos de problemas que se encuentran en el procesamiento de consultas para alcanzar un porcentaje de éxito aceptable en el proceso

de traducción.☞

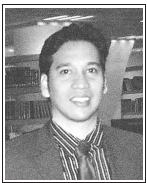
---

## REFERENCIAS

1. Sethi V. (1986) "Natural Language Interfaces to Databases: MSI Impact, and Survey of their Use and Importance". University of Pittsburgh, 1986.
2. Cimiano P., Haase P., Heizmann J. (2007) "Porting Natural Language Interfaces Between Domains: an Experimental User Study with the ORAKEL System". En: Proc. 12Th International Conference on Intelligent User Interfaces. Honolulu, Hawaii, pp. 180–190
3. BBC News (2009) Bill Gates Says: Mouse is Out, Touch Screen and Natural Language Interface Are In. [http://news.bbc.co.uk/player/nol/newsid\\_7170000/newsid\\_7174300/7174330.stm?bw=bb&mp=wm&asb=1&news=1&bbcws=1](http://news.bbc.co.uk/player/nol/newsid_7170000/newsid_7174300/7174330.stm?bw=bb&mp=wm&asb=1&news=1&bbcws=1) Recuperado en 23 de Junio de 2012.
4. Popescu A, Armanasu A, Etzioni O et al. (2004) "Modern natural language interfaces to databases: composing statistical parsing with semantic tractability". Proc. 20th International Conference on Computational Linguistics.
5. López V., Pasin M. y Motta E. (2005) "AquaLog: An Ontology-Portable Question Answering System for the Semantic Web". A. Gómez-Pérez and J. Euzenat (Eds.): ESWC 2005, LNCS 3532, Springer-Verlag, Berlin, Heidelberg 2005, pp. 546–562.
6. Pazos R., González J., Aguirre M., Martínez J., Fraire H. (2012) "Natural Language Interfaces to Databases: An Analysis of the State of the Art". In Proc. International Seminar on Computational Intelligence 2012. Por aparecer en 2012.
7. Pazos R., Pérez J., et al. (2005) "A Domain Independent Natural Language Interface to Databases Capable of Processing Complex Queries". Lecture Notes in Artificial Intelligence, Vol. 3789. Springer-Verlag, pp. 833–842.
8. Pazos R., Rojas J., et al. (2010) "Dialogue Manager for a NLIDB for Solving the Semantic Ellipsis Problem in Query Formulation". Lecture Notes in Artificial Intelligence, Vol. 6277. Springer-Verlag, pp. 203–213.
9. Pazos R., Gonzalez J., Aguirre M. (2011) Semantic Model for Improving the Performance of Natural Language Interfaces. Proc. 10th Mexican International Conference on Artificial Intelligence Vol. 7094, pp. 277–290.

---

## SOBRE LOS AUTORES



**Marco Antonio Aguirre Lam** es un estudiante de doctorado en el Departamento de Estudios de Posgrado e Investigación en el Instituto Tecnológico de Cd. Madero. El obtuvo su grado de maestría en Ciencias en Ciencias de la Computación en el Instituto Tecnológico de Cd. Madero en 2009. Sus intereses de investigación incluyen procesamiento de lenguaje natural y optimización inteligente.



**Rodolfo A. Pazos Rangel** obtuvo el grado de doctor en Ciencias de la Computación en la University of California at Los Angeles en 1983. Actualmente es profesor investigador en el Inst. Tecnológico de Cd. Madero. El Dr. Pazos es miembro del Sistema Nacional de Investigadores Nivel II. Sus intereses científicos incluyen procesamiento de lenguaje natural y algoritmia.

---