ANÁLISIS DE TEXTO PARA LA IDENTIFICACIÓN AUTOMÁTICA DE MARCADORES LINGÜÍSTICOS DEFINICIONALES EN RECETAS DE GASTRONOMÍA DE COSTA RICA

Text analysis for automatic identification of definitional linguistic markers in Costa Rican gastronomy recipes

Sharon Corrales Montero[*]
Karen Miranda Hernández[**]
Édgar Casasola Murillo[***]
Jorge Antonio Leoni de León [****]
Mario Hernández Delgado[*****]

RESUMEN

El análisis de contextos definicionales permite clasificar y sistematizar las informaciones definicionales pertenecientes a un dominio específico y, posteriormente, identificar estándares de las formas en que se definen las palabras y términos en tal dominio. En este artículo, se describe el proceso realizado para automatizar el análisis de contextos definicionales en el dominio gastronómico de Costa Rica. La labor se realizó mediante el uso de herramientas computacionales para el procesamiento de lenguaje natural. La automatización permite el análisis sobre grandes volúmenes de datos y obtener resultados en menos tiempo del requerido por el análisis manual. Ahora bien, el procedimiento consta de dos módulos, uno de clasificación de documentos en textos con recetas o sin ellas y un segundo módulo de identificación de los ingredientes de cocina con base en patrones lingüísticos formales.

Palabras clave: análisis lingüístico de recetas, análisis de contextos definicionales, patrones definicionales, marcadores definicionales, procesamiento del lenguaje natural.

[*] Universidad de Costa Rica. Estudiante de la Maestría en Computación. Costa Rica. Correo electrónico: sharoncm.1691@gmail.com
[**] Universidad de Costa Rica. Estudiante de la Maestría en Computación. Costa Rica. Correo electrónico: karenmh09@gmail.com
[***] Universidad de Costa Rica. Escuela de Computación e Informática y Posgrado en Computación. Costa Rica. Correo electrónico: casasola@gmail.com
[****] Universidad de Costa Rica. Escuela de Filología, Lingüística y Literatura y Posgrado en Lingüística. Costa Rica. Correo electrónico: antonio.leoni@ucr.ac.cr
[*****] Universidad de Costa Rica. Programa Estudios de Lexicografía. Costa Rica. Correo electrónico: pdfmario@gmail.com
Recepción: 28/5/2017 Aceptación: 15/7/2017.

ABSTRACT

The analysis of definitional contexts allows to classify and systematize the definitional information belonging to a specific domain, and then to identify standards for the forms in which words and terms are defined in this domain. This paper describes the process implemented to automate the analysis of definitional contexts in the gastronomy domain in Costa Rica. The automation was done by using computational tools for natural language processing. The automation enables analysis of large quantities of data and results in less time than required by manual analysis. Automation consists of two modules, the first one is for the classification of documents in texts with or without recipes and the second one is for the identification of recipe ingredients based on formal linguistic patterns.
Key Words: linguistic analysis of recipes, analysis of definitional contexts, definitional patterns, definitional markers, natural language processing.

1. Introducción

El análisis de contextos definicionales o definitorios (en adelante, análisis de CD) es una línea de investigación que permite, en primer lugar, clasificar y sistematizar las informaciones definicionales relativas a un dominio restringido. Posteriormente, esa organización conceptual puede servir tanto para la recuperación de relaciones semánticas definitorias a partir de textos como para la estandarización de las formulaciones definicionales del dominio de especialidad estudiado (cf. Alarcón, 2003; Sierra, Alarcón y Aguilar, 2006; Alcina y Valero, 2008).

A causa del interés en las posibilidades de este tipo de estudios, surge el proyecto "Análisis de contextos definicionales en corpus de gastronomía tradicional en Costa Rica (CODEGAT)", investigación que pretende examinar la información gastronómica presente en textos de recetas costarricenses, con el fin último de aportar a la sistematización del conocimiento gastronómico socializado.

