# Exploring MWEs for Knowledge Acquisition from Corporate Technical Documents

**Bell Manrique Losada**
Universidad de Medellín
Cra. 87 30-65 Belén
Medellín, AQ, Colombia
bmanrique@udem.edu.co

**Carlos M. Zapata Jaramillo**
Universidad Nacional de Colombia
Cra. 80 65-223 Robledo
Medellín, AQ, Colombia
cmzapata@unal.edu.co

**Diego A. Burgos**
Wake Forest University
Greene Hall, P.O. Box 7566
Winston Salem, NC 27109, USA
burgosda@wfu.edu

## Abstract

High frequency can convert a word sequence into a multiword expression (MWE), *i.e.*, a collocation. In this paper, we use collocations as well as syntactically-flexible, lexicalized phrases to analyze 'job specification documents' (a kind of corporate technical document) for subsequent acquisition of automated knowledge elicitation. We propose the definition of structural and functional patterns of specific corporate documents by analyzing the contexts and sections in which the expression occurs. Such patterns and its automated processing are the basis for identifying organizational domain knowledge and business information which is used later for the first instances of requirement elicitation processes in software engineering.

## 1 Introduction

In software engineering, business knowledge and the needs of a system's users are analyzed and specified by a process called requirement elicitation (RE). Traditionally, RE has been carried out by human analysts through techniques such as interviews, observations, questionnaires, etc. The information obtained by the analyst is then converted to a controlled language used further stages of software implementation. These techniques, however, necessarily increase costs and imply a certain degree of subjectivity. Sometimes, as an alternative approach for RE, human analysts elicit requirement from documents instead of from clients or users. The present work, proposes the use multiword expressions (MWEs) such as collocations and syntactically-flexible, lexicalized phrases to detect relevant patterns in 'job specification documents' (a kind of corporate technical document). The approach contributes to the task of generating controlled language used in subsequent automated knowledge representation.

MWEs are lexical items which can be decomposed into multiple lexemes with lexical, syntactic, semantic, pragmatic, and/or statistical idiomaticity (Baldwin *et al.*, 2010). According to Bauer (1983), MWEs can be broadly classified into lexicalized phrases and institutionalized phrases. Institutionalized phrases, or collocations, basically require a high frequency of co-occurrence of their components. Lexicalized phrases (LP), on the other hand, may present other kind of idiomaticity, but not only statistical. Along with collocations, out of the set of lexicalized phrase types, we find syntactically-flexible, lexicalized phrases and semi-fixed phrases of special interest for the present work.

Based on an experimental corpus, we identify when and how a MWE is used in order to identify patterns, infer organizational relationships, and generate corporate information and/or conceptual models for further requirement elicitation.

We propose context analysis—in which MWEs occur—would contribute by adding essential information to the pattern definition. Such patterns are conceived from the structural and functional components inherent to corporate documents. This means that we classify MWEs according to the section in the document where they prevail. We expect the automated processing of such patterns helps in the identification and understanding of domain knowledge and business information from an organization.

The remainder of this paper is organized as follows: in Section 2 we describe the conceptual



*Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013)*, pages 82–86,
Atlanta, Georgia, 13-14 June 2013. 

framework and background. Section 3 presents examples and analysis of the MWEs used for this study. Last, Section 4 draws conclusions and outlines future work.

## 2 Conceptual Framework and Background

Two main lines converge on this study, namely requirements elicitation belonging to software engineering and linguistic description and parsing related to natural language processing.

Requirements elicitation (RE) is the initial process from requirement engineering in the software development process lifecycle. RE involves seeking, uncovering, capturing, and elaborating requirements, based on activities of the business analysis initially performed. This process comprises functional, behavioral, and quality properties of the software to be developed (Castro-Herrera *et al.*, 2008). In order to accomplish RE, an analyst should increasingly and iteratively develop several actions involving natural language analysis and modeling (Li *et al.*, 2003).

