# Extracción automática de contextos definitorios en textos especializados

**Gerardo Sierra Martínez**
Universidad Nacional Autónoma de México
Instituto de Ingeniería
gsierram@ii.unam.mx

**Rodrigo Alarcón Martínez**
Universidad Pompeu Fabra
Instituto Universitario de Lingüística Aplicada
ralarconm@ii.unam.mx

**César Antonio Aguilar**
Universidad Nacional Autónoma de México
Instituto de Ingeniería
caguilar@ii.unam.mx

Uno de los problemas recurrentes de cualquier área de conocimiento es la organización y explicación de los términos que incluye en su campo de estudio. El reciente avance en el desarrollo de nuevas tecnologías para el trabajo terminológico ha aportado diversas herramientas para tratar de resolver este problema. Una de estas herramientas son los corpus de textos especializados en los cuales se pueden extraer automáticamente términos y definiciones.

Dentro de éste marco, el Grupo de Ingeniería Lingüística desarrolla un proyecto en torno a la descripción y extracción automática de *contextos definitorios* (CDs), los cuales definiremos como aquellos fragmentos de un texto especializado que aportan información útil para entender un término en su contexto real, y que pueden ser puntos de inicio para la elaboración de ontologías, glosarios, diccionarios electrónicos, entre otras importantes aplicaciones.

Los CDs incluyen un término, una definición y patrones definitorios, como patrones verbales (*se define como*, *constituido por*), o bien elementos estilísticos como la presencia de marcas tipográficas y variaciones en la tipografía textual que ayudan a resaltar la presencia del término o la definición (*comillas*, *cursivas*).

El estudio de los CDs involucra diversas líneas de investigación que pueden ser divididas en estudios descriptivos y estudios aplicados. Por un lado, es necesario describir el comportamiento lingüístico de los elementos constitutivos de los CDs, y por otro lado es necesario elaborar una metodología para su extracción automática.

Así, encontramos que el estudio lingüístico involucra un análisis descriptivo de los distintos tipos de definiciones que suelen introducir los patrones verbales definitorios. A su vez, es común que en un texto especializado no se repitan constantemente los términos. En su lugar suelen aparecer referencias anafóricas que los sustituyen y que en muchos casos ocupan el lugar del término en el contexto definitorio.

A partir de estos trabajos descriptivos podemos observar que es necesario, en primer lugar, elaborar una herramienta de búsqueda para la extracción automática de CDs, y en segundo lugar, identificar automáticamente en estos contextos los elementos constitutivos: el término y la definición. Asimismo, se requiere identificar automáticamente cuál es el término en el caso en que éste se sustituye mediante una referencia anafórica.

Aunque existen varios enfoques metodológicos para la extracción conceptual en textos especializados, el presente proyecto propone desarrollar un sistema completo y coherente de estructura modular, basado en información lingüística, que sea aplicable a diversos corpus textuales especializados en lengua española con el fin de extraer automáticamente términos y definiciones. Igualmente, este proyecto tiene la finalidad de conformar un Corpus de Contextos Definitorios, esto es, un repositorio electrónico para los términos, definiciones y aquellos patrones definitorios que suelen coocurrir en los CDs.

En el proyecto participan varios grupos de investigación. En el aspecto más teórico, se encuentra un grupo que estudia el concepto de definición. En un aspecto teórico-práctico se analiza desde el punto de vista lingüístico la relación entre el verbo definitorio y el tipo de definición. Otro grupo en el terreno de la terminótica investiga los patrones sintácticos de los términos en español con el fin de identificar automáticamente los términos presentes en los contextos definitorios. Otro grupo más estudia el comportamiento de anáforas en CDs. Finalmente, otro grupo busca desde la lingüística computacional elaborar un extractor

de CDs, analizando de manera integral los estudios anteriores.

## Publicaciones


Alarcón R., Sierra G. (2003) "The Role of Verbal Predications for Definitional Contexts Extraction". *5º Encuentro Terminología & Inteligencia Artificial (TIA)*. Estrasburgo, Francia.

Alarcón R. y Sierra G. (2003) "El rol de las predicaciones verbales en la extracción automática de conceptos". *Estudios de Lingüística Aplicada* 38.

