Specialized Text Mining Through Specific Keywords

Maira Alejandra Pulgarín, Maria Cecilia Plested Alvarez, Adriana Lucia Diaz

SPECIALIZED TEXT MINING THROUGH SPECIFIC KEYWORDS

Maira Alejandra Pulgarín[1,2,3], Maria Cecilia Plested[1,2], and Adriana Lucia Díaz[1,2]

[1]Research Group for Terminology and
Translation- GITT, University of Antioquia,
[2]Colombian Terminology Network,
[3]Calasanz' School, Medellin, Colombia
leomaja5323@hotmail.com
[1]Research Group for Terminology and
Translation- GITT, University of Antioquia,
[2]Colombian Terminology Network, Medellin, Colombia
plested@quimbaya.udea.edu.co
[1]Research Group for Terminology and
Translation- GITT, University of Antioquia,
[2]Colombian Terminology Network, Medellin, Colombia
adiaz.trad@gmail.com

ABSTRACT

The aim of this paper is to present the structure of a methodological procedure for the specialized text mining through specific keywords applicable to different LSP- text typologies in specialized areas, where the use of a foreign language is unusual. These specific keywords allow those experts in these disciplines - without a deep knowledge of foreign language - to determine a set of informational contents of the mined texts which provide the experts with a better contextualized conceptual understanding of LSP- documentation of their disciplines.

KEYWORDS: Text mining, terminological work, terminological analysis, LSP-text, cognitive processes, reading literacy, specialized knowledge, WIKO model, social circus, forensics, thesauri, methodological procedure, block contents.

INTRODUCTION

According to the research carried out on cognitive autonomy and reading literacy [1], readers of specialized fields (the most, postgraduate students) claimed that their level of appropriation of a foreign language was weak, because they don't have the ability to determine keywords derived of a concept system that allow them to build textual relationships in order to understand a text in a foreign language.

Therefore, they had to resort to the use of dictionaries to try to understand the corresponding conceptualization. In addition, these readers were unaware of use

appropriate search engines to select the right document of the conceptual definition or the specific keywords necessaries to make an effective text mining in order to classify the specialized texts they needed. In that case, the benefits of a digital interactive text mining for reading are reduced.

In the same way, as one result of the project Terminological Thesaurus in Translation and Interpretation Terminology 'TETIT' [2], it was ratified that at the students researchers level was necessary to have a concept system of keywords, in order to complete a text mining based on the facets that shape the thesaurus as a basic tool for the specialized search procedure. In particular, for those who need the help of keywords as a starting point for your specialized documentary research, it is a great help.

Similarly, for the second phase of the Forensics Project [3], where five specialized subareas were determinate from a wide corpus of documents in Spanish, the difficulties to correlate the definitions in the contexts in which the terms endorsing the specific concept in the two foreign languages (English and French) in which the research was also carried out, was further revealed. Both languages were worked in parallel to the Spanish concept system, for this reason the team of co-researchers was settled by translators and terminologists for whom the procedure to demarcate the corresponding definitions and apply keywords to an LSP-text mining field in those foreign languages, it means in a complex corpus of documentation in those foreign languages, was an every-day-work.

In the case of Social Circus [4], the determination and finding of equivalents in five languages (French, English, Spanish, Portuguese and Russian) had the same difficulty: to identify the specialized keywords for text mining in a parallel work about the concept definitions in these five languages at the same time. Therefore, the researchers were called to set a specialized scientific terminological work for solving this problem.

In all these processes, one of the research question was how to establish specific keywords for LSP-texts mining in particular subject fields for special expert groups, not only with a basic level of foreign language knowledge, but also with a wide competence in their LSP disciplinary research field.

In this order of ideas, it determines the methodological strategies which lead the search for the readers (researchers) with limited knowledge of a foreign language for the construction of a concept system. The keywords were précised according to their disciplinary necessity for the text mining process.

THEORETICAL SUPPORT- METHODOLOGY

All the applied procedures in these case studies were developed based on WIKO-Model [5], COLTERM methodologies [6] and cognitive processes in order to achieve definitions and their equivalents and to determinate the precise concept systems for LSP-text mining in particular subject fields. The concepts that guide the methodology are the result of another of our research. This allows them to dial with a specific conceptual parameter in the terminological analysis in context.

