Method for Selecting Specialized Terms from a General Language Corpus

Gilberto Anguiano Peña* and Catalina Naumis Peña**

*Diccionario del Español de México, El Colegio de México, Camino al Ajusco #20 Pedregal de Santa Teresa, Tlalpan, C. P. 10740, México D. F., <ganguia@colmex.mx>
**Instituto de Investigaciones Bibliotecológicas y de la Información, Universidad Nacional Autónoma de México, Piso 12, Torre II de Humanidades, Avda Universidad 3000, Ciudad Universitaria, Delegación Coyoacán, C. P. 04510, México D. F., naumis@unam.mx

Gilberto Anguiano Peña, is a doctoral candidate at the National Autonomous University of Mexico (UNAM). His undergraduate thesis (UNAM 1991) was *La relevancia de la información bibliográfica en la documentación de un diccionario* (The Relevance of Bibiographic Information in Dictionary Documentation), and his master's (UNAM 2007) *Indización semiautomática para almacenar y recuperar información del léxico del español usado en México* (Semiautomatic Indexing for Storing and Retrieving Information from the Spanish Language Used in Mexico). Since 1992, he has been project researcher for the *Diccionario del Español de México* (Dictionary of Spanish in Mexico) at El Colegio de México.

Catalina Naumis Peña has undergraduate and master's degrees from the National Autonomous University of Mexico (UNAM), and a PhD in Information Sciences (Documentation) from the Complutense University of Madrid. Specializes in analysis and content representation within the area of information and knowledge organization, at the UNAM Library and Information Science Research Institute. She also teaches in the UNAM Faculty of Philosophy and Literature. Her most extensive projects have been the *Tesauro Latinoamericano en Ciencia Bibliotecológica* (Latin American Thesaurus in Library Science and Information) and the *Macrotesauro mexicano para contenidos educativos* (Mexican Macrothesaurus for educational content).

Anguiano Peña, Gilberto and Naumis Peña, Catalina. **Method for Selecting Specialized Terms from a General Language Corpus.** *Knowledge Organization. 42(3)*, 164-175. 27 references.

Abstract: Among the many aspects studied by library and information science are linguistic phenomena associated with document content analysis, for purposes of both information organization and retrieval. To this end, terms used in scientific and technical language must be recovered and their area of domain and behavior studied. Through language, society controls the knowledge available to people. Document content analysis, in this case of scientific texts, facilitates gathering knowledge of lexical units and their major applications and separating such specialized terms from the general language, to create indexing languages. The model presented here or other lexicographic resources with similar characteristics may be useful in the near future, in computer-assisted indexing or as corpora monitors, with respect to new text analyses or specialized corpora. Thus, using techniques for document content analysis of a lexicographically labeled general language corpus proposed herein, components which enable the extraction of lexical units from specialized language may be obtained and characterized.

Received: 19 March 2015; Revised: 2 July 2015; Accepted: 3 July 2015

Keywords: language units, language corpus, lexical analysis, specialized terms

1.0 Introduction

The overall goal of this paper is to determine methodologies and strategies for processing a linguistic corpus from the general language in order to extract specialized terms from one discipline and those shared by several so they may be automatically separated from the mass of lexical units in the general language. This type of work makes it possible to gather terms and later clarify their meaning, learn about their uses in the texts in which they

appear and utilize them when constructing indexing languages. Alexiev (2006, 96) wrote:

> Terminologies come under various forms (indexed, thesauri, termbanks, specialized dictionaries, glossaries, etc.) and are designed to meet the needs of translation, Language for Specific Purposes (LSP) teaching, information retrieval, controlled indexing, document consulting and navigation, technical authoring, or merely to help the understanding of technical documents.

Is the language of science a secret language? The author of a published scientific article titled "The Secret Language of Science" (Ryan 1985, 91) says that:

> Too many of the students enrolled in introductory courses at colleges and universities do not understand the concepts being discussed even though they make a conscientious effort to study and to master the technical vocabulary … the secret language of science consists of the many common words that have been appropriated into biology, chemistry, physics, psychology, and even mathematics, with highly specific technical meanings.

Essentially, the problem becomes evident when people try unsuccessfully to understand words, phrases, sentences, etc. found in scientific texts in order to interpret their meaning. Consequently, they are unable to benefit from scientific and technological knowledge, the outcome being that the code of the sciences is opaque, dark and foreign to most of the population, reinforcing the idea that the language of science is practically a secret language. This is actually quite understandable, because individuals engaged in scientific communication need to have the basic elements to codify, transmit, decode and interpret the scientific or technical meanings, and anyone lacking those elements is automatically excluded from the specialized communication held between senders and receivers of science and technology, as discussed in Gemma (2013).

