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.

Introduction

The general objective of this paper is to determine the methodologies and strategies for processing a linguistic body of a general language to obtain the specialized terms used in a field and the terms shared among several fields, and thereby separate them automatically from the mass of lexical units of the general language. This kind of work allows terms to be obtained and their meanings later clarified. It also allows one to learn of the uses of the texts in which such terms appear and use these terms in the construction of documental languages.

Is the language of science a secret language? In a digital newspaper in Murcia, Spain, the following scientific dissemination article was published: *El lenguaje secreto de la ciencia*, whose author contends: "[S]ince algebra all the way to geometry and on to aerodynamics, mathematics have been at the center of all scientific inquiry and are now basic to our daily lives" (Moreno, 2011. Translated from Spanish); thereby suggesting that mathematics is the secret language permeating all science. The Research into basic mathematics she presents justifies this asseveration is supported by maths professor Manuel Saorín, who states: "Withdrawing money from and ATM or sending an email would not be possible without algebra" (Moreno, 2011. Translated from Spanish). Doubtless, he is not referring to the clarity of the terms used in science to communicate information; yet an attitude of complacency can be observed with regard to the use of language hidden in the language of science.

This causes one to reflect on the fact that nowadays humanity enjoys many valuable resources and new technologies allowing us to acquire and disseminate information. These technologies include telecommunications, radio, television, telephony, data transmission, and interactive digital books, etc. Of course, the internet, with its bourgeoning social media applications, is the most outstanding of these new technologies, allowing human beings to enjoy access to global knowledge from mobile phones, tablets, laptops and televisions, etc.

This access has been possible for many years. So, how is it that science and technical knowledge are growing and spreading, but the world's population has not exploited it to improve its well-being? The answer is that there

is a sort of functional illiteracy[1] that does not allow a large segment of people to decode the meaning of many scientific dissemination messages, which is exacerbated by the use of terms that are unknown to the general population, something that should be addressed by actions such as those proposed by López-Barajas (2009) regarding virtual literacy.

The essence of the problem is patent when a population needs to understand certain terms, phrases or sentences contained in scientific literature in order to fully understand meanings, but fails to do so. Consequently, these persons are barred from the benefits of the scientific and technological knowledge, which appears to them as something impenetrable, obscure and distant, thereby reinforcing the notion that science is a secret language.

This is nothing out of the ordinary because scientific communication requires interlocutors to possess the basics of encoding, decoding, interpretation and transmission of scientific and technical messages. Those without these basic skills are automatically excluded from specialized communication between senders and receptors of such messages, as described by Sánchez González (2010). The problem is that the lack of clarity seems to be repeated in dissemination texts comprising the linguistic corpus of the general language. The *Corpus del Español Mexicano Contemporáneo 1921- 1974* (*CEMC*) provided the baseline for this study. This work contains scientific dissemination formal education texts at the undergraduate level in which specialized terms appear that are recovered from the mass corpus to be analyzed. This work does not address the semantic aspect of terms. It focuses exclusively on the methodology for teasing such terms out of the general language.

Focus on communication and other aspects of the text

Library and Information Sciences hold that in order to aid users to access information, the starting point must always be the notion that communication of a knowledge laden message depends entirely on grasping its mean-

1 Jiménez del Castillo defines this term thus: "The functional illiterate is a person who presented with information (or knowledge in alphabetic code) is incapable of putting it into operation through consequential actions. In this sense, we say that such a person does not possess the ability to process the information in a way that society to which he belongs expects" (2005: 290. Translated from Spanish). With regard to technology, Wikipedia provides this perspective: "The condition of functional illiteracy also seriously limits the quality of interactions a person can achieve with information and communications technologies, and their abilities to use efficiently a word processor, spread sheet, web navigator or mobile phone. ("Analfabetismo funcional", 2014.)

ing. In this light, one observes what happens with the linguistic sign and its components: i.e., signifier, referent and signified, in order to achieve effective communication.

Where texts are involved, as is the case with Library and Information Science, one must establish that there are several inherent facets to the need to communicate something, as explained by authors such as Luis Fernando Lara (1977, 1984, 1996, 1999, 2001; and Jetta Zahn, 1973), Ana María Cardero García (1998, 2003, 2004, 2005 and 2009) and Catalina Naumis Peña (1997, 1999, 2000 and 2003) in Mexico, and foregin specialists such as Juan Carlos Sager (1993), Juana Marinkovich (2008), Rosa Estopà (1998), Rita Temmerman's (2000) in the socio-cognitive theory of terminology; and María Teresa Cabré and Rosa Estopà (2002); and finally María Teresa Cabré (1999a, 1999b, 2002), who posited two proposal regarding the Doors Theory and the Theory of Communicative Terminology (TCT). These specialists argue that one must take into account the context in which a lexical unit is used and its correlation with the rest of the language in order to understand its true meaning, something which in itself is framed by the consensus of native speakers.

If the theories of these specialists are deemed pertinent, one must then admit socio-linguistic concepts, including situational context, field, tenor and mode (Halliday, 1979), along with quantitative urban sociology or *variationist theory*[2], which acknowledges the socio-economic and cultural status of the speaker. As such, the act of communication in scientific language will be shown to be comprised of spatial and temporal circumstances and this requires the linguistic context of the text, those factors associated with the production of an utterance, to be taken into account, for this context will affect the interpretation, propriety and meaning of the message, in terms of grammar, syntax, lexicon and context. One must also take into account the extra-linguistic context or situation, which is the subset of potential participants in the communication, including place, type of register and moment in which the linguistic act occurs.

The study and maintenance of linguistic registers is very important to the task of clarifying terms, because it includes the subset of contextual and socio-linguistic variables that modulate the mode in which the language is used in a concrete socio-linguistic instance. That is to say: analysis of a linguis-

2 "*Urban quantitative sociolinguistics* or *variationism* (this field studies the linguistic variations of a speaker or community of speakers in the context of social factors)". (*DTCE*, 2014. Italics in the original. Translated from Spanish).

tic register defines whether the communication is deployed as standard language or in a non-standard, cultured or sub-cultured usage; and whether it is formal or informal in nature (among other possible usages), as can be seen in *Corpus del Español Mexicano Contemporáneo, 1921-1974* (CEMC) (Lara and Ham Chande, 1979: 7-39), from which the results used in this study were obtained.

