Quantitative evaluation of the movement from complexity toward simplicity in the structure of thesaurus descriptors

Maziar Amirhosseini[1*] and Juhana Salim [1,2]
[1]Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, MALAYSIA.
[2]Centre for Collaborative Innovation
Universiti Kebangsaan Malaysia (UKM-CCI), MALAYSIA.
e-mail: *mazi_lib@yahoo.com(corresponding author); js@ftsm.ukm.my

ABSTRACT

The concepts of simplicity and complexity play major roles in information storage and retrieval in knowledge organizations. This paper reports an investigation of these concepts in the structure of descriptors. The main purpose of simplicity is to decrease the number of words in the construction of descriptors as this idea affects semantic relations, recall and precision. ISO 25964 has affirmed the purpose of simplicity by requiring splitting compound terms into simpler concepts. This work aims to elaborate the standard methods of evaluation by providing a more detailed evaluation of the descriptors structure and identifying effective factors in simplicity and complexity results in the structure of thesauri descriptors. The research population is taken from the descriptors of the Commonwealth Agricultural Bureaux (CAB) Thesaurus, the Persian Cultural Thesaurus (ASFA) and the Chemical Thesaurus. This research was conducted using the statistical and content analysis method. In this research we propose a new quantitative approach as well as novel indicators and indices involving Simplicity and Factoring Ratios to evaluate the descriptors structure. The results will be useful in the verification, selection and maintenance purposes in knowledge organizations and the inquiry method can be further developed in the field of ontology evaluation.

Keywords: Thesaurus quantitative evaluation; Simplicity; Complexity; Descriptors' structure; Simplicity ratio; Factoring ratio.

INTRODUCTION

Gell-Mann (2002) defines "simplicity" through its opposite meaning i.e. "complexity" based on plectics i.e. the study of complexity and simplicity. In this article, simplicity and complexity concepts are analyzed in the structure of descriptors in knowledge organizations. Simplicity concept implies to the condition of decreasing the number of words in descriptors structure. In contrast, an increase in the number of words in descriptors structure results in constructing complex descriptors. In this matter, single-word descriptors are simpler than multi-word ones. Word is a single distinct conceptual unit of language (Word 2012). In this text, "Word" is considered as a smaller indexing unit in terms of information storage and retrieval. For instance, "*Information Management Systems*" is a more complex indexing unit than

"*Information*" as a single-word. Therefore, the number of words in descriptors structure plays a tremendous role as a factor to measure the simplicity and complexity in descriptors structure.

The establishment and development standards of thesauri have taken into account the structure of descriptors in terms of simplicity and complexity. ISO 25964 emphasizes that **"**absolute consistency in the admission of complex concepts" (i.e., compound terms) is difficult to achieve and is not always necessary. It is a general rule that descriptors should represent simple concepts in the form of single words (NISO 2005; ISO: 25964, 2011), and compound terms should split into simpler or unitary concepts to become easily understandable (ISO: 25964 2011; Holm 2007). Nevertheless, identifying the simple concept is a complex process in descriptors structure and is closely related to its subject characteristics. For instance, unit concepts such as "*Vitamin B*" and "*Black Market*" are the exceptions in the process of compound terms splitting. Consequently, the standards emphasized the use of single words as descriptors and decrease of the number of words in descriptors structure for the sake of simplicity to "compensate for the vocabulary control problems" (NISO 2005) which are caused by complex descriptors.

The extraordinary role of the single descriptors for the sake of simplicity to improve information retrieval effectiveness (Leveling et al. 2011) can be cleared by these examples selected from UNESCO Thesaurus (2004): "*Information*", "*Management*", "*Resources*", "*Sciences*" and the like. In this matter, single descriptors consist of generic concepts to develop the rate of recall. Additionally, the composition of the single descriptors in IR time results in making the semantic linkages between different subject fields, such as "*Information Resources Management*", "*Information Management*", "*Management Information Resources*", "*Management Information*", "*Resources Management*" and so on. Moreover, these compositions are reasons for increasing precision on basis of syntactic relations in information retrieval time. Furthermore, a few numbers of descriptors can be connected to one another through the numerous linkages. Consequently, an increase in the number of simple or single descriptors could be a reason for increasing the semantic relations, recall and precision while there is a decrease in the number of descriptors in vocabulary.

In addition to the role of the single descriptors in improving information retrieval, single words and the number of words in descriptors structure can be considered as the criteria (i.e. factors) to measure the simplicity and complexity in the structure of descriptors in the field of thesaurus quantitative evaluation. In this matter, this investigation attempts to measure simplicity concept on basis of the number of words in descriptors structure in the Commonwealth Agricultural Bureaux (CAB) thesaurus (CABI 1995), the Persian Cultural Thesaurus (ASFA Thesaurus) (ASFA 2001) and the Chemical Thesaurus (SDIC 2004).

LITERATURE REVIEW

This section explains quantitative investigations, especially proportional and metric analyses in analyzing the structural domains in knowledge organizations. The quantitative evaluation of thesaurus began with Kochen and Tagliacozzo's (1968) works using connectedness ratio and accessibility measure. The 1960s and 70s were the start of the construction of some original ratios which measured some factors in the structure of controlled vocabularies. Thesaurus

quantitative evaluation has been carried out much earlier than 1976. Some statistical measures had been proposed before 1976. Ratios, in fact, are a part of statistical measures in thesaurus quantitative evaluation (Lancaster 1986).