Knowledge organization should clarify terms so as to help users in their information searches. The key words, descriptors or subject headings used to index contents are validated terms from texts written about each subject. Thus, the corpus of a general language dictionary of the Spanish used in Mexico can also be a source for extracting indexing term candidates. The problem is that the lack of clarity would seem to be repeated in texts that are involved in forming the linguistic corpora from the general language. The corpus called *Corpus del Español Mexicano Contemporáneo* (*CEMC*, Corpus of Contemporary Mexican Spanish) has been utilized as the basis for developing this study. It has incorporated scientific, as well as formal, undergraduate-level education texts in which terms from the specialized language appear that are retrieved from the corpus mass in order to be analyzed. This paper does not touch on semantic aspects of the terms, only the methodology for isolating them from the general language corpus. According to Alexiev (2006, 14), "The concept <corpus> is used in modern linguistics to refer to both running text and lists of lexical items excerpted from running text for various, including terminographic, purposes."

2.0 Communication Focus and Other Aspects of the Text

Information studies currently considers, among other things, that to help users access information, the starting point is to make clear that the main idea of the communication of a message is for its meaning to be understood, so it is necessary to observe what happens with the linguistic sign and its components: signifier, referent and meaning, for effective communication to exist. When working with texts, as is usually the case in information studies, it must be established that there are several aspects inherent to the need to communicate something, as explained by such authors as Lara (2001), Temmerman (2000), with her socio-cognitive theory of terminology, and Cabré (1999; 2003), with her two propositions: the Theory of Doors and the Communicative Theory of Terminology (TCT, Spanish initials). These specialists argue that the context in a lexical unit is used and its correlation with the rest of the language must be taken into account in order to understand its true meaning, which, in turn, is designated by common consensus among speakers.

If the theoretical approaches of these specialists are considered pertinent, it would also be important to examine those of sociolinguistics concerning the context of situation, field, tenor and mode (Halliday 1977) and of quantitative urban sociology or variationism (DTCE 2014), which looks at the speaker's socioeconomic position, as well as cultural background. This will make it clear that scientific language communication encompasses the spatial and temporal circumstances in which it develops. So, studying the object called text must also take into account linguistic context, meaning the factors linked to sentence production. The same context affects the interpretation, adaptation and meaning of the message (through grammar, syntax, vocabulary and context). Furthermore, the context or extra linguistic situation must also be considered, as it is the set of potential participants in the communication, such as the place, type of registry and moment when a linguistic act occurs.

The study and maintenance of linguistic registries is extremely important for clarifying terms, since it includes the

set of contextual, sociolinguistic variables that condition the way in which a language is used in a concrete socio-economic context. In other words, analysis of a linguistic registry must define whether the communiqué is in standard, non-standard, proper or semi-proper language, if it is a formal or informal communication, etc., as established in the *CEMC* stratification (Lara and Ham 1979, 7-49) from which the results for this article came.

Likewise, when scientific texts are analyzed, it is important to indicate that science is a type of communication based on registries of use and formal situations where the sender chooses suitable linguistic resources, in specialized registries, aimed at a receiver whose common link is an interest in a specific or professional specialized activity. These characteristics help to differentiate and identify it from registries pertaining to other sociocultural contexts such as the one studied in this case. Professional situations typically use a technical vocabulary specific to the area of interest and expressions with a special meaning. The messages transmitted tend to be in writing. Nevertheless, scientific authors in real life are often unable to communicate their message, as set down by Wüster in the General Theory of Terminology (GTT, see Wüster 1979), that is with terminological units exclusive of their discipline, since they also need to use lexical units from the general language and even specialized lexical units from other disciplines.

Important aspects must be considered when deciding on analyzing the lexical units in one of a particular author's works or texts. Since the writer is like an authority to be followed and respected, a productive author (among the most cited in a field) should be chosen. Other aspects should also be considered, such as birthplace, socioeconomic status, individual experiences, culture, ideology, religion, political stance, verbal tradition, language, professional training, previous individual and group research, experience, freedom of expression, individual interests, updatedness, scientific specialization and the type of documents or texts produced, since they could be as varied as letters, communiqués, reports, theses, research papers, articles, books, speeches, decisions, norms, laws, regulations or general documents. In fact, to situate the production of the terms to be analyzed, the type of document or scientific text must be identified, as well, of course, whether it is by an authority on the subject, is an oral or written communication, was produced quickly or put together slowly, was a freely chosen or assigned topic, etc.

Another aspect to consider is the use of specialized expressions, because scientific authors are generally meticulous when selecting lexical units for their texts, so as to minimize the ambiguities in the scientific and technical communication. Nevertheless, an author may or may not be good at choosing the most precise words to fulfill the communication goal, since countless ideas may guide the choice of lexical and terminological units in the discourse, among them the very situation in which the discourse is produced, the language in which it is written and the correct use of nomenclature, proper names, abbreviations, acronyms, initials, pat expressions, codes, passwords, concepts, numbers written out and in figures, symbols, formulas, conventions, etc., which may or may not contribute to a terminological unit taking shape in specialized texts.

Clearly, numerous factors can influence an author's selection of lexical and terminological units. Moreover, simple forms, syntagmas, pat expressions and multi-verbal terminological phrases exist, and in addition, another type of information, often found in academic, technical and specialized texts and in greater proportions is the use of quotes and transcriptions that are studied to find out about what others have said either as thoughts or scientific proofs (Cabré et al. 2014). They often appear in the language in which they were originally produced, such as Latin, Greek, English, French, etc., and along with their respective critical baggage.