Similarly, when scientific texts are analyzed, it is important to indicate that science is kind of communication based on registers of formal use and situations, where the sender selects the appropriate linguistic resources, using specialized registers, targeted to a receptor whose nexus is shared interest in the specialized activity of a specific profession. These characteristics help differentiate it from the registers belonging to other socio-cultural fields such as the one studied in this case. A professional exchange is characterized by the use of its own technical vocabulary and expressions carrying special meaning, very often these messages are written. In real life, scientific authors, however, cannot in reality send a message as Wüster (2003) would have it in his General Theory of Terminology (GTT), by using only specialized terminology of the specific discipline; since they also need to employ lexical units from the general language and units belonging to other specialized disciplines.

This type of lexical analysis must select authors who are recognized as authorities in their respective fields. These authors shall be highly productive and widely cited. One must also take into account the author's place of birth, socio-economic status, life experiences, culture, ideology, religion, political leanings, verbal tradition, language, professional training, individual and team research, experience, freedom of expression, individual interests, current relevance, scientific specialty and type of documents or texts produced, which may include letters, memoranda, reports, degree theses, research reports, articles, books, conference scripts, resolutions, standards, laws, regulations or dissemination papers.

To situate this production of terms to be analyzed from a documentary standpoint, one must identify the type of document or scientific text, establish if it comes from an authority in the field and whether it represents spoken or written language, if it was written in a hurry or subject to several drafts or whether it was a free discourse or commissioned work to name only a few aspects to be weighed. One must also consider the use of specialized expressions, because scientific authors generally employ diction meticu-

lously in order to reduce ambiguity. The author, however, may or may not be successful in this endeavor, because the mind may have many reasons for choosing one lexicon or set of terminology over another. These motives can include situational factors entailed in drafting the discourse, the language used, the correct use of nomenclature, proper names, abbreviations, acronyms, signs, fixed idioms, codes, passwords, concepts, numbers (in numeral form and written out), symbols, formulas, conventions, etc.; all of which may or may not favor the deployment of a terminological unit in specialized texts. As can be seen, many factors can influence and author's word or term choices; since there are simple forms, syntagmas, fixed expressions or phrasal construction. Other types of information of greater proportion entailed in specialized academic and technical texts may also be added to all of this. This later type of information includes citations and transcriptions revealing the existence of a large number of mentions by others, whether in terms of thoughts or scientific proofs (Cunha, 2014). Many times these data appear in the original language, such as Latin, Greek, English and French, etc. and are included in the bibliography and/or footnotes.

Content Analysis

It is time to situate ourselves in the idea that in order to solve scientific problems, scientific research uses analytical techniques that are field specific. The field of Library and Information Science employs several techniques that are complementary or similar to analytic methods. Of course, there are many disciplines that could contribute to this matter, but the fields of Linguistics, Applied Linguistics and Computer Science are much closer to the point.

These fields, moreover, are often involved in multidisciplinary studies of, for example, content analysis, discourse analysis, grammatical analysis, qualitative analysis, analytic definitions analysis, contrastive phraseological analysis, lexicological analysis, document analysis, conceptual relationship analysis, analysis of texts, syntagma unit analysis, analysis and design of linguistic corpus, term analysis and, finally, the method of document content analysis used for sending information.

In the introduction to *La ciencia del texto: un enfoque interdisciplinario*, Teun A. Van Dijk explains to what degree discourse analysis uses an interdisciplinary "transversal connection." Van Dijk starts with the assumption that language interactions achieve communication and meaningful exchanges

through texts and discourses. Linguistics studies a part of language use, but other sciences do the same, including socio-linguistics, communications, cognitive psychology, pedagogy, jurisprudence, political science, sociology and, of course, Library Science. The textual and discursive relationships occur between diverse kinds of texts, the underlying textual structures, their diverse conditions and functions, contents and effects produced in the speakers (Van Dijk, 1992: 9-10).

The diverse types of texts and the relationships between them and society exhibit connections of diverse kinds in accord with the field from which they issue. Science examining texts is interested in grasping the common properties and characteristics in the use of language across the spectrum of fields comprising the social sciences and humanities.

The area of information analysis and systematization within Library and Information Science is manifested in the act of describing text types, data and informational content whereby such texts can be located in the systems. The use of processes in common with other disciplines, however, is undeniable. This study, in fact, shall use terminology and lexicographical analysis.

Documentation in Lexicography

Document analysis as presented by Rubio Liniers (2004) is also applied in the field of Lexicography, because it is a part of the process of creating dictionaries. Gómez González-Jover (2005) stresses that this method is indispensable in the task of representing content of documents comprising a corpus in which lexicographic units defined in the dictionary are included. The representation of contents performed allows users to consult and retrieve from diverse points of access. Moreover, the results of this type of analysis can almost always be used to create new lexical information products, such as concordances, statistical data, indexes and dictionaries.

Document content analysis aids user to decode messages and the retrieve relevant information from the document system of the *Diccionario del español de México* (*DEM*).[3] This situation is supported by the fact that the author has already emitted his message and it is contained in the documental sup-

3 This Project was launched in 1973. As stated by Barcala Rodríguez (2010) for other corpora, the *DEM* structured a system of lexicographic retrieval on the basis of the Linguistic Corpus.

port, largely in written texts belonging to the same specialization. As such, it is the job of information centers to ensure that the content of these documents, which may be be candidates for terms, are available to and easily retrievable by users.

The preparation of *DEM* rests on the basis of a Linguistic Corpus, which establishes the guidelines for maintaining and offering a great capacity and versatility in the management of the information it contains. Like any other information system, the Linguistic Corpus defines entries and access points that must be included. Even though nowadays there are multimodal corpuses (voice, image, text, etc.) the corpus generally used up until recently by science and technology has focused on the diverse modalities and features of words or lexicons contained in both general and specialized languages. In the case under study here, this is applied to support scientific dissemination communication.