Our research's theoretical starting point of the underpinning terminology theory must be understood as an organizational focus of expertise constituted by conceptual units that make up the subject field or discipline in every spoken or written process. The terminology in the context of a foreign language must be seen as an additional tool in the learning of that language, which improves the acquisition basis and gives the possibility to elaborate and acquire new knowledge [7]. In addition, one of the most important aspects is to consider each concept definition through its specific diachronical tracking [8]. Budin defines terminology as an epistemic, informational, communicative system from a specific discipline that confirms the principles of determinate concept organization based on pragmatic criteria (functions, goals and purposes), and in which the terminology unit, as a relationship of correspondence between concepts and representation, is the primary unit of reference [5].

Other important point in this study is the terminological analysis for text mining through specific keywords. For the Research Group for Terminology and Translation -GITT terminological analysis is "the procedure used to argue about informational relations on a field of specific knowledge. It consists of decomposed the concept, as object of knowledge, into a contextualized statement configured by the specific characteristics using the object definition to delimited them from other concepts or terms". In fact, it contains conceptual resources that allow establishing relationships between concepts to build conceptual schemes or corresponding ontologies which, in turn, are guided by the default keywords in their thinking. The above permits the directed text mining for obtaining secure results according to the needs of the expert.

In this point, it is important to take into account the onomasiological approach of the terminology production in each knowledge field, which also will allow the application of strategies for integrated multilingual content development and terminological knowledge management in academic digital environments, as it was ratified by Pulgarín, Plested [1]. Therefore, the terminological analysis permits the selection of the most relevant content or context at the time of the construction of knowledge by the expert. Concerning terminology as a specialized tool, we consider that achieving a specialized text mining through specific keywords, it is an important part that integrates prior knowledge with a new one. In fact, it contains conceptual resources

that allow establishing relationships between concepts to build conceptual schemes or corresponding ontologies which, in turn, are guided by the keywords that the experts could determine. The above permits the directed text mining for obtaining secure results according to the needs of the expert. Because of an excessive information access and the implementation of digital literacy, the experts don't know how to reach the information they need, they feel lost. [1], [3]. Hypermedia and interactivity makes that readers in specialized fields distort their reading, due to the specialized contexts. In turn, this specialized context is essential to read in a foreign language for specific purpose and to select a relevant specific knowledge corpus in their work field [5]. When reading an analogue text in a foreign language, certain specific contexts are options for the access to a better understanding.

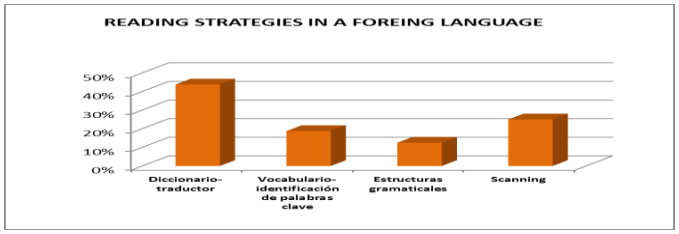

Fig. 1. Reading Strategies in a Foreign Language (Pulgarin, Plested 2012, p. 13)

According to the above, the specialized readers require establishing strategies that enable qualify these skills to read analog formats. However, a reading proficiency in digital media such as videos would be remedied by an efficient key words concept mining [4]. González [7] supports that "an autonomous model is based on a methodology of independent learning in which the learner applies their academic experiences along with the theoretical and practical foundation that has received from teachers". Consequently, the WIKO modeling [2] and the mining theory in specific subject fields allowed the development of methodological procedures and search strategies derived from the LSP – Methodological Matrix [9], applicable to each LSP-discipline.

This is also based on the analysis of specialized terminological contexts [10], [6] as well as from some methodological applications supported on cognitive autonomy and reading competence of specialized texts. Therefore, Pulgarín and Plested [1] consider that the issue of understanding is made more complex because of the reader lacks of appropriate methods to read different texts and to limit the access to specialized documents in a foreign language. In addition, the reader does not have strategies to find that kind of specialized information or the necessary autonomy to produce new learning about keywords for text mining starting from what she/he is reading. Since the saturation of information delimits the efficiency of what is relevant in terms in

contexts and for specialized expert search. In this case, Cassany [11] proposes that there are new communicative practices which emerge in everyday interaction and constitute the permanent flow of information; it means, modern structures, namely, Hypertext and Intertextuality [5], [11] are the new ways of records and particular linguistic forms, therefore, also increase the cognitive processes involved in the interaction, reading or writing for which dare to be significant changes in the culture and ways of thinking of the societies [12] (expansion of democracy, increased capacity of communication and liberties, etc.).