Parte importante de esa sistematización es la adecuada identificación de la lista de productos/ingredientes que serán objeto de las diversas acciones y procesos, así como la precisa descripción de las tareas paralelas y secuenciales en las que aquellos se utilizarán. Desde el enfoque del análisis de CD, que es el que aquí se sigue, lo fundamental es identificar las formas recurrentes que se utilizan efectivamente en los textos para la expresión de las relaciones conceptuales pertinentes (cf. Sierra y Alarcón, 2002; Soler, 2005; Sierra, 2009; Valero, 2009; Valero y Alcina, 2009). A esas formas recurrentes se les llama, en esta perspectiva investigativa, "patrones definicionales", cada uno de los cuales asocia una **clase de contenidos semánticos** con una **clase de formas que sirven para introducirlos** dentro de la cadena textual (y que funcionan como marcadores, señalizadores, indicadores).

A pesar de tener ya varios lustros en desarrollo, la línea de análisis de CD no cuenta aún con paradigmas metodológicos de empleo universal.[1] Sin embargo, una característica esencial de su planteamiento es la automatización de los procedimientos de identificación y validación de los patrones definitorios propuestos, así como de la recuperación de las relaciones conceptuales pertinentes. Esta automatización permite trabajar sobre grandes volúmenes de datos y obtener resultados en menos tiempo que el requerido por el análisis manual. Debido a lo anterior, el equipo de trabajo de CODEGAT incluye tanto a lingüistas como a especialistas con conocimientos en procesamiento del lenguaje natural.

Ahora bien, en relación con el proceso del análisis de texto, este se dividió en dos módulos (v. Figura 1). El primer módulo corresponde a la clasificación de los documentos en aquellos que contienen información de recetas y aquellos que no la contienen. Una vez así clasificados, se toman solamente los documentos contenedores de recetas y se aplica el segundo módulo. En este, las palabras de cada documento son etiquetadas según su categoría gramatical. Posteriormente, sobre el texto etiquetado se buscan marcadores lingüísticos y se genera un documento de resultado, el cual contiene marcados los ingredientes dentro de la receta.

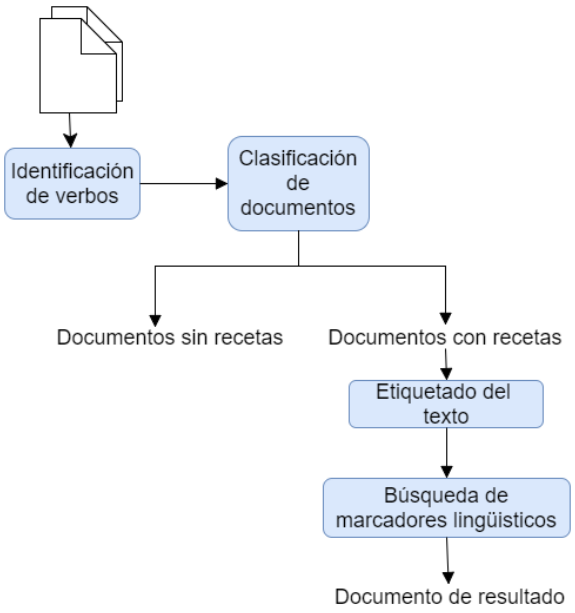


Figura 1. Descripción del proceso realizado

[1] Precisamente por la búsqueda de modelos metodológicos eficaces con vista en los objetivos del análisis de CD, parte importante de los estudios enmarcados en esta línea de investigación es la propuesta y continuo afinamiento de los procedimientos aplicados a las diversas etapas.

Antes de aplicar el procedimiento de dos módulos, se realizó una descarga masiva de documentos mediante una araña de búsqueda. Esta araña estaba guiada por hipervínculos brindados por los lingüistas del equipo.

En cuanto a las herramientas computacionales utilizadas en el proceso descrito en la Figura 1, estas corresponden a agentes automáticos de recolección de información, expresiones regulares y etiquetador de partes del discurso:

- **Agente automático de recolección de información:** proceso que se ejecuta de forma continua, sigue enlaces y busca adelante por información inferida (cf. Casasola y Gauch, 1997).
- **Expresión regular:** secuencia de caracteres utilizada como patrón para describir, manipular y realizar búsquedas dentro del texto. Es una herramienta sumamente flexible y eficiente para el procesamiento de texto (cf. Habibi, 2004; Friedl, 2006; Fitzgerald, 2012).
- **Etiquetador de partes del discurso (POS tagger):** *software* para asignar a cada palabra dentro del texto una etiqueta con base en la función que asume en la oración (referida especialmente a la clase léxica o la clase morfológica). Este etiquetado es importante en el área de recuperación de información y procesamiento de lenguaje natural porque encapsula datos propios de la palabra (número, género, tiempo verbal, entre otros), así como de sus palabras vecinas (cf. Hasan, UzZaman y Khan, 2007).