The documental process in lexicography requires basic compliance with certain steps such as:

- Planning activities, entailing setting goals and objectives, and the organization and methodology to be implemented.
- The selection and acquisition of the documents, including making transcriptions of oral reports.
- Treatment of documents in terms of physical appearance, which implies physically preparing the material in order to obtain the corresponding file for later analysis.
- Drafting the bibliographic description of the document, highlighting the points of access allowing its identification with regard to other documents. For printed texts, this description includes author, title, legal notice and physical description of the material. The lexicography also includes external data of interest to the fields of socio-linguistic, pragmatics and semiotics, data which generally correspond to the communication unit analyzed in which the issuer, the situation from which the communication occurred and the channel used are highlighted. There is also an extra-linguistic context or register of speech by which formality or informality of written documents can be distinguished. This is also useful in understanding the target audience, i.e., whether it is for general readership or specialists. From the subset of register, the subsequent situational and thematic identifications with regard to the lexical units are performed. Consequently, this aids us-

ers of the system assign meanings to the lexical units contained in the information retrieved.

- With regard to the text or the strictly linguistic context, texts written in a scientific discipline generally must exhibit the components of the linguistic sign (signifier, signified and referent), with the smallest portion of text consisting of a paragraph separated by a period and return, or one item. Texts are analyzed by means of programs and algorithms previously determined for the purpose of gathering information from the document. In general, this analysis brings forth the graphic forms of words or lexical units, just as they are found in the texts of the natural language, whether common or specialized.

In Library and Information Science, where indexing is done with the natural language, the term is isolated from its context. A textual analysis approach is used[4] to analyze the scientific document and later a documental analysis is made of the content, with the indexing of natural language as the central objective, whereby the same text is used to extract indexing terms. This process supplies lists of signifiers or lexical units separated from their signifieds and referents. In this way the linguistic sign is fragmented, which complicates the user's retrieval efforts and necessitates additional search support.

In contrast to this method, the extraction of terms to constitute a linguistic corpus provides different lists, comprised of simple or compound words with their respective designation of part of speech, morphologies, internal structures, syllable divisions, placements, phraseological units, compound syntagmas, phraseological utterances, meaningful worlds, key words, vacuous words, technical terms, neologisms or term candidates.

In general terms lexical units obtained from a linguistic corpus are accompanied by quantitative data (range and frequency) and the contextual origin can be identified by means of usage register, especially when such units come from specialized language.

4 An *ad hoc* corpus can be made, or commercial text analysis programs, such as *WordSmith, AntConc, Notepad, Atlas.ti* and *Sketch Engine* can be used.

Terminological exclusion using general language subsets

When one wishes to extract term candidates from general or specialized texts, it is useful to keep in mind the information that already exists on the lexicon in infometric, bibliometric, scientometric and lexicometric studies. The Luhn cutoffs and the determination of TF-IDF weights (Blázquez Ochando, 2013) must also be considered in order to filter out common language usage and retrieve the term candidates

In addition to the indicator cited above, this article proposes that other similar indicators based on natural language can be employed to exclude subsets of general language. These indicators include fundamental vocabulary (similar to a frequency index and the Zipf model), common lexicons (based on the dispersion index) and lists of grammatical words (the equivalent of empty words) for the purpose of achieving maximum isolation of the specialized units searched for in the text. This is to say, the lexicographic knowledge produced by the *DEM* project and its *Corpus of Contemporary Mexican Spanish 1921-1974* (*CEMC*, 1975), can be reused to simplify the information one wishes to analyze.

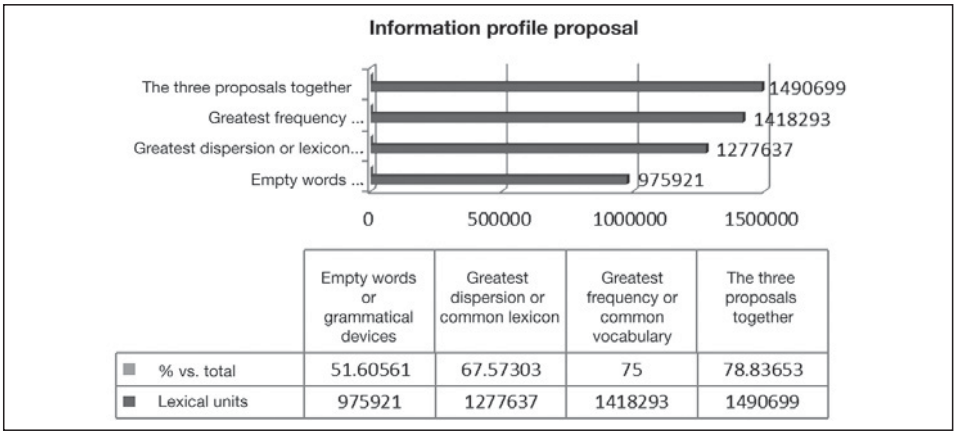
These pages use some of the results of the content analysis of *CEMC*, which is comprised of nearly two million grammatically labeled words. This corpus, in turn, supplied a lexicographical product, which properly speaking is a statistical index of natural language with lexical, grammatical and sociolinguistic information, as well as records of language usage and quantitative data. This product is called the *Diccionario estadístico del español de México* (*DEEM*, 2005). The results obtained from DEEM regarding empty words, greater dispersion and greater frequency are as follows:

1) Grammatical lexical units or *empty words*: These are largely articles, preposition, interjections, pronouns, etc., adding up to 292 lemmas and accounting for 51.60% of the total information in the corpus. This is the third group of terms to be excluded when one attempts to extract scientific and technical terms.
2) The lexical unit exhibiting greatest dispersion or the most *common lexicons* (Anguiano Peña, 2013a): These units consist of 994 distinct lemmas and account for 67.57% of the total corpus. Upon performing a search of specialized terms, these kinds of lexicons tend to be separated out from the document content analysis.