3.0 Content Analysis

It is a good time to focus on the idea that to solve scientific problems, both the sciences and technical fields carry out their research by means of the analysis method, with a particular type of analysis used in each discipline or field of human knowledge. Certain analysis methods and techniques are related or complementary to information studies. Of course countless disciplines offer useful knowledge on the subject, but in fact, traditionally information is sought from the closest areas, such as: linguistics, the gamut of applied linguistics and computer science, among many others. Analyses are developed in these disciplines that could be applied to multidisciplinary studies, for example: content analysis, analysis of discourse, grammatical analysis, qualitative analysis, quantitative analysis, analysis of analytical definitions, analysis of contrasting phraseology, lexicological analysis, document analysis, analysis of conceptual relationships, analysis of texts, analysis of syntagmatic units, analysis and design of linguistic corpora, analysis of terms and lastly the document content analysis method used to transfer information.

In the introduction to the book *Text and Context*, Van Dijk (1977) explains how discourse analysis is studied by different scientific disciplines and the extent to which there is an interdisciplinary "transverse connection." Van Dijk starts with the assumption that language use, communication and interaction are produced through texts or discourses. Linguistics studies a part of language use, as do other sciences: sociolinguistics, communications, cog-

nitive psychology, pedagogy, jurisprudence, political science, sociology and, of course, information studies. Textual or discursive relationships are formed among different types of texts, the underlying textual structures, their distinct conditions and functions, the contents and the effects they have on speakers (Van Dijk 1977, 12-13). The various types of text, the relationships among them and with society have various kinds of connections that are analyzed from different viewpoints, depending on the field where they are carried out. The sciences of text attempt to delve into the common properties and characteristics of language use within the spectrum of disciplines that encompass the social and human sciences.

The area of information analysis and systematization that comprises library-and-information-science is confined to describing the types of texts, data and informative contents that lead to their localization in systems. However, the use of processes common to other disciplines is undeniable, among them the terminological and lexicographic analysis used in this paper.

4.0 Documentation in Lexicography

The compilation of dictionaries presented by Varantola (2003) is the method and is essential for representing the content of the documents that make up the corpus where the lexicographic units defined in the dictionary are included. The representation of contents carried out allows for consultation and retrieval through different access points, plus, with the information produced by this type of analysis, new products may almost always be generated to satisfy lexical information needs, such as concordances, statistical data, indices and dictionaries. Document content analysis helps in message decoding and retrieval of information pertinent to users of the *Diccionario del Español de México* (*DEM*) indexing system project (This project was launched in 1973 and from the outset, as Varantola (2003) mentions for other corpora, the *DEM* also structured its lexicograph information retrieval system taking corpus linguistics as its base). This is based on the fact that the author previously created his/her message, which is contained in a support document, usually a specialty text. Therefore, it is up to the information centers to ensure that the contents of such documents, as term candidates may be, are easy for users to consult and retrieve.

Corpus linguistics provides the underpinnings of the *DEM*, to maintain general and specialized corpora that offer great capacity and versatility in managing the information it contains, as with any other information system. A corpus of this type, however, presents entries and points of access defined by corpus linguistics. Although multimodal corpora (voice, image, text, etc.) now exist, the corpora used generally by the sciences and technical fields until fairly recently are meant to analyze the words or lexical units contained in general texts or specialized languages in their different modalities and characteristics and especially in the general communications case being studied.

The indexing process in lexicography basically requires fulfilling certain stages for their application, such as:

- Planning activities, including defining goals, organization and methodologies to be implemented,
- Document selection and acquisition mark the beginning of the process. In the case of recordings with informants, transcriptions are made.
- External document treatment involves physically preparing material and thus creating the respective file, for subsequent analysis.
- Next comes the bibliographic description of the document, highlighting the access points that will enable its identification with relation to other documents. The description of printed texts includes authors, title, printer information and physical description of the material. Additionally, in lexicography, data external to the document and of interest to sociolinguistics, pragmatics and semiotics are included. Generally, such data correspond to the communication unit analyzed, in which the sender stands out, the situation in which the communication was generated and the channel utilized. Also, an extra linguistic context or spoken registry is added, with which the formality or informality of the written documents may be ascertained, as well as whether the text targeted a general or specialized audience. Later situational and thematic identification concerning the use of lexical units will depend on these registries as a whole, which will help system users assign meaning to lexical units from the retrieved information.
- Regarding the text or strictly linguistic context, texts written in a scientific discipline should have the components of the linguistic sign (signifier, meaning and referent), with the smallest text equaling a paragraph, or an item. They are analyzed through previously determined programs and algorithms in order to obtain the information contained in the document. Usually, analysis of these texts produces graphic forms of the words or lexical units, as found in natural language texts, either from common or specialized language texts.

In information studies, when pre-coordinated systems of indexing are used, terms are isolated from their contexts. The work method is textual analysis of the scientific document, with subsequent document content analysis, keeping pre-coordinated terms as the main objective. Indexing terms are extracted from the same text. (A corpus may be created ad hoc, or commercial text analysis programs such as WordSmith, AntConc or Notepad, Atlas.ti,

Sketch Engine, etc. may be used.) Lists of signifiers or lexical units, separated from their meanings and referents, were thus derived. The linguistic sign is thus fragmented, which causes information retrieval complications for users, meaning they need backup in their searches.