2. Procesamiento

2.1. Clasificación de documentos

Este módulo es el encargado de analizar los documentos para clasificarlos en dos categorías: documentos con recetas y documentos sin recetas. La implementación de la lógica del módulo se realizó en dos etapas.

2.1.1. Etapa 1

En esta etapa, se utilizaron únicamente documentos de recetas previamente analizados por los lingüistas involucrados en la investigación. Estos documentos se analizaron mediante el

uso de un etiquetador de partes del discurso (POS tagger) para identificar de manera automática los verbos presentes en los textos. El resultado del proceso mostraba la lista de los verbos identificados, con su correspondiente forma en infinitivo (lema) y su frecuencia absoluta de aparición en los textos analizados.

A partir de este resultado, se consideró el papel desempeñado por cada verbo en las recetas de cocina. De esta manera, se clasificaron los verbos en dos categorías de significancia (media y alta), basándose en qué tanto es exclusivo cada verbo del dominio de la gastronomía y las recetas de cocina, lo cual es útil para una identificación automática de recetas. Por ejemplo, algunos verbos encontrados en los textos que se estudiaron son de uso común en otros dominios y, por lo tanto, no pueden asociarse de manera única al contexto de cocina. Sin embargo, otros verbos son claramente exclusivos del discurso gastronómico, hecho que permite pensar en la posibilidad de utilizarlos instrumentalmente para la identificación de recetas de cocina dentro de un corpus textual inicialmente indiferenciado.

- **Significancia media**: verbos comunes en diversos dominios y, por tanto, no exclusivos de los contextos de recetas. Por ejemplo: cocinar, servir, hacer, mezclar.
- **Significancia alta**: verbos que pueden asociarse comúnmente al contexto de la gastronomía y las recetas de cocina. Por ejemplo: amasar y picar.

2.1.2. Etapa 2

Para la segunda etapa, se tomó como base de la clasificación de documentos el resultado obtenido en la etapa 1, correspondiente a la significancia de los verbos. Además, se procedió a la creación de un agente automático de recolección de información (v. Casasola y Gauch, 1997) para la descarga masiva de documentos de internet. Este agente utiliza, inicialmente, un conjunto de páginas web a las que se les conoce como *semillas* del agente. Estas semillas son visitadas y cualquier enlace incluido en ellas es agregado a la lista de páginas web por visitar y descargar.

Por un lado, para asegurarse de que la mayoría de los documentos descargados sean gastronómicos, este agente confronta los documentos por descargar contra una lista de páginas web ya validadas. Esta lista de páginas válidas fue previamente brindada por los lingüistas del equipo, quienes las analizaron según criterios de contenido y las valoraron como páginas con gran densidad de recetas de cocina.

Por otro lado, para la clasificación automática de documentos se construyó un programa que recibe como parámetro la ubicación de la carpeta donde se encuentran los documentos descargados por el agente y procede a analizar cada uno de ellos. El primer paso del proceso es el etiquetado del texto de los documentos mediante un POS tagger para identificar los verbos presentes.

Posteriormente, se verifica si los verbos identificados se encuentran clasificados en la lista de verbos de significancia media o alta. Si el documento contiene al menos un verbo de significancia alta o si contiene al menos cuatro verbos de significancia media, el documento es clasificado como contenedor de recetas. Por lo tanto, todo verbo no clasificado previamente tendrá impacto nulo en la clasificación del documento.

Con respecto a la elección de un mínimo de cuatro verbos de la categoría de significancia media, esto se debe a que cada uno de estos verbos por sí solo no permite asegurar que el documento contiene recetas. Sin embargo, la aparición de varios de estos verbos en un mismo documento aumenta la posibilidad de que tal documento efectivamente contenga descripciones culinarias.