Sierra G. y Alarcón R. (2002) "Identification of recurrent patterns to extract definitory contexts". En *Lecture Notes in Computer Science,* Springer, No. 2276.

Sierra G., Medina A., Alarcón R. y Aguilar C. (2003) "Towards the extraction of conceptual information from corpora". En D. Archer, P. Rayson, A. Wilson & A. McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 conference*, UCREL Technical Paper, No. 16, Lancaster University.

Sierra, G., Alarcón, R., Medina, A. y Aguilar, C. (2003): "Definitional Contexts Extraction from Specialised Texts". En B. Lewandowska-Tomaszczyk, (ed.), *PALC 2003 Proceedings: Language, Corpora and E-Learning*, , Frankfurt: Peter Lang Publish.


## Grupo de Investigación Principal

*Grupo de Ingeniería Lingüística*, Instituto de Ingeniería, Universidad Nacional Autónoma de México:
http://www.iling.unam.mx/

**Líder del Proyecto**:
Doctor Gerardo Sierra, gsierram@ii.unam.mx (II-UNAM).

**Estudiantes de doctorado**:
Rodrigo Alarcón, ralarconm@ii.unam.mx (IULA-UPF). Línea de trabajo: Elaboración de un extractor de contextos definitorios en textos especializados.
César A. Aguilar, caguilar@ii.unam.mx (II-UNAM). Línea de trabajo: Análisis lingüístico de patrones verbales definitorios y de definiciones.

**Estudiantes de maestría**:
Alberto Barrón, alberto@pumas.ii.unam.mx (II-UNAM). Línea de trabajo: Elaboración de un identificador de términos en contextos definitorios.
Grizel Delgado, mdelgador@ii.unam.mx (ISI-University of Düsseldorf). Línea de trabajo: Identificación automática de relaciones anafóricas en contextos definitorios.

**Estudiantes de licenciatura**:
Itzia Bacca, ibaccai@ii.unam.mx (II-UNAM). Línea de trabajo: Tipología y descripción lógica de definiciones.
Valeria Benítez. vbenitezr@ii.unam.mx (II-UNAM). Línea de trabajo: Análisis de relaciones anafóricas en contextos definitorios.

### Colaboraciones

*Grupo IULATerm,* Instituto Universitario de Lingüística Aplicada, Universidad Pompeu Fabra (Barcelona, España):
http://www.iula.upf.es/iulaterm/tpresca.htm
Doctor Lluis de Yzaguirre, (IULA-UPF) lluis.deizaguirre@upf.edu Asesoría en programación para el extractor de contextos defintorios.
Doctora Carme Bach, (IULA-UPF) carme.bach@upf.edu Asesoría en terminología.
*Groupe Éclectik,* del Observatoire de Linguistique Sense-Texte, Universidad de Montreal (Canadá):
http://www.olst.umontreal.ca/eclectikfr.html
Doctora Marie-ClaudeL'homme, (OLST-UMontreal) mc.lhomme@umontreal.ca.
Asesoría en terminológia.
PatrickDrouin, (OLST-UMontreal) patrick.drouin@umontreal.ca Asesoría en extracción terminológica.
*National Centre for Text Mining*, School of Informatics, University of Manchester (Inglaterra): http://www.nactem.ac.uk
Doctora Sophie Anaiadou, (NaCTeM-UM) Sophia.ananiadou@manchester.ac.uk Asesoría en extracción terminológica.
Investigador Naoaki Okazaki, (NaCTeM-UM) okazaki@mi.ci.i.u-tokyo.ac.jp Asesoría en extracción terminológica.


**Financiamiento**: Consejo Nacional de Ciencia y Tecnología, México. Proyecto 46832 "Extracción de conceptos en textos de especialidad a través del reconocimiento de patrones lingüísticos y metalingüísticos".
**Co-financiamiento**: Programa Transdisciplinario en Investigación y Desarrollo de la Secretaría de Desarrollo Institucional, UNAM, México. Macro-Proyecto "Tecnologías para la Universidad de la Información y la Computación".