Unlike this method, when terms are extracted to form linguistic corpora in lexicography, various lists are produced, which may be of simple or compound words, with their grammatical category, through morphology, according to their internal structure, according to the number of syllables they have, or also as: placements, phraseological units, compound syntagmas, phraseological sentences, significant words, key words, stop words, technical terms, neologisms or term candidates. Generally, the lexical units obtained from corpus linguistics are joined by quantitative data (range and frequency), and the realm of their origin may be recognized through the registry of their use, as long as they mainly pertain to specialized languages.

5.0 Terminological Exclusion Through Subsets of the General Language

When intending to extract term candidates from general or specialized texts, it is extremely helpful to keep in mind prior existing information on the language in general, taken from measurement information studies such as informetry, bibliometry, scienciometry and lexicometry, also Luhn's cut-offs and obtaining TF-IDF weights (Schultz 1968) so as to create filters that exclude the common language and mainly retrieve term candidates.

Furthermore, besides the aforementioned indicators, we suggest that other very similar indicators based on the natural language may be used to exclude subsets of the general language, such as: the basic vocabulary (similar to the highest frequency index and the Zipf model (Zipf 1949), the common language (based on the dispersion index) and the list of grammatical words (the equivalent of empty words), with the goal of isolating the specialized units searched for in the text as much as possible. In other words, the lexicographic knowledge that, in this case, produced by the *DEM* project (2012) and its *CEMC* (1975), may be reutilized, as a way of simplifying the information to be analyzed.

Some results of the *CEMC* content analysis, which was structured with close to two million grammatically labeled words, are used in this paper. That corpus yielded a lexicographic product, the *Diccionario Estadístico del Español de México* (*DEEM* 2005). It is a statistical index of natural language with lexical grammatical and sociolinguistic information, language use registries and quantitative data.

The results from the DEEM regarding empty words, greatest dispersion and highest frequency were:

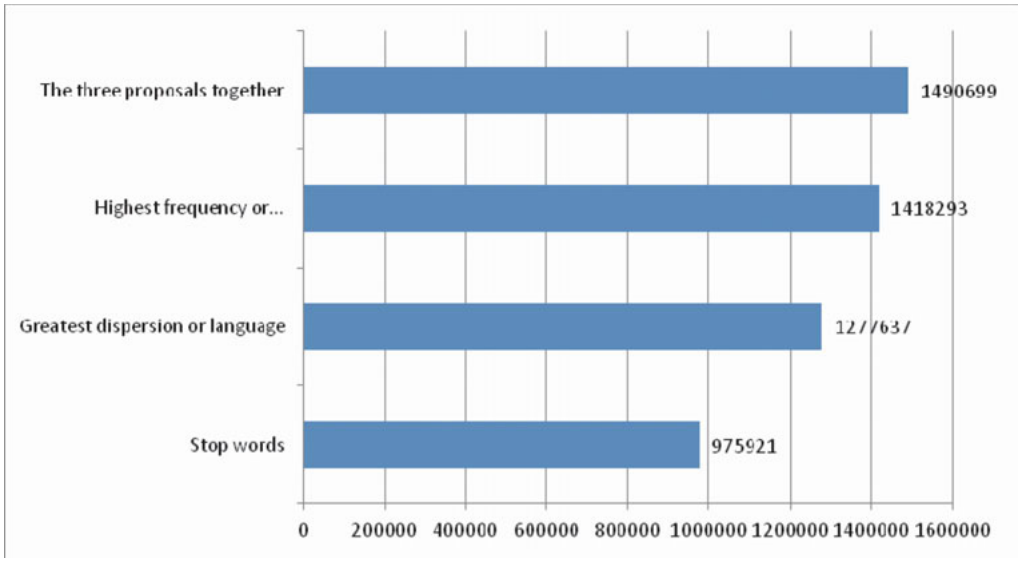
1) Grammatical lexical units or empty words. They are mainly articles, prepositions, interjections, pronouns, etc. and come to 292 headwords, or 51.60% of the total information in the corpus. This is the third group of excluded terms when seeking scientific and technical terms to extract.
2) The lexical units with the greatest dispersion, or common language, (see Anguiano 2013) are 994 different headwords that corresponded to 67.57% of the total information in the corpus. When a specialized term search is carried out, this type of lexical unit tends to separate from the document content analysis.
3) The most frequent lexical units, or basic vocabulary, as Lara introduced (2007). Here it is helpful to realize that studies of lexicometry, informetry, the Zipf model, etc. pointed out that there is an economic phenomenon in language use, known as "least effort." It basically describes how people use a huge amount of graphic words that correspond to a very small amount of headwords, the result being a very small number of lexical units with very high frequency. Following this reasoning, it is understood that the basic vocabulary or that with the highest frequency index is the most used in texts and speeches, as in the *CEMC*, where merely 861 headwords comprise 75% of the total information in the corpus. It is recommended that this type of lexical units also be eliminated.