Extrapolating Coiro [13] in different studies of context school has ratified changes in the cognitive processes that have produced new forms of communication challenges in the ways of interpreting. In his studies of American school contexts, he has managed to determine that there are notable differences in the processes that people develop to understand a text in an analog context with respect to the digital context.

ANALYSIS AND RESULTS

The analysis procedure and methodology for projects studied have followed specific steps:

1. Selection of bibliographical sources, as corpus, for tracking
2. Tracking and terminology compilation
3. Search for conceptual equivalences in Spanish, English and French
4. Selection and terminological comparison in at least these three languages
5. Precision of the findings and relationships between the terms in context
6. Terminological register in the database Multiterm of Trados using the COLTERM format
7. Verification of the use of key words in the specific text mining.

Besides, the following COLTERM methodology [6] can be used as the axe point of the text mining:

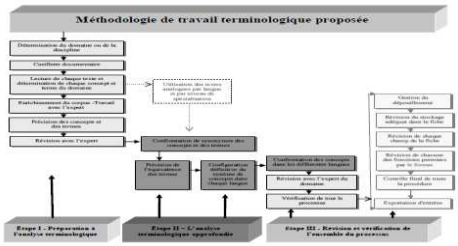

Figure 2. Methodology for Terminology Work (Source: Díaz, 2007, p. 114)

The above methodological procedure for terminology work can be used by researches, experts, teachers and students of a specific field in order to determine the necessary keywords for specialized knowledge extraction and conceptual precision in contexts through text mining and applying autonomous reading strategies in foreign language [1], [13].
As final step after the end of the research project it's useful to develop a 'Validation Workshop' with recognized experts of the specific field [14].

These steps were applied in all of the following studied projects:

Análisis diacrónico de los conceptos: "definición, concepto, análisis terminológico y rastreo terminológico". Hacia una precisión conceptual en Colombia: The result of this project allowed the conceptual path for the entire analysis process in all other projects; i.e., the definitions of those concepts are the guide for the entire study of the documentation that was collected as corpus for each project. Similarly, it served to analyses the contexts and definitions of other concepts needed in every field of studied knowledge in order to determinate the key words that should be used by experts for information retrieval [8].

'TETIT': The first part of this project reflects the documental and terminological dimension with alphabetical and hierarchical order, and the second part shows the terminological and documentary results with thematic development of the contextualized terms in the COLTERM format. For the terms treatment a process in stages has been used:
a) Analysis of the relevance of each term in relation to the thesaurus
b) Selection of synonyms and quasi-synonyms in context
c) Location of the term in the corresponding sub-facet. This point is of great importance for the analysis, because a term may belong to two or more facets, therefore you must take the right decision to locate it in the facet of greater degree of relationship, although within our Thesaurus, it was studied the possibility of placing a term in several facets if necessary [2].

From this project was derived a thesaurus about translation and interpretation. This thesaurus contains the initial proposal and an additional database in COLTERM format.**Conceptualización Metodológica en Clínica Forense: análisis de inconsistencias y ambigüedades (Forensics project):** The corpus of the studied areas of knowledge allowed determining disciplinary actions and to revise cognitive understanding difficulties in highly specialized thematic contexts. According to the terminology and the resultant concept systems as a basis for the determination of the specialized ontology to specify the keywords, allowing the group of experts to perform an adequate text mining [3].

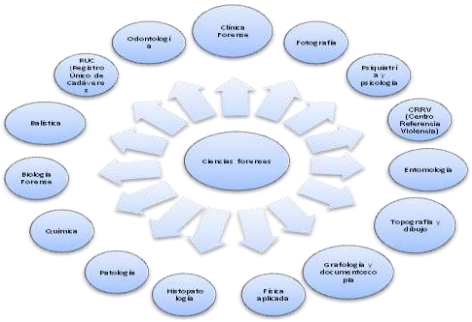

Fig. 2. Ontologies on Forensics (Source: Tobón et all., 2012, p. 5)

Determining an ontology according to the particular needs of the readers- researchers to promote processes of understanding in specific fields of knowledge, requires a high level of cognitive autonomy, because readers- researchers must be able to establish relevant metadata encouraging the understanding in context of the texts that dealt with. From this project we have obtained a terminological database in three languages: Spanish, English and French with definitions and contexts, all of these were extracted from originals documents from each language.