3) The lexical unit most frequently used or *fundamental vocabulary* presented by Lara (2007): In this category one must be aware of the phenomenon of economy of language known as "least effort," identified in lexicometric and infometric studies and the Zipf model (Zipf, 1949), among others. This concept describes how people use an enormous number of graphic words that correspond to a small number of lemmas, which results in a very small number of lexical units with very high frequency of use. In line with this reasoning, we understand that fundamental vocabulary, i.e., that which is most frequent, is the vocabulary most often exploited in texts and discourses. In the *CEMC,* 861 lemmas account for 75% of the total of the information of the corpus. We suggest that this type of lexical unit also be eliminated.

Figure 1 and *Table 1* show the significant savings secured over these three headings, which is why they were excluded from the analysis.

Figure 1. Proposed cutoffs: empty words, most frequent, most dispersed and the three together, against 1 891 058 lexical units (%) Source: Created by Gilberto Anguiano Peña for doctoral thesis (2015)



Source: Gilberto Anguiano Peña for PhD dissertation (2015)

Table 1. Summary of Empty Words and lexicons with greatest frequency and dispersion, and all three together

Concept	Graphic words	% of total
Empty or grammatical words	975 921	51.60561
Greatest dispersion or common lexicon	1 277 637	67.57303
Greatest frequency or fundamental vocabulary	1 418 293	75
The three proposal together	1 490 699	78.83653

Source: Gilberto Anguiano Peña for PhD dissertation (2015)

As can be observed, the value of three groups taken together is not the sum of the three groups, because there are lexical units that belong to two and even three groups. Because the combined value stands at 78.83% of the global information analyzed, this recommends the implementation of filters using general language before attempting an analysis of content. In this way, there is a savings of nearly 80% in the retrieval of term candidates, which is in line with calculations presented in other information retrieval studies. Moreover, in order to retrieve scientific and technical terms more efficiently, a minimum number of valid appearances is set. In this way, those lexical units that appear very infrequently and with scant literary warrant are filtered out.

Documental process of disambiguation of meanings and defining the use of candidate terms

This paper proposes the following process for retrieving a text in which a user is interested: After the index of signifiers is obtained, which is the same as terminology candidate list, these are simplified and lemmatized. Then one must retrieve each as per the use register within the thematic area the document analyzed belongs to and in which it was documented. In practice this would be something like signaling the lexicon available in the text.[5] In this way, the user receives help in "clearing up ambiguity of meaning and finding the proper use of certain voices" (Estopà, 1998: 360. Translated from Spanish), and the user may subsequently request the information retrieval system to provide the referent closest to the search query by simplifying the search to a minimum. Any such effort notwithstanding, the real meaning shall always depend on the reader's interpretation.

Much like the Library Science indexing process, the term candidates or key words can be adjusted to a controlled language in order to improve content retrieval. This can be done through the use of subject headings or thesauruses. In this way the words of the natural language are converted from the indexation of expressions and concepts acquired from a controlled language. At the end of the documental process, information is released to the users so they might appropriate it. Documental analysis in lexicographic

5 In the view of López Morales "The *available* lexicon is the subset of words speakers possess as mental lexicon and whose use is contingent upon the concrete topic of the communication. What we want to know is which words a speaker would be capable of using in certain themes of communication" (2013. Italic in original).

projects supplies diverse informative products which are targeted to internal and external users. These may be separate or combined as a system. The components can be a data base of concordances, like the *Key Word in Context* (KWIC), quantitative information, document card catalogues, the dictionary being prepared or the distinct interfaces for consulting lexicographical information.

As part of the long process of documentary text content analysis from the lexicographical standpoint, what one expects to obtain after concluding the indexing and classification by natural language is a list of lexical units that are meaningful in both the general language and the scientific-technical sociolect. This can be expected because of the inclusion of texts from such specialized fields.

Exploitation of use markings from lexicographical documentation

The *DEEM* results provide the basis to build another data base: the *Sociolinguistic lexicon model of Spanish used in Mexico* (Anguiano Peña, 2006). After assigning a semi-automated index to the lexical units of the *DEEM*, a summary of partial results of the previous data base was possible; and with the completed data, the total results of the lexical unit used in the general language in Mexico could be identified by means of their socio-linguistic registers (*Table 2*).

Table 2. Sample record used for identifying term candidates in the sociolinguistic model

Lemmas	Part of speech	Total Frequency	% total	Use of Spanish	Language level	Speech registers	Most frequent	Best distribution	Text key	Use registry 1	Use registry 2	Use registry 3
acción [action]	nom	4	0.00021	standard	cultured							
actitud [attitude]	n	259	0.01370	standard			basic vocabulary					
activación [activation]	n	14	0.00074	standard	cultured	science			420, 427, 428, 454, 469, 473, 477, 478	Chemistry	Medicine and veterinary med	Medicine
activado [activated]	adj	2	0.00011	standard	cultured	science			389, 478	Electronics and electricity	Medicine	
activamente [actively]	adv	12	0.00063	standard	cultured							
actividad [activity]	n	511	0.02701	standard			basic vocabulary					
activista [activist]	adj; n	6	0.00031	standard	cultured							
acto [act]	n	308	0.01629	standard			basic vocabulary					
actor [actor]	adj; n	133	0.00704	standard								

Source: Gilberto Anguiano Peña for PhD dissertation (2015)

Proposal for limiting term candidates

For the purpose of specialized information search and retrieval and on the basis of previous analysis of documental content of general and specialized texts, we propose eliminating the following data originating in the quantitative data and use marks of the general language:

- The most frequently used lexical units
- The most widely dispersed lexical units
- Lexical units belonging to the empty words group
- Lexical units in a non-standard language
- Lexical units deemed uncouth

If this list of quantitative and socio-linguistic lexical units are filtered out of the analysis, the process of retrieving term candidates can be streamlined considerably. What is important is that once a list of such elements is secured it can be compared to the registers of the use of language that already exists in the S*ocio-linguistic model of the lexicon of Mexican Spanish.* This comparison will help both the information user and the Library and Information Science professional to reconstruct the meaning of the linguistic sign and the creation of a controlled language. This new comparison can find term candidates used exclusively in a field, which would prove they are key words and could become terms in the strict sense after expert verification. Consequently, candidates used in two or more fields can be recognized, indicating they are terms in the broad sense and in fact may have multiple meanings, which in lexicography means that are technical terms. We may also find candidates that belong to science that are also technical terms. These could be considered technical terms, but dictionary entries may include the label "Scientif." indicating they belong to scientific language.