Chart 1 shows the results for these three divisions, explaining their exclusion from analysis as a drastic saving.

Clearly, the three subgroups together do not total themselves, because some lexical units are repeated in two, or even all three, subgroups. Considering that together they might total 78.83% of all the information analyzed, it is then interesting for information retrieval to create filters with the information from the general language prior to content analysis, which would save about 80% of term candidate retrieval, which coincides with calculations made in other information retrieval studies. To make the work of retrieving scientific and technical terms more efficient, a minimum of valid lexical unit appearances is set, to avoid filtering very low frequency lexical units, because terms whose meanings have no literary guarantee may appear.

6.0 Document Process for Clarifying Meaning and Defining Term Candidate Use

The way we propose to retrieve text of user interest is that once the index of signifiers is obtained, equivalent to the list of terminological unit candidates, they must be simplified and made into headwords. Next, each unit is retrieved through the thematic use registry to which the



	Stop words	Greatest dispersion	Highest frequency	The three proposals together
in relation to the total	51.60	67.57	75	78.83
lexical units	975921	1277637	1418293	1490699

Chart 1. Proposal of cut-offs; empty words, highest frequency, greatest dispersion and of the three together, with respect to 1,891,058 lexical units (%)

Concept	*Graphic words*	*% respecto al total*
Empty or grammatical words	975921	51.60561
Greatest dispersion or common language	1277637	67.57303
Highest frequency or basic vocabulary	1418293	75
The three together	1490699	78.83653

Table 1. Summary of empty words, greatest dispersion, highest frequency and the three together

documents analyzed belong, making the practice into something like indicating the available text language; as this would help users "clarify the meaning and find the appropriate use of certain words" (Estopà 1998, 360). López (2013, 1) wrote: "The available vocabulary is the set of words that speakers have in their mental language and whose use is conditioned by the concrete topic of communication. The idea is to discover what words a speaker would be able to use for specific topics of communication." Furthermore, users may subsequently ask the information retrieval system for the referent closest to what is being sought, simplifying searches as much as possible. Nonetheless, and despite all efforts, the true meaning will always be the reader's interpretation. Like the indexing process, term candidates or key words may be adjusted to a controlled language to improve content retrieval. This can be done by employing subject headings and thesauri. Words from the natural language obtained through indexing are converted into expressions and concepts in a controlled language.

At the end of the indexing process, the information is dispersed so that it gets to users who can utilize it. In the case of lexicographic projects, there are different information products derived from the document analysis that target internal and external users. They may be separate or joined as a system. The components may be the database of concordances, similar to KWIC (Key Word in Context), quantitative information, index cards, the very dictionary being put together, or the various interfaces generated to consult the lexicographic information. Among the results of the long lexicographic process of document content analysis, what should result upon con-

cluding the indexing or natural language classification is a list of lexical unit signifiers from the general language, but also from the sciences and technical fields based on the presence in texts related to this field of work.

7.0 Application of Use Labels from Lexicographic Documentation

Based on the results from the *DEEM*, another database could be put together, the sociolinguistic model of the Spanish language used in Mexico (Anguiano 2006). After assigning the lexical units from the *DEEM* semi-automated indexing, it was possible to get the sum of the partial results that the prior base showed. Then with the complete data, the total results of the lexical units used in the general language in Mexico could be identified through their sociolinguistic registries (see table 2).

Pre-coordinated terms may appear in texts from different specialties, because they are not exclusive to one knowledge area but are, rather, shared terms.

8.0 Proposal for Limiting Term Candidates

For the specialized information search and retrieval, we suggest eliminating the following information from the quantitative data and general language use labels prior to the general and specialized text document content analysis:

– The most frequent lexical units
– The most widely dispersed lexical units
– Lexical units pertaining to the empty word group
– Lexical units from non-standard language
– Lexical units from semi-proper language

Eliminating the quantitative and sociolinguistic units listed above from the analysis would economize substantially on term candidate information retrieval.

But most importantly, after coming up with the list of term candidates, comparisons may be made with the pre-existing language use registries in this same sociolinguistic model of the Mexican Spanish language. Such an examination would help both information users and library and information science professionals in the reconstruction of the meaning of the linguistic sign and the elaboration of a controlled language.

This new correlation exercise could reveal term candidates that are used exclusively in a particular discipline, which would confirm, first off, that they are key words and that later, after expert validation, they might turn into terms in the strict sense. The comparison may also lead to identifying candidates used in two or more disciplines, which would indicate that they are terms in a loose sense and may even have polysemy, in other words that for lexicography they are technical terms. Candidates may also be found that pertain to both the sciences and technical fields, so could be considered technical terms while bearing the (sci.) label in dictionaries, denoting they pertain to the scientific language.

Furthermore, we propose reutilizing lexicographic processes to differentiate lexical units and extract scientific and technical terms through content analysis of specialized texts, by employing use labels or spoken registries as suggested by Rey-Debove (1971, 91-92) who describes three fundamental aspects for achieving this goal:

1) the set of words (lexical units) that belong to a language;
2) the sociolinguistic information from lexical units; and,
3) the use labels agreed upon by the community.