Social Circus of the Cirque du Soleil, Montréal Canada: From this project, the translation in five languages of the Basic Techniques in Circus Arts (TBAC) and of the Social Worker´s Guide (in Social circus) has permitted to create the first lexicons and glossaries in Circus Arts and Social Circus built on a harmonized conceptual basis [4].

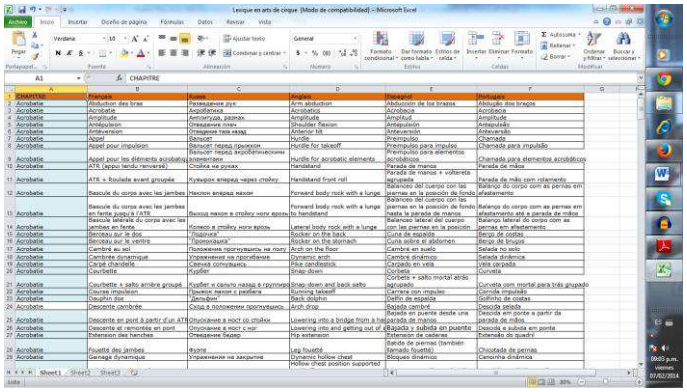

Figure 4. Screen from TBAC's glossary. (Source: Díaz, 2013)

These TBAC and Social Worker´s Guide's lexicon and glossary, with more than 300 terms in five languages (French, English, Spanish, Portuguese and Russian), is serving as a multidisciplinary tool that is used not only in the circus and social area

but also in the area of sports. Besides, it offers definitions to the terminology of the social circus, as it is used by Cirque du Soleil. These definitions ensure a common understanding of the terms and expressions resulting from the set of documents on social circus of Cirque du Soleil. Both, lexicons and glossaries, are spread through the platform Box, which is used by more of 400 organizations involved in social circus, and its aim is to share information and specialized knowledge in this area [14].

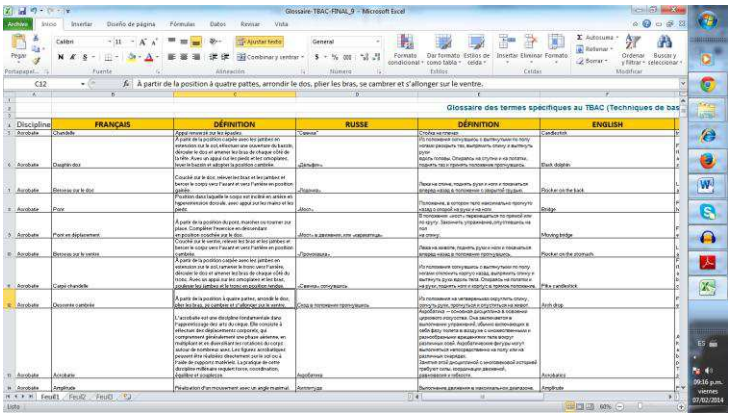

Figure 5. Screen from TBAC's glossary with definitions. (Source: Díaz, 2013)

CONCLUSIONS

The application of COLTERM methodology for terminological work combine with reading strategies allow a better use of text mining in order to obtain specialized keywords by experts, students or researchers, which knowledge in foreign language was weak. This improved their own research work.

These terminology analyses have permitted effective transfer of knowledge in order to have a harmonized and effective communication among peers.

The production of specialized glossaries and lexicons favors the creation of an interdisciplinary awareness and scientific community. All the processes in the projects were corroborated through workshops of autonomous learning.

The specialized text mining through specific keywords allow experts in different subject fields to determine a set of informational contents which provide them with a better contextualized conceptual understanding of LSP-documentation of their disciplines.