On the other hand, a user of a language has available a large number of pre-constructed phrases conforming single choices, even though they might appear to be analyzable into segments (Sinclair, 1991). Such phrases are known as lexical phrases (LPs) and may have a pragmatic function. According to Pérez (1999), the importance of LPs lies in their usage and domain, which constitute an integral part of the communicative competence. In the same line of thought, López-Mezquita (2007) categorizes LPs into polywords, institutionalized expressions, phrasal constraints, and sentence builders.

For this study, we use the classification of MWEs proposed by Baldwin *et al.* (2010). This and other classifications have been used in natural language processing techniques for text-mining and information extraction. They also have been applied to the analysis of many kinds of documents, *e.g.*, technical documents, patents, and software requirement documents.

Cascini *et al.* (2004) present a functional analysis of patents and their implementation in the PAT-Analyzer tool. They use techniques based on the extraction of the interactions between the entities described in the document and expressed as subject-action-object triples, by using a suitable syntactic parser.

Rösner *et al.* (1997) use techniques to automatically generate multilingual documents from knowledge bases. The resulting documents can be represented in an interchangeable, reusable way. The authors describe several techniques for knowledge acquisition from documents by using particular knowledge structures from particular contexts. Breaux *et al.* (2006) describe the extraction of rights and obligations from regulation texts restated into restricted natural language statements. In this approach, the authors identify normative phrases that define what stakeholders are permitted or required to do, and then extract rights and obligations by using normative phrases.

For knowledge acquisition, several authors have applied NLP techniques for handling MWEs. Jackendoff (1997) and Aussenac-Gilles *et al.* (2000) extract knowledge from existing documents and demonstrate its usage on the ontological engineering research domain.

Some other contributions are related to the extraction of multiword expressions from corpora, empirical work on lexical semantics in comparative fields, word sense disambiguation, and ontology learning (Bannard, 2005). In the intersection of NLP and requirement elicitation, Lee and Bryant (2002) use contextual techniques to overcome the ambiguity and express domain knowledge in the DARPA agent markup language (DAML). The resulting expression from the linguistic processing is a formal representation of the informal natural language requirements.

For processing technical and organizational documentation, Dinesh *et al.* (2007) propose the description of organizational procedures and the validation of their conformance to regulations, based on logical analysis. Lévy *et al.* (2010) present an environment that enables semantic annotations of document textual units (e.g., words, phrases, paragraphs, etc.) with ontological information (concepts, instances, roles, etc.). This approach provides an ontology-driven interpretation of the document contents.

Some work has been also developed to perform corpus-based analysis from several technical documents, as follows: for the use of frequency and concordance data from a corpus, Flowerdew (1993) work on English biology lectures; Lam (2007) propose the processing of English tourism

documents looking for pedagogical implications of its usage; and Henry and Roseberry (2001) observe English application letters.

In other lines of thought, we found language models accounting for documents oriented to audit linguistic expertise and analyze communicative and health texts (Fernández & García, 2009).

## 3 Exploration of MWEs in Corporate Documents

### 3.1 Corpus and Analysis Tools

We collected and analyzed a set of documents from the corporate domain in different subject fields such as medicine, forestry, and laboratory. The corpus used as the basis for this preliminary study consists of 25 English-written documents with independence of its variety.

The documents selected are a small sample belonging to the 'Job Specification Document' (JSD) category and were collected following representativeness and ecological criteria, *i.e.*, looking for the collection of documents produced, created, or promoted in the corporate or business environment. All the documents were taken from different corporations and sum 31627 tokens and 3839 types.

The initial exploration of this experimental corpus was supported by AntConc 3.3.5w® (Anthony, 2009) and TermoStatWeb™ (Drouin, 2003). AntConc was used to manually and systematically find frequent expressions and select their contexts, and TermoStatWeb™ was used to list most frequent verbs, nouns, and adjectives which could become part of MWEs.

### 3.2 Identification of Relevant MWEs

Relevant MWEs are identified in the experimental corpus according to the flow chart shown in Figure 1. From each technical document belonging to the corpus, we carried out the task of LP extraction (institutionalized expressions or lexicalized expressions) and classification (analysis by categories).

We classify the extracted expressions based on the document section where they prevail (see Table 1). Each section corresponds to a structural component of the JSD which also reflects the communicative intention of the writer.