By incorporating these three guidelines into the information analysis, the expectation is that when the same lexical units from the general language, identified by consensus, are contrasted with the specialized language, they may first be used to classify the technical terms, and since the latter behave very similarly to terminological units, may also be designated term candidates. (From a linguistic standpoint, terms may also be called technical terms, as stated in the following definition translated from the Spanish (*DMLE* 2007 s.v.):

> TECHNICAL TERM. *m.* 1 A term that has a concrete and specific meaning within the language of a trade, science, art or industry: the word "algorithm" is a technical term in mathematics.

9.0 The Search for Terminological Units in Texts

To access specialized terms with the help of a corpus from the general language such as the cited model, first, term candidates with a spoken registry related to a specialized text are separated. At this stage of the term search process, it is normal to find, in the lists produced by the automated analysis, lexical units that pertain to use in a discipline at its various communication levels, even if all these units belong to the standard language, the proper language and a science or technical field. This, in itself, means that the following lexical units from a general or scientific text may be obtained from the content analysis: 1) units from the general language; 2) units pertaining to the style of the analyzed discipline; 3) term candidates in a loose sense and 4) term candidates in the strict sense, as shown in image 1.

Headwords	*Gram. cat.*	*Frequency total*	*% total*	*Use of Spanish*	*Language level*	*Spoken registries*	*Greatest frequency*	*Best distribution*	*clave de text*	*Registry 1 use*	*Registry 1 use*	*Registry 1 use*
action	noun	4	0.00021	standard	proper language							
attitude	noun	259	0.01370	standard			basic vocabulary					
activation	noun	14	0.00074	standard	proper language	sciences			420, 427, 428, 454, 469, 473, 477, 478	Chemistry	Medicine and veterinary science	Human medicine
activated	adj	2	0.00011	standard	proper language	sciences			389, 478	Electronics and electricity	Human medicine	
actively	adv	12	0.00063	standard	proper language							
activity	noun	511	0.02701	standard			basic vocabulary					
activist	adj; s	6	0.00031	standard	l proper language							
act	noun	308	0.01629	standard			basic vocabulary					
actor	adj; s	133	0.00704	standard								

Table 2. Example of Use Registries in the Identification of Term Candidates in the Sociolinguistic Model

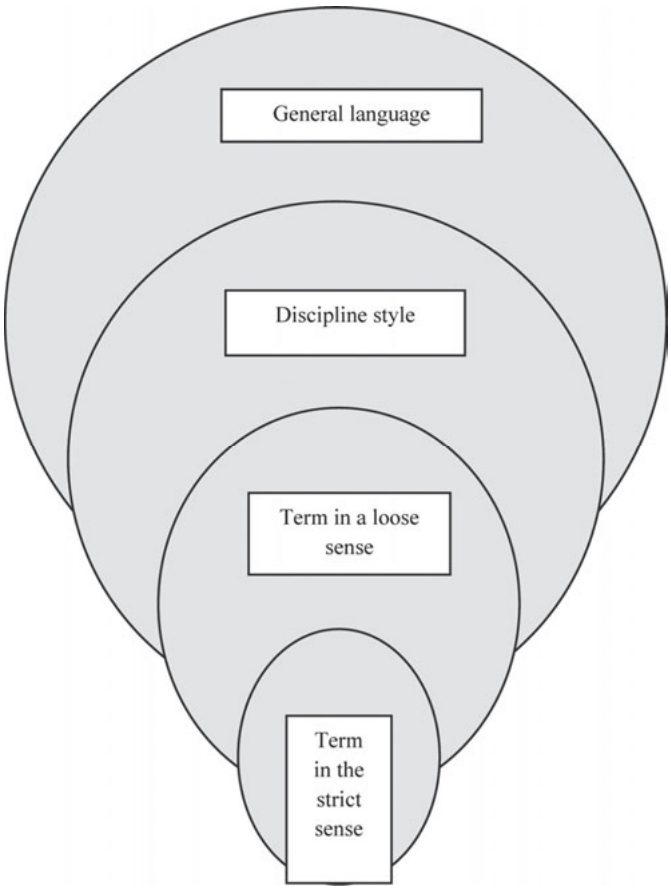


Image 1. Lexical Units in the Document Content Analysis of a Text

To delve further into the discussion from the previous paragraph, a more detailed explanation regarding the units to be found in texts follows:

1) Lexical units that belong to the general language and that appear in texts from the sciences and technical fields but that are also identified sociolinguistically as pertaining to: the general standard language, non-standard language, semi-proper language and proper language. In other words, they are not exclusive to the sciences or technical fields, and it is advisable to exclude them from the term candidate list.
2) Lexical units that correspond to the writing style typical of the discipline under analysis. These units tend to be lexical units from the discipline's verbal tradition, pat phrases and locutions. Their appearance corresponds to a very low frequency index with respect to an analyzed text, yet since they are characteristic of certain scientific disciplines it is not advisable to eliminate them prior to content analysis. Here we may find locutions, phraseological units, Latinisms, etc.
3) Specialized lexical units or technical terms. Their use and meaning are typical of the discipline to which the analyzed text corresponds, though they may also have the same signifier in the general language and even in other disciplines; in other words, they may have synonyms. This type of lexical units are recorded in general language dictionaries, (as in DRAE 2001 or DEM 2012) and in fact such units are terms in a loose sense. Forms of graphic words, just as they appear in the original text, tend to be few: feminine, masculine, singular and plural; they have a very low document content analysis frequency index in the common language, but as lexical units made into headwords (canonically grouped words) their percentage with respect to the total of the analyzed sample increases. In other words, a small number of lexical units is grouped within a large number of headwords. In terms of their dispersion index in the DEEM, what was seen is that while they may be concentrated by their use in a discipline, they may also show up in other disciplines within the sciences or technical fields or belong to scientific language, which encompasses both knowledge areas. Among other things, they may be recognized, because even if they have a known signifier, the meaning differs from that in the natural language, which is why the average reader does not

grasp their meaning and it seems like a secret. These units may appear simply or multi-verbally as, for example, in syntagmas, pat phrases or as phraseological units.

4) Candidates to be terminological units of the discipline analyzed. Their document behavior is very similar to technical terms, but they do not have synonyms and presuppose a univocal meaning. These units belong to standard and proper language, are used exclusively in the sciences or technical fields and have a spoken registry that situates them in a formal communication mode so they are used exclusively in a specialized language, having no meaning or equivalent in common language. Such candidates may be simple lexical units or units composed of several words. They pertain exclusively to one discipline. In principle, candidates may be considered key words; after being validated by an information specialist they may become part of the indexing language, and in the best-case scenario they may be terms of a particular discipline, in the strict sense (Rey-Debove 1971). With a low appearance frequency level in text analysis, when lexical units are grouped together, they have a high percentage of headwords in relation to the total analysis. Since their data are concentrated in a single discipline, they do not undergo dispersion.

Considering all of this, it is also likely that in any document content analysis of a text, be it general, on science or on technology, and keeping in mind Rey-Debove's perspective (1971), the lexical units have the following characteristics regarding signifier, meaning and type of communication they belong to (see table 3):

*Signifier**	*Meaning***	*Language Type*
A common signifier	and a common meaning	form part of the general language.
An uncommon signifier	and a common meaning	would be a technical term of signifier, for ex., close up, stock shot, fade out.
A common signifier	with an uncommon meaning	is a technical term in a loose sense, for ex., winder, optic, camera.
An uncommon signifier	and an uncommon meaning	would be a technical term in the strict sense, for ex., magnetic eraser, projection lamp, translucent screen system, animation technique.

Table 3.

* Signifier is that which indicates something, in this study a word or lexical unit given to a person, animal, thing or concept, tangible or intangible, concrete or abstract, to distinguish it from others.

** Meaning is that which is indicated, and for our purposes, the representation or mental concept of something.

Despite this coexistence of lexical and terminological units in a scientific text, they may be differentiated if their spoken registry is verified, confirming whether it is found in a form of communication or in a text that belongs exclusively to a specialized language, that is whether it is proven to be the product of a formal communication between specialists in a particular discipline to ensure effective communication among themselves.

As seen in the previous paragraph description of the process carried out, the lexical units analyzed originate from an empirical study developed by lexicography, which shows that something similar to what happens with any general language text occurs with texts from a specialized language. Both types of texts are composed, to a greater or lesser degree, of lexical units from the general language and not only specialized units from the sciences or technical fields. And while it may not seem so, these differences are actually useful for information retrieval, as the terms to be extracted from the texts are not typical of the common language.

9.1 Example of the type of analysis made with the model

Following the steps proposed in this article and isolating the headwords that correspond to the sciences and technical fields contained in the model, the following results were produced (see chart 2).

This graph illustrates how 346,284 graphic words were automatically extracted from scientific texts that were manually indexed as 16,296 headwords in science texts. Of them, however, only 4,876 headwords were identified as exclusive to that area. As shown, the sciences obtained 16% of the total of 30,899 headwords taken from the corpus. With technical texts, 202,667 graphic words were automatically extracted and manually indexed as 10,821 headwords in technical texts, 1,574 of which were defined as exclusive to technical fields. This was 5% of the 30,899 total headwords taken from the corpus. With just 6,450 headwords, sciences and technical fields together covered 21% of the total of 30,899 headwords.