ACKNOWLEDGMENTS

This paper is derived from some research projects developed by the GITT – Grupo de Investigación en Terminología y Traducción – School of Languages, University of Antioquia and the Colombian Terminology Network - Colterm in Medellín, Colombia: Análisis diacrónico de los conceptos: "definición, concepto, análisis terminológico y rastreo terminológico". Hacia una precisión conceptual en Colombia; Tesauro Terminológico en Terminología, Traducción e Interpretación "TETIT". Both supported by GITT-CICINF, UdeA. Conceptualización Metodológica en Clínica Forense: análisis de inconsistencias y ambigüedades. Informe final de Investigación, supported by GITT-CEDED, Universidad de Antioquia, Medellín, Colombia; Mr. David Simard, Social Circus of the Cirque du Soleil, Montreal, Canada who gave the authorization of use reserved material for this article, Ms. Adriana Lucía Díaz, that supported all the research with the Social Circus Department at the Cirque du Soleil, Canada and who contributed to the development of the methodology of terminology work for Colterm,

REFERENCES AND CITATIONS

[1] Pulgarín Rodríguez, M. A., Plested, M. C. (2012). Autonomía Cognitiva y Competencia Lectora en Lengua Extranjera. In: Universidad de Medellín CD ROM 6948, Medellín

[2] Restrepo, R. et al. (2007). Tesauro Terminológico en Terminología, Traducción e Interpretación "TETIT". Informe final. GITT-CICINF, Universidad de Antioquia, Medellín, Colombia.

[3] Tobón, F.A., Plested, M.C., Betancur, A., López, L. A., Mejía, M.L., Gutiérrez, G.P. (2012) Conceptualización Metodológica en Clínica Forense: análisis de inconsistencias y ambigüedades. Informe final de Investigación. GITT-CEDED, Universidad de Antioquia, Medellín, Colombia.

[4] Díaz, A. L. (2013) Aplicación de la Terminología para la Transferencia de Conocimiento: Caso específico sobre Circo Social. Paper presented in the International Conference on Terminology and Multilingual Structured Content for Industry and Trade –Best Practices in Research and Applications– VI Seminario Nacional de Terminología. EAFIT, Colombia INCOMPLETO.

[5] Budin, G. (1996) Wissensorganisation und Terminologie : Die Komplexität und Dynamik wissenschaftlicher Informations- und Kommunikationsprozesse. Tübingen**:** Narr Verlag.

[6] Díaz, A. L. (2007) La méthodologie de travail terminologique au Québec et en Colombie: étude comparative. Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de Maître des Arts (M.A.) en Traduction option Recherche-Terminologie. Département de linguistique et de traduction, Faculté des arts et des sciences Université de Montréal, Canada.

[7] González, A.(2007) La enseñanza aprendizaje del inglés con fines profesionales, http://hera.ugr.es/tesisugr/16649801.pdf

[8] Plested, M. C., Giraldo, B. S., Villamizar, C., Piraquive, E., Sandoval, A., Castrillón, E. R., Aristizábal, Y. L., Pérez, J.M. (2003) Análisis diacrónico de los conceptos: "definición, concepto, análisis terminológico y rastreo terminológico". Hacia una precisión conceptual en Colombia. Informe de Investigación CODI, Universidad de Antioquia, Medellín.

[9] Giraldo, B. S. (2005) Terminology and Knowledge Management Principles to Orient ESP Courses. Tesis de Maestría, presentada en la Facultad de Artes y Humanidades de la Universidad de Caldas, Manizales, Colombia

[10] Plested, M. C., Casals, S., Vallejo, G. C. (2010) Un enfoque onomasiológico de la Ciencia. V Coloquio Internacional sobre la Historia de los lenguajes Iberorrománicos de especialidad: Comunicación y transmisión del saber entre lenguas y culturas, 27 -30 Mayo 2010, Universidad de Leipzig

[11] Cassany, D. (2002).De lo analógico a lo digital, el futuro de la enseñanza de la composición- Lectura y Vida. Revista Latinoamérica de lectura. Año 25, N°3 de septiembre 2004.

[12] Galinski, C. (2000). Terminology Infrastructures in Europe. Weltgesellschaft, Weltverkehrssprache, Weltkultur. Globalisierung versus Fragmentierung. Tübingen: Stauffenburg Verlag Brigitte Narr GmbH.

[13] Coiro, J. (2011). Predicting reading comprehension on the Internet: Contributions of offline reading skills, online reading skills, and prior knowledge. Journal of Literacy Research, 43(4)352-392

[14] Plested-Alvarez, M. C., Begué, J., Castrillón, E. R., Flórez, S., Ospina, I. C. (2006) Concept Units Organization: A Must. TSTT' 2006. International Conference on Terminology, Standardization and Technology Transfer. Proceedings, Encyclopedia Of China Publishing House, Beijing