This paper also proposes the reuse of lexicographic processes to differentiate lexical units and extract these through content analysis of specialized texts, employing usage markings or speech registers as posited by Josette Rey-Debove (1971), when she examined three fundamental aspects for achieving this goal:

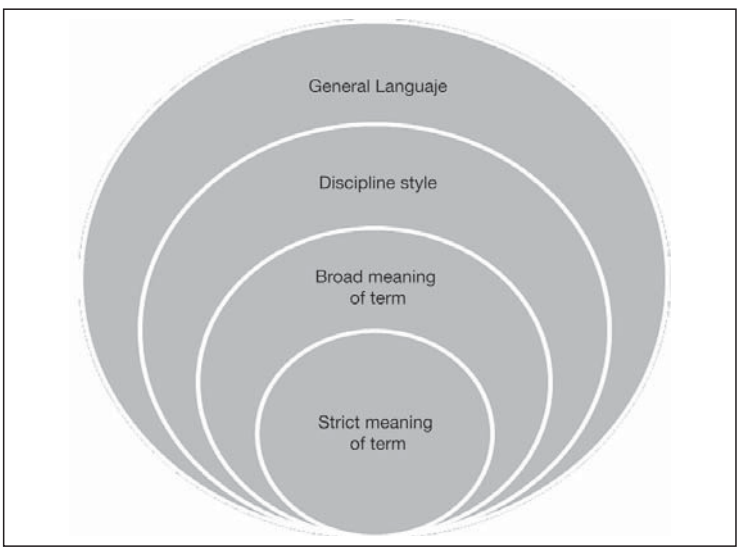
1) The subset of words (lexical units) belonging to a language or dialect.
2) The socio-linguistic information of the lexical units.
3) The consensual usage markings made by the community of speakers.

The aim of incorporating these guidelines in the information analysis is to have the same lexical units of the general language, identified by consensus and standing in contrast to the specialized language, serve, in the first place, to classify technical terms.[6] Since these terms behave similarly to terminological units, they can be designated term candidates.

Text search of terminological units

To obtain specialized terms with the help of a corpus such as the aforementioned model, one must first separate the term candidates with registers of usage in a specialized text. During this stage of the term search process it is common to find lexical units in the lists produced by automated analysis that belong to a discipline across several of its communication levels, even though all these units belong to the standard language, to cultured usage and a to a science or technical field. This means that the content analysis can provide the following lexical units from a general or scientific text: 1) Units that belong to the general language. 2) Units belonging to the style of the field under analysis. 3) Term candidates in the broad sense, and 4) Term candidates in the strict sense, as shown in *Figure 2*.

Figure 2. Lexical units in the documental content analysis of a text



Source: Gilberto Anguiano Peña for PhD dissertation (2015)

6 From the standpoint of linguistics, a term is deemed "technical" in accord with the following definition: "Technical terms. *n.* 1 Any term that has a concrete and specific meaning within the language of a trade, science, art or industry: the word "algorithm" is a technical terms from the field of mathematics." (*DMLE*, 2007. Translated from Spanish.)

In order to take a deeper look at the proposals made in the previous paragraph, the following observations are in order:

1) While also identified socio-linguistically as belonging to standard usage, non-standard usage, uncouth usage (vulgate) and cultured usage –which is to say they are not exclusive to science or technology– the lexical units that belong to the general language and also appearing in scientific and technical texts should be excluded from the term candidate list.
2) Lexical units belonging to the characteristic writing style of the field under analysis. These units generally belong to the verbal tradition of the field and consist of stock phrases and utterances. They appear at a in the text analyzed at very low frequency, but they are characteristic features in certain scientific fields. As such, it is best not to exclude them before performing the content analysis. These units may include set phrases and Latinisms, etc.
3) Specialized lexical units or technical terms. These are used with very narrow meanings in the field to which the text under analysis belongs. Such units may have the same signifier in general usage or in other fields; which is to say, they can have synonyms. This type of lexical unit will have an entry in the dictionary of general language,[7] and are in fact terms in the board sense.[8] The forms of graphic words, as they appear in the original text, are generally few and include the cases of masculine, feminine, singular and plural. Documental content analysis finds these words appear very infrequently in the common language, but as lemmatized lexical units (words grouped under their canonical form) they attain an additional percentage above the total of the sample analyzed. In other words, a small number of lexical units are grouped in a high number of lemmas. As for the dispersion index in *DEEM*, we observe that while such units are concentrated in a given field, they may appear in other scientific or technical fields or otherwise belong to scientific language bridging both areas of knowledge. These can be recognized, because even while having an acknowledged signifier, they have a meaning distinct from that ascribed in the natural language. As such, the average reader does not grasp the meaning and the term will seem obscure or secret. These

7 Such as in *DRAE* (2001) or *DEM* (2012).

8 In this study the term is used in its broad sense. Cardero García's proposal (2004: 42-43. Translated from Spanish), contained in a work on the control of satellites, argues that technical terms are "[…] designations from the general language that specialize their signified or designations that are common across several areas of knowledge […]". This would correspond to an infrequent signified with a frequent signifier.

units can appear in simple form of in phrases as syntagmas, set phrases or phraseological units

4) Term candidates in field under study. In documental behavior, these units are very similar to technical terms, but they do not have synonyms and presuppose a single, unequivocal meaning. These units belong to standard usage and are cultured. They are used exclusively in science or technology and have a formal register for exclusive use in the specialization. As such, they do not have meaning or equivalent in the common language. These candidates may take the form of simple or compound lexical units. In principle, candidates may be considered key words and once validated by an information specialist they may become part of the documental language. In the best of cases, they may become terms of a field in the strict sense.[9] In text analysis their frequency of appearance is low, but when these lexical units are grouped they have a high percentage of lemmas. They are without dispersion because their data are concentrated in a single filed.

