Extraction of candidate terms from a corpus of non-specialized, general language

Gilberto Anguiano Peña *
Catalina Naumis Peña **

Paper submited: October 21, 2013.
Accepted: October 9, 2014.

Abstract

Linguistic phenomena associated with the analysis of document content and employed for the purpose of organization and retrieval are well-visited objects of study in the field of library and information science. Language often acts as a gatekeeper, admitting or excluding people from gaining access to knowledge. As such, the terms used in the scientific and technical language of research need to be kept up and their behavior within the domain examined. Documental content analysis of scientific texts provides knowledge of specialized lexicons and their specific applications, while

* El Colegio de México, México. ganguia@colmex.mx
** Instituto de Investigaciones Bibliotecológicas y de la Información de la UNAM, México. naumis@unam.mx

differentiating them from common use in order to establish indexing languages. Thus, as proposed herein, the application of lexicographic techniques to documental content analysis of non-specialized language yields the components needed to describe and extract lexical units of the specialized language.

Keywords: Content Analysis; Term Extraction; Scientific Language; Corpus of General Language.

Resumen

Extracción de candidatos a términos de un corpus de la lengua general

Gilberto Anguiano-Peña y *Catalina Naumis-Peña*

Entre los objetos de estudio de la Bibliotecología e Información se incluyen los fenómenos lingüísticos asociados al análisis de contenido documental tanto para organizar la información como para recuperarla. Para ello, se deben rescatar los términos usados en el lenguaje científico y técnico, estudiar su ámbito de dominio y comportamiento. A través de la lengua se controla y se excluye el conocimiento que una población pueda obtener. El análisis documental del contenido, en este caso de los textos de difusión científica, permite obtener un conocimiento de las unidades léxicas, sus aplicaciones significativas y separar los términos de la lengua general para crear lenguajes de indización. Es así que por medio del análisis de contenido documental en un corpus de lengua general marcado con los métodos de la lexicografía se obtienen y caracterizan los componentes que permiten extraer unidades léxicas del lenguaje especializado mediante las técnicas propuestas en el presente trabajo.

Palabras clave: Análisis de contenido; Extracción de términos; Lenguaje científico; Corpus de lengua general.