Por último, el programa genera un archivo de resultado, en el cual se indica la clasificación asignada a cada documento analizado. Este archivo de resultado presenta, además, los verbos detectados en el texto del documento durante el análisis, tal como se ejemplifica en la Tabla 1:

Tabla 1. Ejemplos de resultado de clasificación de documentos

RESULTADO
El archivo 0 es receta. Verbos: [cocer] Verbos [comer, cocinar, adornar, servir, probar]
El archivo 102 es NO receta. Verbos: [ ]

Verbos: [servir]
El archivo 135 es receta. Verbos: [hornear, batir] Verbos: [ ]
El archivo 179 es receta. Verbos: [ ] Verbos: [moler, majar, arrollar, pelar, enfriar, cocinar, cortar]

Asimismo, con el fin de evaluar la precisión del programa, se realizó una clasificación manual de los documentos, de modo que se pudieran comparar los resultados automáticos contra los manuales. Según la matriz de confusión resultante (Tabla 2), la precisión del sistema resultó ser del 77 %.

Tabla 2. Matriz de confusión de la clasificación de documentos: clasificación obtenida contra clasificación real

		Clasificación obtenida	
		Receta	No-Receta
Clasificación real	Receta	301	82
	No-Receta	88	37

2.2. Identificación de ingredientes

Este módulo trabaja únicamente sobre los documentos clasificados en el módulo anterior como documentos con recetas. Antes de iniciar el procesamiento propio de este componente, se aplica sobre el texto un preprocesamiento para normalizarlo y estandarizarlo. La normalización

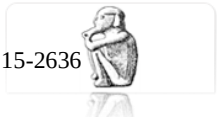
consiste en la eliminación o sustitución de los caracteres especiales, conversión de todo el texto a minúscula, eliminación de espacios múltiples entre palabras, entre otros. Por su parte, la estandarización consiste en la transformación de caracteres especiales de representación de fracciones (por ejemplo, ½, ¼) a su forma normal, es decir, mediante caracteres individuales (por ejemplo, 1/2, 1/4, respectivamente). Es conveniente señalar que en un inicio este preprocesamiento no se realizaba. No obstante, la falta de estandarización en el texto provocaba problemas en la identificación de los marcadores lingüísticos.

Ahora bien, el proceso realizado en este segundo módulo ha tenido dos etapas. La mayor diferencia entre ambas es el paso de usar expresiones regulares basadas en reglas con palabras específicas a expresiones regulares basadas en reglas con categorías gramaticales.

2.2.1. Etapa 1

Los lingüistas del equipo brindaron un corpus de prueba (CP) constituido por texto plano rico en información gastronómica (y, especialmente, denso en recetas). El CP fue extraído de internet por medio de un proceso automatizado y, posteriormente, depurado y prenormalizado de forma masiva para ser utilizado como corpus anónimo. También ofrecieron una lista de formas lingüísticas postuladas como marcadores definicionales de ingredientes –en la Figura 2 se pueden ver algunos ejemplos–:

2_barras_de_
2_botellas_de_
2_cabezas_de_
2_[X]_grandes
2_[X]_medianas
2_[X]_pequeñas
2_[X]_picadas
2_[X]_tiernos



2_cucharadas_de_

2_ cucharaditas_de_

2_dientes_de_

2_hojas_de_

2_[X]_batidos

2_[X]_duros

Figura 2. Ejemplos de marcadores lingüísticos iniciales

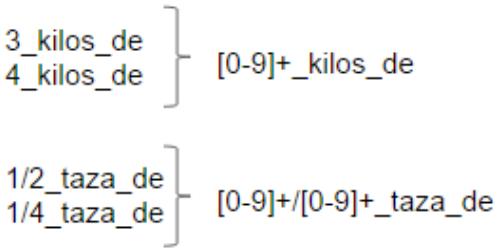
Estos candidatos a marcadores habían sido previamente identificados mediante un proceso manual de análisis de un corpus base (CB) –diferente del corpus de prueba CP ya mencionado–; luego fueron generalizados (o pregeneralizados), utilizando una simbología que pudiera servir de transición hacia la posterior formulación mediante expresiones regulares. La simbología de esa generalización inicial se puede observar en el Tabla 3:

Tabla 3. Simbología de los marcadores lingüísticos iniciales

Notación	Significado
[X]	Conjunto obligatorio y variable de 1 o más letras
[[X]]	Conjunto opcional y variable de 1 o más letras
[[elemento]]	Elemento opcional
_	Espacio en blanco