In view of these considerations and the proposal of Cardero García (2004: 37), we can also expect that any documental content analysis of a general or specialized text will very likely exhibit lexical units with the features (in terms of signified, signified and communication type) shown in *Table 3*.

Table 3. Features of lexical units analyzed

Signifier*	Signified**	Language type
A common signifier	and a common signified	These are part of common usage.
An uncommon signifier	and a common signified	This would be a technical term of signifier, for example: *close up, stock shot, feidear.*
A common signifier	with an uncommon signified	This is a technical term in the broad sense; for example, bobbin winder, optical, camera.
An uncommon signifier	and an uncommon signified	This would be a technical term in the strict sense; for example, magnetic eraser, projection lamp, translucent screen system, animation technique.

* Signifier is that which indicates something. In this study it is a word or lexical unit given to a person, animal, thing, or tangible or intangible concept, and/or concrete or abstract object for the purpose of distinguishing it from others.

**Signified is the object indicated by the signifier. For our purposes, it is the mental concept or representation of something.

Source: Gilberto Anguiano Peña for his PhD dissertation (2015)

9 We can take also what is proposed by Cardero García (2004: 43), who views strict a term as belonging exclusively to a single discipline, and consequently an infrequent signified and signifier.

Despite the coexistence of lexical units and terminological units in a scientific text, it is possible to differentiate these by examining speech register in order to tell whether it exists in a form of communication or in a text belonging exclusively to a specialized language; that is, by verifying that they are products of formal communication used by specialists of a given field in order to ensure effective communication among them. As can be observed in the description of the process performed and described in the previous paragraphs, the lexical units analyzed are drawn from an empirical lexicographical study, which shows that something similar to that which occurs with general language texts also occurs with texts from the specialized language, in that both types of texts are comprised largely of lexical units from the general language. Even though it may seem to be the contrary, these differences are actually useful in the task of information retrieval, because the terms one wishes to extract from texts are not part of the common language.

Illustration of this type of analysis performed with the Model

In accord with the proposal herein and by isolating the lemmas corresponding to science and technical fields contained in the Model, the following results were obtained:

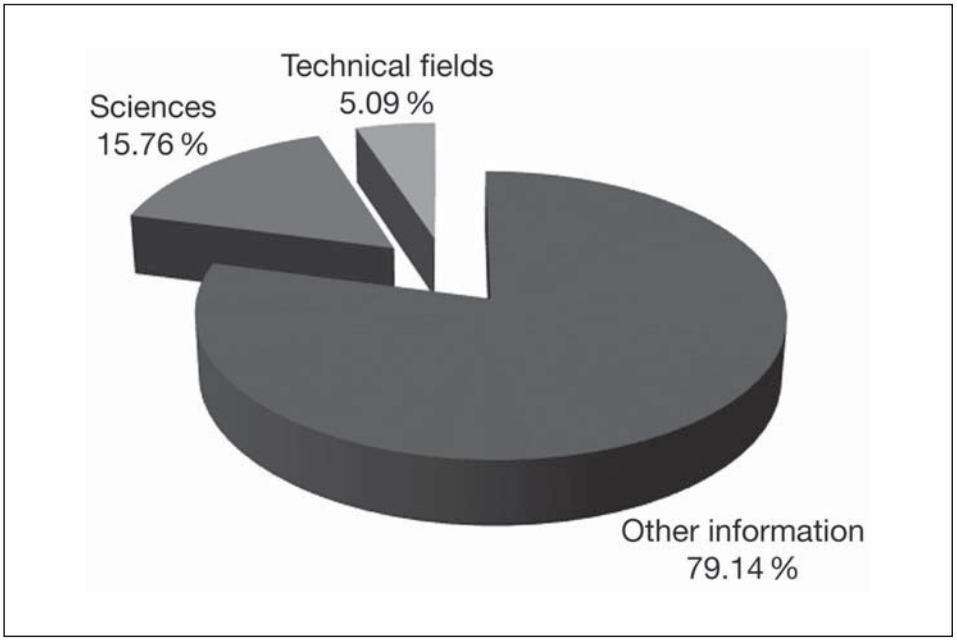


Chart 1. Of a total of 30, 899 lemmas (words that could be headings in a dictionary entry with associated definition) assigned in *CEMC*: 4,871 science lemmas and 1,574 technical lemmas were retrieved as terms candidates.

The *Chart 1* is derived from 30,899 lemmas in general use. To obtain the corpus of 15.76 % of terms exclusive to science, the lemmas were restricted twice (See subsection 4 of the Section: "Text search of terminological units"). First the 30,899 lemmas were reduced to 16,296 lemmas within the field of science and trades. From this number, the lemmas exclusively used in science and technical trades were drawn. The 6,450 lemmas used exclusively in these areas constitute 20.85 % of the corpus of lemmas.

Final considerations

The Model shown here, or other lexicographic resources with similar features, can be useful in the near future for computer assisted indexing or as a corpus monitoring resource in new analyses of specialized texts or of a corpus. Its use will allow rapid generation of signifier term candidates, which can also be useful for representing and retrieving content from the original text. These will also be valuable in the development stage of controlled language when working on terms, uniterms, subject headings or descriptors comprising the terminology of a given discipline analyzed in this way.

Moreover, it is important to understand that natural language and specialized language are constantly evolving, which makes it difficult to control and retrieve specialized language and associated terminology; but this is all the more reason for Library and Information Science to be involved in aiding users and readers decode the language of science.

Works cited

Aguilar, C. A.; Alarcón, Rodrigo; Rodríguez, Carlos and Sierra Martínez, Gerardo (2006), "Reconocimiento y clasificación de patrones verbales definitorios en corpus especializados." In *La terminología en el siglo XXI: contribución a la cultura de la paz, la diversidad y la sostenibilidad*, edited by María Teresa Cabré, Rosa Estopà and Carles Tebé Soriano, 259-269. Barcelona: Institut Universitari de Lingüística Aplicada-Universidad Pompeu Fabra.