10.0 Final Considerations

The model presented here or other lexicographic resources with similar characteristics may be useful in the near future, in computer-assisted indexing or as corpora monitors, with respect to new text analyses or specialized corpora. Their utilization would facilitate rapidly generating lists of term candidate signifiers, which, besides being useful for representing and retrieving original text content, will also be of great value in the stage of development of the controlled language when working with the terms, uniterms, subject headings or descriptors that comprise the terminology of a

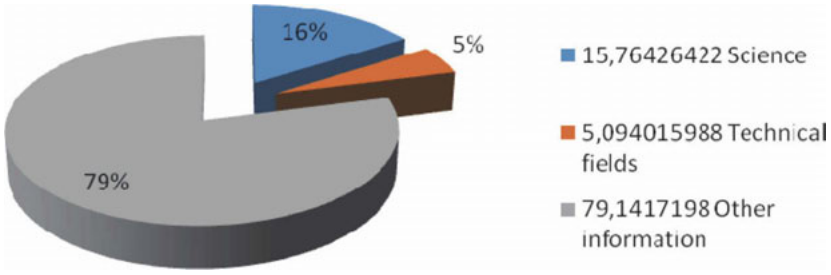


Chart 2. Term Candidates Retrieved from a Total of 30,899 Headwords: 4,871 Lexical Units from the Sciences and 1,574 from Technical Fields

discipline analyzed this way. Finally, we must keep in mind that natural and specialized languages are constantly evolving, so there are, of course, difficulties in controlling and retrieving specialized languages and their terminologies. Consequently, the presence of library and information science and its development are that much more necessary, to help users and readers decode the language of science.

References

Alexiev, Boyan. 2006. "Terminology Structuring for Learner's Glossaries." *Knowledge Organization* 33, no. 2: 96-118.

Anguiano Peña, Gilberto. 2006. *Modelo sociolingüístico del léxico del español usado en México.* Mexico: El Colegio de México, Centro de Estudios Lingüísticos y Literarios, Diccionario del Español de México.

Anguiano Peña, Gilberto. 2013. *El léxico común del español de México.* Mexico: El Colegio de México, Centro de Estudios Lingüísticos y Literarios, Diccionario del Español de México.

Cabré, M. Teresa, Iria da Cunha, Eric San Juan, Juan-Manuel Torres-Moreno and Jorge Vivaldi. 2014. "Automatic specialized vs. non-specialized text differentiation: the usability of grammatical features in a Latin multilingual context." In *Languages for Specific Purposes in the Digital Era*, edited by E. Bárcena, T. Read and J. Arús, 223-41. Berlin: Springer.

Cabré, Maria Teresa. 1999. *Terminology, Theory, Methods, and Applications*. Barcelona: Benjamins.

Cabré, Maria Teresa. 2003. "Theories of terminology, their description, prescription and explanation." *Terminology* 9, no. 2: 163–99.

CEMC. 1975. *Corpus del Español Mexicano Contemporáneo, 1921-1974*. María Isabel García Hidalgo, Luis Fernando Lara, Roberto Ham Chande et al., Mexico: Diccionario del Español de México.

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

DEEM. 2005. *Diccionario estadístico del español de México. Lematizado.* Gilberto Anguiano Peña, Francisco Segovia and Erika Flores (eds.). Mexico: El Colegio de México, Centro de Estudios Lingüísticos y Literarios, Diccionario del Español de México.

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

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

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

Estopà, Rosa. 1998. "El Léxico Especializado en los Diccionarios de Lengua General: Las Marcas Temáticas." *Revista Española de Lingüística* 28, no. 2: 359-87.

Gemma, Will. 2013. *The Elements of Communications a Theoretical Approach*. https://blog.udemy.com/elements-of-communication/#wrap.

Halliday, Michael A. K. 1977. "Text as Semantic Choice in Social Contexts". In *Linguistic Studies of Text and Discourse*. Volume 2 in the Collected Works of M.A.K. Halliday, edited by J, J. Webster, 23-81. London: Continuum.

Lara, Luis Fernando. 2001. *Ensayos de Teoría Semántica: Lengua Natural y Lenguajes Científicos*. Mexico: El Colegio de México.

Lara, Luis Fernando. 2007. *Resultados Numéricos del Vocabulario Fundamental del Español de México.* Mexico: 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,* edited by Luis Fernando Lara, Roberto Ham Chande and

María Isabel García Hidalgo, 7-39. Mexico: El Colegio de México. http://www.scielo.cl/scielo.php?pid=S0718-09342008000200002&script=sci_arttext.

López Morales, Humberto. 2013. ¿Qué es la disponibilidad léxica?. En DispoLex: investigación léxica. Recuperado de: http://www.dispolex.com/info/la-disponibilidad-lexica.

Rey-Debove, Josette. 1971. *Etude Linguistique et Semiotique des Dictionnaires Français Contemporains*. Paris: Mouton the Hague.

Ryan, Janet N. 1985 "The Secret Language of Science or, Radicals in the Classroom." *The American Biology Teacher* 47, no. 2: 91.

Schultz, Claire K. 1968. *H.P. Luhn: Pionner of Information Science; Selected Works*. New York: Spartan Books.

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

Van Dijk, Teun A. 1977. *Text and Context: Exploration in the Semantics and Pragmatism of Discourse*. Linguistics Library, no.21. London: Longman.

Varantola, Krista. 2003. "Linguistic Corpora (Database) and the Compilation of Dictionaries." In *A practical guide of Lexicography,* edited by Piet van Sterkenburg, 228-39. Amsterdam: John Benjamin.

Wüster, Eugen. 1979. *Introduction to the General Theory of Terminology and Terminological Lexicography*. Wien: Springer.

Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort*. Oxford, England: Addison-Wesley Press.