Los marcadores lingüísticos brindados fueron, entonces, transformados por medio de simbología utilizada por las expresiones lingüísticas del lenguaje de programación seleccionado

para la automatización. Este paso a expresiones regulares permitió abstraer los patrones iniciales y, por ende, se disminuyó la cantidad de patrones por evaluar. La abstracción se logró, en su mayoría, al pasar de números específicos (0, 1, 2, 3, 4, 5,... n) a una expresión regular de un conjunto de números, como en:

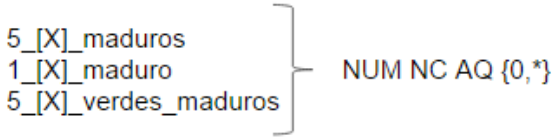


Estas expresiones regulares se buscaron en el texto para identificar los puntos de inserción de ingredientes en las recetas de cocina. Por último, se generaba un documento de resultados en el que se señalaban los marcadores encontrados y se desplegaba la frecuencia absoluta de aparición de cada uno de ellos. De esta manera se lograron identificar también marcadores que no resultaban útiles para la investigación, debido a su baja o nula frecuencia absoluta de aparición. Además, se logró identificar los casos de superposición de patrones; una vez identificadas y evaluadas estas superposiciones, se eliminaron las expresiones regulares que producían las redundancias.

A pesar de lograr identificar la mayoría de los ingredientes de cocina por medio de marcadores, también había aquellos que no lograban ser detectados. Una vez analizados los resultados, se observó que muchos de los casos de ingredientes no identificados se debían a falta de coincidencia entre el género o número gramatical de las formas que aparecían en el texto y el género o número de las formas postuladas por los marcadores. Asimismo, no todas las medidas y sus diversas formas de escribirse estaban consideradas en los marcadores. A partir de la revisión de esos resultados, se llegó a la conclusión de que era necesario generalizar de manera un poco diferente los marcadores, de modo que siempre incluyeran al menos todas las posibles inflexiones de género y número.

2.2.2. Etapa 2

Esta segunda versión del módulo se desarrolló con el fin de solucionar el problema de las inflexiones nominales (número) y adjetivales (género y número), además de otras generalizaciones pertinentes. Para esto se empleó un etiquetador de partes del discurso (POS tagger) con un modelo correspondiente al lenguaje español. Este etiquetador permitió pasar de las expresiones regulares de la primera etapa a expresiones regulares basadas en categorías gramaticales.



Para la construcción de estas nuevas expresiones regulares se utilizaron los marcadores brindados inicialmente. Además, las categorías de interés corresponden únicamente a los valores numéricos, sustantivos comunes, adjetivos calificativos, signos de puntuación y preposiciones. Esto permitió disminuir más la cantidad de marcadores por evaluar en el texto, ya que las categorías gramaticales generalizaron valores específicos.

Primeramente, este módulo toma cada uno de los documentos y etiqueta su texto según las partes del discurso. En el siguiente paso, se analiza el texto etiquetado para identificar la presencia de los marcadores definicionales de interés. Por último, el proceso genera un documento de resultados por cada documento analizado. En este documento de resultados, se presentan señalados los marcadores definicionales e ingredientes encontrados en el documento.

En cuanto a los resultados, estos se evaluaron cuantitativamente según la cantidad de ingredientes identificados correctamente con respecto al total de los ingredientes en los documentos analizados. Este módulo identificó correctamente y en forma automática el 53 % de los ingredientes.

3. Conclusiones

El uso de herramientas de computación para el procesamiento automático de texto demostró ser de utilidad para la recolección, clasificación automática e identificación de ingredientes de recetas de gastronomía con base en marcadores lingüísticos. En relación con el POS tagger, este permitió identificar los verbos dentro del texto para hacer una discriminación automática de documentos, así como considerar las inflexiones nominales y adjetivales en los marcadores lingüísticos. De la misma forma, el uso de esta herramienta permitió disminuir la cantidad de tiempo en el desarrollo del proceso, ya que no se requirió realizar manualmente la clasificación en categorías gramaticales de cada una de las palabras en los documentos.

Asimismo, la combinación entre expresiones regulares y categorías gramaticales permitió la generalización y expansión de los marcadores que los lingüistas del equipo habían brindado pregeneralizados.