"Analfabetismo funcional" (2014). *Wikipedia. La enciclopedia libre.* Last modified: October 23, 2014. http://es.wikipedia.org/w/index.php?title=Analfabetismo_funcional&oldid=77706574

Anguiano Peña, Gilberto (2006), *Modelo sociolingüístico del léxico del español usado en México.* México: El Colegio de México, Centro de Estudios Lingüísticos y Literarios, *Diccionario del español de México.* [Unpublished paper]

—— (2013a), *El léxico común del español de México.* México: El Colegio de México, Centro de Estudios Lingüísticos y Literarios, *Diccionario del español de México.* [Unpublished paper]

—— (2013b), *Palabras vacías del español de México.* México: El Colegio de México, Centro de Estudios Lingüísticos y Literarios, *Diccionario del español de México.* [Unpublished paper]

Barcala Rodríguez, Francisco Mario (2010), "Corpus lingüísticos estruturados de grandes dimensións: Metodoloxía e sistemas de recuperación de información." PhD diss., Universidade da Coruña, Departamento de Computación. http://ruc.udc.es/dspace/bitstream/2183/7171/1/tese_mario_barcala.pdf

Blázquez Ochando, Manuel (2013), *Técnicas avanzadas de recuperación de información: procesos, técnicas y métodos.* Madrid: Universidad Complutense de Madrid. http://mblazquez.es/wp-content/uploads/ebook-mbo-tecnicas-avanzadas-recuperacion-informacion1.pdf

Bogomilova Lozanova, Elena, (2009), "Posibilidades y límites del análisis cuantitativo de corpus especializados." In *Memoria del I Simposio Internacional sobre Organización del Conocimiento: Bibliotecología y Terminología,* compiled by Catalina Naumis Peña, 63-78. México: UNAM, CUIB.

Cabré, María Teresa (1999a), "La terminología hoy: concepciones, tendencias y aplicaciones." In *La terminología: representación y comunicación: elementos para una teoría de base comunicativa y otros artículos,* 17-37. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada.

—— (1999b), "La terminología y documentación." In *La terminología: representación y comunicación: elementos para una teoría de base comunicativa y otros artículos,* 231-247. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada.

—— (2002), "Textos especializados y unidades de conocimiento: metodología y tipologización." In *Texto, terminología y traducción*, edited by Joaquín García Palacios and M. Teresa Fuentes, 15-36. Salamanca: Ediciones Almar. http://www.upf.edu/pdi/dtf/teresa.cabre/docums/ca02te.pdf

Cabré, María Teresa, and Estopà, Rosa (2002), "El conocimiento especializado y sus unidades de representación: diversidad cognitiva." *Sendébar* 13:141-153.

Cardero García, Ana María (1996), "La integración del corpus de la terminología de control de satélites en México." In *Actas del V Simposio Iberoamericano de Terminología*, 106-111. México: UNAM.

—— (1998), "Algunas observaciones de los conceptos, sus áreas temáticas. La sinonimia y la polisemia en tres vocabularios especializados en México." In *Actas del VI Simposio Iberoamericano de Terminología*, 137-154. La Habana.

Cardero García, Ana María (2003), "Unidad y variedad del español de América. Los vocabularios especializados." In *Estudios de lingüística y filología hispánicas en honor de José G. Moreno de Alba,* compiled by Ignacio Guzmán Betancourt and María del Pilar Máynez Vidal, 299-322. México: UNAM.

—— (2004), *Lingüística y terminología.* México: UNAM, Facultad de Estudios Superiores Acatlán.

—— (2005), "Algunas características lingüísticas de las denominaciones de una terminología." *Lingüística Mexicana* 2 (1): 141-152.

—— (2009), "El descriptor y el término. Los conceptos y la lingüística." In *Memoria del I Simposio Internacional sobre Organización del Conocimiento*: *Bibliotecología y Terminología*, compiled by Catalina Naumis Peña, 53-62. México: UNAM, CUIB.

CEMC (*Corpus del Español Mexicano Contemporáneo, 1921-1974*) (1975), María Isabel García Hidalgo, Luis Fernando Lara, Roberto Ham Chande *et al.* México: *Diccionario del español de México.* [Unpublished paper]

—— (*Corpus del Español Mexicano Contemporáneo, 1921-1974. Lematizado)* (2005), by Gilberto Anguiano Peña, Francisco Segovia and Erika Flores García. México: El Colegio de México; UNAM, Instituto de Ingeniería. www.corpus.UNAM.mx/cemc/

Cunha, Iria de (2014), "Análisis discursivo, textos especializados y traducción." In *Manual de traducción de textos especializados. Nuevos enfoques, nuevas metodologías*, edited by Marisela Colín, 32-45. México: UNAM.

DEEM (Diccionario estadístico del español de México. Lematizado) (2005), edited by Gilberto Anguiano Peña, Francisco Segovia and Erika Flores. México: El Colegio de México, Centro de Estudios Lingüísticos y Literarios, *Diccionario del español de México*. [Unpublished paper]

DEM (Diccionario del español de México) (2012). México: El Colegio de México, Centro de Estudios Lingüísticos y Literarios. http://dem.colmex.mx/moduls/Default.aspx?id=8

DMLE (Diccionario Manual de la Lengua Española Vox) (2007). Larousse Editorial. http://es.thefreedictionary.com/tecnicismo

DTCE (*Diccionario de términos clave de ELE)* (2014). España: Biblioteca Virtual Cervantes. http://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/sociolinguistica.htm

DRAE (Diccionario de la lengua española) (2001). 22nd edition. España: Real Academia Española; Espasa Calpe.

Estopà, Rosa (1998), "El léxico especializado en los diccionarios de lengua general: las marcas temáticas." *Revista Española de Lingüística* 28 (2): 359-387.

Faber Benítez, P.; Moreno Ortiz, A., and Pérez Hernández, C. (1998), *Lexicografía Computacional y Lexicografía de Corpus*. http://www.ontoterm.com

Gómez González-Jover, Adelina (2005), "Terminografía, lenguajes profesionales y mediación interlingüística. Aplicación metodológica al léxico especializado del sector industrial del calzado y de las industrias afines." PhD diss., España: Universidad de Alicante, Facultad de Filosofía y Letras, Departamento de Filología Inglesa. http://rua.ua.es/dspace/bitstream/10045/760/1/tesis_doltoral_adelina_gomez.pdf

Halliday, M. A. K. (1979), *El lenguaje como semiótica social*. México: FCE.