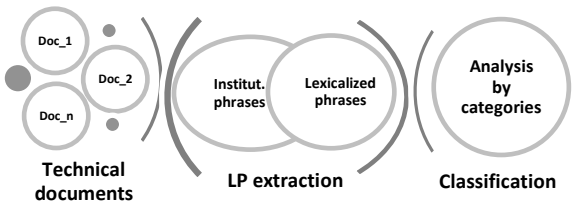


Figure 1. Flow chart for identifying MWEs

| *No.* | *Description section* |
|---|---|
| i | Job purpose / objective |
| ii | Main responsibilites / functions |
| iii | Knowledge and skills |
| iv | Requirements |

Table 1. Sections of JSD

Table 2 shows the relevant MWEs identified, as follows: i) the selected expressions with the corresponding MWE category (C) according to the classification proposed by Baldwin *et al.* (2010); ii) the frequency (F) of occurrence for each expression; and, iii) the section number (S) where the expression prevails in the JSD (from the Table 1).

| *C* | *MWEs* | | | *F* | *S* |
|---|---|---|---|---|---|
| 1. Statistically-idiomatic phrases | *be* | *Able* | *to* | 13 | iii |
| | *be* | *required* | *to* | 13 | ii |
| | *are* | *required* | *to* | 7 | iv |
| | *be* | *responsible* | *for* | 5 | ii |
| | - | *knowledge* | *of* | 49 | iii |
| | - | *experience* | *in* | 15 | iv |
| | - | *ability* | *to* | 61 | iii |
| | *related* | *duties* | *as* | 11 | Ii |
| | *the* | *duties* | *of* | 6 | ii |
| | *skills* | *and* | *abilities* | 11 | iii |
| | *level* | *experience* | - | 12 | iv |
| | *job* | *code* | - | 4 | i |
| | *job* | *description* | - | 9 | i |
| | *job* | *specification* | - | 7 | i |
| | *office* | *equipment* | - | 5 | ii,iii |
| | *working* | *relationships* | *with* | 12 | ii,iii |
| | *at* | *all* | *times* | 10 | ii |
| | *as* | *well* | *as* | 11 | ii |
| 2. Syntactically-flexible phrases | *be* | *[adquired]* | *on* | 5 | iv |
| | *to* | *[support]* | *the* | 29 | ii |
| | *the* | *[priority] and [schedule]* | *of* | 24 | ii,iii |
| | *the* | *[work] of* | *[others]* | 12 | iii,iv |
| | *by* | *[giv]ing* | *[time]* | 11 | iii,iv |
| | *in* | *[contacts] with the* | *[public]* | 13 | ii |
| 3. Semi-fixed phrases | - | *work* | *in* | 7 | ii,iii |
| | - | *work* | *of* | 6 | ii |
| | - | *work* | *with* | 5 | iii |
| | - | *may* | *be* | 30 | ii |
| | - | *may* | *have* | 5 | iv |
| | - | *follow* | *up* | 4 | i,ii |
| | - | *carry* | *out* | 9 | i, |

Table 2. Extracted MWEs

We use brackets for indicating semi-fixed phrases or variable uses of the expression (they can take values with the same conjugation). In this way, we identify and prioritize the most frequents MWEs and patterns in each category, as follows:

1. *ability to*, *knowledge of*, *experience in*, *be able to*, *be required to*
2. *to-V-the*, *the-N-and-N-of*, *in-N-with-the-N*
3. *may be*, *carry out*, *work in*, *work of*

Likewise, we also found useful identifying the most frequent lexical items that could become part of MWEs and alternate with the expressions and patterns presented above. For that purpose, TermoStatWeb was used to generate a map with the most frequent verbs, nouns, and adjectives. Some examples are shown in Figure 2.