3.1. Trabajo futuro

La investigación presentada en este artículo proyecta extenderse y optimizarse para la obtención de mejores resultados. El plan es dividir las tareas computacionales en dos temas a ser desarrollados como trabajos finales de investigación aplicada (TFIA) en la Maestría en Computación. El primero se refiere a la clasificación automática de documentos, utilizando aprendizaje de máquina y marcadores lingüísticos. Este trabajo se enfocará en la clasificación de textos para diferenciar entre archivos que tienen información gastronómica (específicamente, recetas) y archivos que no la tienen; esto requerirá trabajar en conjunto con los expertos en el área de lingüística para asiganarles pesos (probabilidades) a los verbos utilizados para la clasificación.

Por otro lado, el segundo tema corresponde al análisis automático de textos de recetas de cocina para la identificación de procesos paralelos y secuenciales. Este trabajo podrá utilizar como base el proceso realizado para la identificación de los ingredientes en las recetas de cocina por medio de patrones definicionales y, además, requerirá contar con un listado de marcadores

lingüísticos definitorios asociados a los diversos pasos/tareas/etapas de los procedimientos culinarios.

Bibliografía

Alarcón, Rodrigo. (2003). *Análisis lingüístico de contextos definitorios en textos de especialidad* (Tesis de licenciatura). Universidad Nacional Autónoma de México, México.

Alcina, Amparo y Valero, Esperanza. (2008). Análisis de las definiciones del diccionario cerámico científico-práctico. Sugerencias para la elaboración de patrones de definición. *Debate Terminológico*, 4. Recuperado de http://seer.ufrgs.br/index.php/riterm/article/download/23841/13830

Casasola, Édgar y Susan Gauch. (1997). Intelligent Information Agents for the World Wide Web. *Technical report ITTC-FY97-111100-1*. Information and Telecommunication Technology Center. Recuperado de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.3628&rep=rep1&type=pdf

Elleithy, Khaled. (Ed.). (2007). *Advances and Innovations in Systems, Computing Sciences and Software Engineering*. Dordrecht, Holanda: Springer.

Fitzgerald, Michael. (2012). *Introducing Regular Expressions*. California, EE. UU.: O'Reilly Media Inc.

Friedl, Jeffrey E. F. (2006). *Mastering Regular Expressions* (3.ª ed.). California, EE. UU.: O'Reilly Media Inc.

Habibi, Mehran. (2004). *Java Regular Expressions: Taming the java.util.regex Engine*. Nueva York, EE. UU.: Apress Media, LLC.

Hasan, Fahim Muhammad, UzZaman, Naushad y Khan, Mumit. (2007). Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla. En K. Elleithy (Ed.), Advances and Innovations in Systems, Computer Sciences and Software Engeneering, 121-126. https://doi.org/10.1007/978-1-4020-6264-3_23

Sierra, Gerardo. (2009). Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos. *linguaMATICA*, 2, 13-37.

Sierra, Gerardo y Alarcón, Rodrigo. (2002). Identification of recurrent patterns to extract definitory contexts. *Lecture Notes in Computer Science*, 2276, 436-438.

Sierra, Gerardo, Alarcón, Rodrigo y Aguilar, César. (2006). Extracción automática de contextos definitorios en textos especializados. *Revista de Procesamiento de Lenguaje Natural*, 37, 351-352.

Sierra, Gerardo, Pozzi, Mara y Torres, Juan Manuel. (Eds.). (18 de setiembre de 2009). Proceedings. *1st International Workshop on Definition Extraction*. Borovets, Bulgaria. Recuperado de https://aclweb.org/anthology/W/W09/W09-4400.pdf

Soler, Victoria. (2005). *Patrones lingüísticos para la búsqueda de información conceptual en el corpus textual especializado de la cerámica TXTCera*. Recuperado de http://repositori.uji.es/xmlui/bitstream/handle/10234/79115/forum_2004_50.pdf?sequence=1

Valero, Esperanza. (2009). *Los marcadores lingüísticos en las definiciones del grupo conceptual 'procesos de fabricación cerámica'*. Recuperado de http://repositori.uji.es/xmlui/bitstream/handle/10234/78051/forum_2008_22.pdf?sequence=1

Valero, Esperanza y Alcina, Amparo. (2009). Linguistic realization of conceptual features in terminographic dictionary definitions. *Workshop On Definition Extraction 2009*, Borovets, Bulgaria, 54-60.