Jiménez del Castillo, Juan (2005), "Redefinición del analfabetismo: El analfabetismo funcional." *Revista Educación* 338: 273-294. http://www.revistaeducacion.mec.es/re338/re338_17.pdf

Lázaro Hernández, Jorge Adrián (2010), "Extracción de la terminología básica de las sexualidades en México a partir de un corpus lingüístico." Thesis. México: UNAM. [Unpublished paper]

Lara, Luis Fernando (1977), "Una base semántica para la lexicografía: la conceptualización del signo lingüístico." *Nueva Revista de Filología Hispánica* 26 (2): 261-275.

—— (1984), "Una caracterización lingüística del discurso científico mexicano." *Discurso: Cuadernos de Teoría y Análisis* 2: 33-42.

—— (1996), "Conocimiento y pragmática en los fundamentos de la semántica." *Estudios de Lingüística Aplicada* 23-24: 236-243.

—— (1999), "Término y cultura: hacia una teoría del signo especializado." In *Terminología y modelos culturales*, edited by en María Teresa Cabré, 39-60. Barcelona: Institut Universitari de Lingüística Aplicada-Universitat Pompeu Fabra.

—— (2001), *Ensayos de teoría semántica: lengua natural y lenguajes científicos.* México: El Colegio de México.

—— (2007), *Resultados numéricos del vocabulario fundamental del español de México.* México: El Colegio de México. http://dem.colmex.mx/moduls/Default.aspx?id=14

Lara, Luis Fernando, and Ham Chande, Roberto (1979), "Base estadística del Diccionario del Español de México." In *Investigaciones lingüísticas en lexicografía*, Luis Fernando Lara, Roberto Ham Chande and María Isabel García Hidalgo, 7-39. México: El Colegio de México.

Lara, Luis Fernando, and Zahn, Jetta (1973), "El tecnicismo en el léxico del español mexicano. Posiciones posibles del DEM." In *Monografías generales del DEM.* México: El Diccionario del Español de México.

López-Barajas, Emilio (2009), "Alfabetización virtual y gestión del conocimiento." *Revista Electrónica Teoría de la Educación. Educación y Cultura en la Sociedad de la Información* 10 (2). http://www.usal.es/teoriaeducacion

López Morales, Humberto (2013), "¿Qué es la disponibilidad léxica?" *DispoLex: investigación léxica*. http://www.dispolex.com/info/la-disponibilidad-lexica

Marinkovich, Juana (2008), "Palabra y término: ¿Diferenciación o complementación?" *Revista Signos* 41 (67). http://www.scielo.cl/scielo.php?pid=S0718-09342008000200002&script=sci_arttext

Medina Urrea, Alfonso, and Méndez Cruz, Carlos (2006), "Arquitectura del corpus histórico del español de México (CHEM)." In *Avances en la ciencia de la computación,* edited by A. Hernández and José Luis Zechinelli Martini, 248-253. México: Sociedad Mexicana de Ciencias de la Computación.

Moreno, María José (2011), *El lenguaje secreto de la ciencia.* http://ababol.laverdad.es/ciencia-y-salud/2985-el-lenguaje-secreto-de-la-ciencia

Naumis Peña, Catalina (1997), "Reconocimiento semi-automático de patrones temáticos y adaptación del lenguaje documental para mejorar la eficiencia en la recuperación del sistema INFOBILA." In *Primer Congreso Interno de la Comunidad Científica del CUIB: los investigadores y sus investigaciones,* 23-27. México: UNAM, CUIB.

—— (1999), *Tesauro latinoamericano en ciencia bibliotecológica y de la información. TELACIBIN.* México: UNAM, CUIB.

—— (2000), "Análisis de la confluencia entre término y descriptor en la elaboración de tesauros." In *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información* 14 (29): 95-113.

—— (2003), "Indización y clasificación: un problema conceptual y terminológico [Indexation and classification: a conceptual and terminologic problema]." *Documentación de las ciencias de la información* 26: 23-40. http://www.ucm.es/BUCM/revistas/inf/02104210/articulos/DCIN0303110023A.PDF

Rey-Debove, Josette (1971), *Étude linguistique et sémiotique des dictionnaires français contemporains.* París: Mouton the Hague.

Rubio Liniers, María Cruz (2004), "El análisis documental: indización y resumen en bases de datos especializadas." In *E-LIS: E-prints in Library and Information Science.* http://www.iberius.org/es/AisManager?Action=ViewDoc&Location=getdocs:///DocMapCSDOCS.dPortal/2519

Sager, Juan C. (1993), *Curso práctico sobre el procesamiento de la terminología.* Madrid: Fundación Germán Sánchez Ruipérez, Ediciones Pirámide.

Sánchez González, Aránzazu (2010), *El proceso de comunicación.* http://zazu897.blogspot.com/2010/10/el-proceso-de-comunicacion.html

Temmerman, Rita (2000), *Towards New Ways of Terminology Description: The Sociocognitive Approach.* Amsterdam, Philadelphia: John Benjamins.

Van Dijk, Teun A. (1992), *La ciencia del texto: un enfoque interdisciplinario*. 3rd edition. Barcelona: Paidós.

Wüster, Eugen (2003) [1998], *Introducción a la teoría general de la terminología y a la lexicografía terminológica,* edited by María Teresa Cabré. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada.

Zipf, George Kingsley (1949), *Human behavior and the principle of least effort.* Oxford: Addison-Wesley Press.

To cite this article as an online journal:

Anguiano Peña, Gilberto y Catalina Naumis Peña. 2015. "Extracción de candidatos a términos de un corpus de la lengua general". *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información*. 67: 19-45. [Include URL here] Consulted on: [Include date here]

To cite this article from an information service:

Anguiano Peña, Gilberto y Catalina Naumis Peña. 2015. "Extracción de candidatos a términos de un corpus de la lengua general". *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información*. 67: 19-45. In: [Include name of information service and URL] Consulted on: [Include date here]