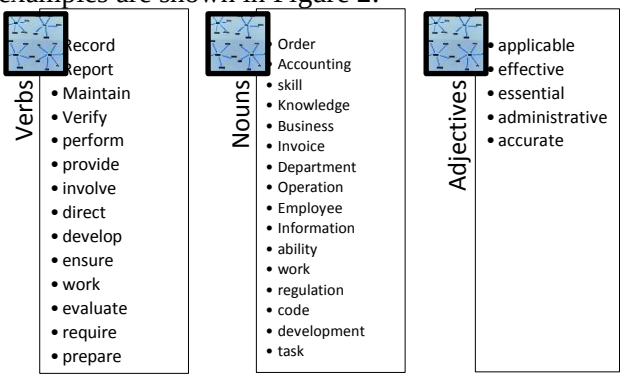


Figure 2. Some frequent verbs, nouns, and adjectives.

The high frequency of these items in the corpus suggests that they could probably be part of MWEs conveying corporate information. Also, when placed in the slots of the patterns observed in Table 2, they increase their chance to become relevant MWEs useful to detect specific corporate knowledge.

The following paragraph is an example of how this can happen. The source text belongs to a JSD from our corpus and shows how two frequent items (*evaluate* and *work*) co-occur in a collocation. Then, identified corporate information is expected to be generated by other means into specific organizational information in a controlled language:

Source paragraph

*…A* City Manager *plans*, *organizes*, *evaluates*, *and controls the* ***work*** *of all* City departments *to ensure that* ***operations*** *and* ***services*** *comply with the* ***policies****…*

Generated organizational information:

[City_manager *plans* ***work***. City_manager *organizes* ***work***. City_manager *evaluates* ***work*** City_manager *controls* ***work***] [City_department *has* ***work***] [City_manager *ensures* ***operations***] [City_department *has* ***operations***] [City_department *has* ***services***] [***operations*** *comply* ***policies***]

In terms of organizational knowledge, an analyst can find information from JSDs about roles, responsibilities, actions, and constraints, as an approach for understanding an organizational domain. Such entities are expressed in a JSD as subject, actions, and object triples, as suggested by some instances in Table 2. This information can be represented either into models or controlled language discourses, among other specifications.

## 4 Conclusions

This study aims at characterizing JSDs by revealing key MWEs used in an English corpus. We proposed a set of MWEs of a JSD, as a corporate technical document, which can be processed as input for further knowledge engineering processes. The appropriateness of JSDs in requirements elicitation was verified with this study.

The analysis shows frequencies and patterns of relevant MWEs as well as their contexts and inflectional forms extracted via a concordance tool. The performed analysis is a preliminary study for knowledge acquisition and understanding of organizational domains. Such knowledge is expected to be readily available to future applications in specific domains in order to validate the findings and then to automate the process.

As future work, we expect to increase the number of documents in the corpus and refine the study of lexical and textual features. Statistical association measures can be also considered as a way to reinforce MWEs and term identification and extraction in the frame of knowledge acquisition from corporate documents. Likewise, given the importance of the syntactic structure given by the triple subject-verb-object, dependency parsing seems to be a promising approach for the identification of roles and responsibilities in JSDs.

## Acknowledgments

This work is funded by the Vicerrectoría de Investigación from both the Universidad de Medellín and the Universidad Nacional de Colombia, under the project: "*Método de transformación de lenguaje natural a lenguaje controlado para la obtención de requisitos, a partir de documentación técnica*".

## References


Anthony, L. 2009. *Issues in the design and development of software tools for corpus studies: The case for collaboration*. Contemporary corpus linguistics, London: P. Baker Ed.: 87-104.

Aussenac-Gilles, N. Biébow, B. and Szulman, S. 2000. *Revisiting Ontology Design: A Method Based on Corpus Analysis. Knowledge Engineering and Knowledge Management*. Methods, Models, and Tools, 1937:27–66.

Baldwin, Timothy and Su Nam Kim (2010) Multiword Expressions, in Nitin Indurkhya and Fred J. Damerau (eds.) Handbook of Natural Language Processing, Second Ed., CRC Press, USA, pp. 267-292.

Bannard, C. 2005. Learning about the meaning of verb–particle constructions from corpora. *Computer Speech & Language*, 19(4): 467-478.

Bauer, L. 1983. *English Word-Formation*. London: Cambridge University Press, 311.

Breaux, T.D., Vail, M.W. and Antón, A.I. 2006. Towards Regulatory Compliance: Extracting Rights and Obligations to Align Requirements with Regulations. North Carolina State University TR-2006-6.

Cascini, G. Fantechi, A. and Spinicci, E. 2004. Natural Language Processing of Patents and Technical Documentation. *Lecture Notes in Computer Science*, 3163:508-520.

Castro-Herrera, C., Duan, C., Cleland-Huang, J. and Mobasher, B. Using data mining and recommender systems to facilitate large-scale, open, and inclusive requirements elicitation processes. *Proceedings of 16th IEEE Inter. Requirements Eng. Conference*, pp.165–168, 2008.

Dinesh, N. Joshi, A. Lee, I. and Sokolski, O. 2007. *Logic-based regulatory conformance checking*. In 14th Monterey Workshop, ScholarlyCommons Penn.

Drouin, P. 2003. *TermoStat Web 3.0*. Désormais utilisable qu'après enregistremen. Available in: http://olst.ling.umontreal.ca/~drouinp/termostat_web/

Fernández, L. and García, F.J. 2009. Texto y empresa. *Applied Linguistics Now: Understanding Language and Mind*, pp.655-665. Universidad de Almería, España.

Flowerdew, J. 1993. Concordancing as a Tool in Course Design. *System*, 21(2): 231-244.

Henry, A. and Roseberry, R.L. 2001. *Using a Small Corpus to Obtain Data for Teaching a Genre*. In Ghadessy/Henry/Roseberry: 93-133.

Jackendoff, R. 1997. *The architecture of the language faculty*. MIT Press, Cambridge, MA, USA.

Lam, P. Y. 2007. A *Corpus-driven Léxico-grammatical Analysis of English Tourism Industry Texts and the Study of its Pedagogic Implications in ESP*. In Hidalgo/Quereda/Santana: 71-90.

Lee, B. and Bryant, B. R. 2002. *Contextual Natural Language Processing and DAML for Understanding Software Requirements Specifications*. In 19th International Conference on Computational Linguistics, Taipei, Taiwan.

Levy, F. Guisse, A. Nazarenko, A. Omrane, N. and Szulman, S. 2010. An Environment for the Joint Management of Written Policies and Business Rules. 22nd IEEE International Conference on Tools with Artificial Intelligence. *IEEE Computer Society*, 2:142-149.

Li, K., Dewar, R.G. and Pooley, R.J. Requirements capture in natural language problem statements. Heriot-Watt University, 2003. Available in http://www.macs.hw.ac.uk:8080/techreps/docs/files/HW-MACS-TR-0023.pdf

López-Mezquita, M.T. 2007. La evaluación de la competencia léxica: tests de vocabulario. Su fiabilidad y validez. *Centro de Investigación y Documentación Educativa*, 177(1): 488.

López Rodríguez, C. I., Faber, P., León- Araúz, P., Prieto, J. A. and Tercedor, M. 2010. La Terminología basada en marcos y su aplicación a las ciencias medioambientales: los proyectos MarcoCosta y Ecosistema. *Arena Romanistica*, 7 (10): 52-74.

Peleg, M. Gutnik, L.A. Snow, V. and Patel, V.L. 2005. Interpreting procedures from descriptive guidelines. *Journal of Biomedical Informatics*, 39(1):184–195.

Perez, C. 1999. *La enseñanza del vocabulario desde una perspectiva lingufstica y pedagogica*. In S. Salaberri (Ed.), Lingutstica Aplicada a las Lenguas Extranjeras, Almería: Univ. de Almería: 262-307.

Rösner, D., Grote, B., Hartmann, K. and Höfling. B. 1997. From Natural Language Documents to Sharable Product Knowledge: A Knowledge Engineering Approach. Journal of Universal Computer Science. 3(8): 955-987.

Sinclair, J. 1991. Corpus, concordance, collocation. Oxford: Oxford University Press.

Soler, C. and Gil, I. 2010. Posibilidades y límites de los tesauros frente a otros sistemas de organización del conocimiento: folksonomías, taxonomías y ontologías. *Revista Interamericana de Bibliotecología*, 33(2): 361-377.