Simplicity analysis in descriptors structure has been taken into account by several research articles in terms of decreasing the number of compound terms via splitting techniques (Pohlmann and Kraaij 1997; De Vries 2001; Hollink et al. 2004) to increase recall and precision (Braschler and Ripplinger 2004; Airio 2006; Lazarinis et al. 2009; Leveling et al. 2011) as well as increase a maximum possible number of semantic relations (Muñoz 1997). Furthermore, a highly specific vocabulary, which is very often clarified by compound descriptors, results in increasing the number of indexing terms (Aitchison et al. 2000).

Burton-Jones et al. (2003) focused on developing a set of syntactic, semantic, and pragmatic constructs to assess concept and its semantic relations through proposed metrics. In 2004, structure complexity measure has been proposed by Kang et al. (2004) to evaluate the complexity of both classes and relationships between the classes in knowledge organizations. His (2005) presented two metrics to measure conceptual coherence and conceptual complexity in semantic network. In the same year, a complexity measure i.e. path-to-term ratio proposed to evaluate concepts and their semantic relations (Mungall 2005).

Airio (2006) believed that if the full form of compound term is used as descriptors in indexing and not its parts, this will cause problems because the query includes only parts of the compounds. In 2006, a research was conducted on the basis of His' (2005) investigation. This inquiry which was done by Zhang, Ye, and Yang (2006) focused on evaluation of ontology complexity with regard to evaluate concepts and their hierarchical relations. Yang, Zhang, and Ye (2006) then published a research in relation to Mungall's (2005) inquiry. They examined the concepts and their hierarchy in an ontology conceptual model which reflects the complexity.

The results of the proposed ratios and metrics were useful in evaluating the structure of knowledge organizations in the form of quantitative evaluation approach. Based on this approach, original measures were designed to analyze some factors for the evaluation of thesaurus structure that focused on one concept alone. However, their failure to consider other relative concepts in the evaluation of thesaurus structure limited the usefulness of this approach for thesaurus structure evaluation. Moreover, the proposed measures have generally focused on evaluating descriptors and their semantic relations. Hence, structural analysis should be evaluated in detail via complement and relevant ratios, especially in assessing the structure of descriptors.

Therefore, in 2007, simplicity and factoring ratios were proposed as two complement ratios to analyze simplicity concept in descriptors' structure (Amirhosseini 2007), and a theoretical basis was proposed to construct ratios in capturing cognition results in the form of modern quantitative evaluation (Amirhosseini 2010). Amirhosseini and Salim (2010) presented two levels of simplicity invisible domains in the structure of Islamic thesauri descriptors. The research results showed that there is an indirect relation between the number of words in the structure of descriptors and simplicity.

RESEARCH OBJECTIVES

In this study, we intend to measure simplicity concept on basis of the number of words in descriptors structure for the sake of generating better thesauri in terms of information storage and retrieval. In this case, analysis of the descriptor structure to identify effective factors on increasing and decreasing simplicity in descriptors structure for the sake of increasing information retrieval performance is the main purpose of this research. The research objectives raised in this paper are as follows:

a) To recognize the condition of simplicity in the descriptors' structure of the thesauri.
b) To analyse the condition of simplicity between descriptors which have various numbers of words in their structures.
c) To clarify the kinds of relations that can be found between the complexity and simplicity in the simplicity and factoring ratios' results.
d) To define the movement from complexity towards simplicity in the simplicity and factoring ratios' results.
e) To identify the important effective factor in the results of simplicity in the descriptors' structure.
f) To explain the significant roles of the simplicity analysis in the descriptors' structure in the information retrieval performance.
g) To demonstrate the position of proportional analyses in the field of ontology evaluation.

METHOD

In this research, the statistical and content analysis method was conducted. Content analysis is a systematic reading of a body of texts that predicts or infers phenomena that cannot be observed directly (Krippendorff 2004). Content analysis includes five main steps: selection of topic, determination of source, identification of data, extraction and analysis of data and reporting of results. These steps are identified in Table 1.

Table 1: Steps in Content Analysis

Description	
Steps involve in the content analysis	Steps identified in the content analysis of descriptors structure
Selection of topic	Analysis of simplicity in the structure of descriptors
Determination of source	Content of the CAB, ASFA and Chemical thesauri
Identification of data	Simple and compound descriptors
Extraction and analysis of data	Quantitative approach involving statistical methods for data extraction. Data analysis involves using proportional analysis
Reporting of results	Results are synthesized and demonstrated in the form of tables, diagrams and figures.

The research topic is the analysis of simplicity in the structure of descriptors. The determined sources are the content of the CAB, ASFA and Chemical thesauri. The research data involves simple and compound descriptors, while the quantitative approach involves statistical

methods for data extraction. Furthermore, data analysis is done through proportional analysis, and lastly, results are synthesized and demonstrated in the form of tables and figures.

The research population consists of the descriptors in the aforementioned thesauri. Sample selection was done based on stratified random sampling. Sample sizes in the CAB, ASFA, and Chemical thesauri were 11127, 4555 and 1828 descriptors, respectively. MS-EXCEL was used for data gathering and analysing.

FINDINGS

In this section, we explained indicators and indices of simplicity and factoring ratios as our data analysis method and demonstrated the results which are derived from the process of the data analysis.

Simplicity Ratio

This ratio analyzes the simplicity in the structure of knowledge organization descriptors. It conveys that Simplicity Ratio measures the amount or proportion of the single and compound descriptors' size in relation to the whole of the descriptors in a thesaurus. The indicators of the simplicity ratio are the number of simple or single descriptors and the total number of descriptors.

$$SR = \frac{a}{b}$$

$a = the\ Number\ \ of\ \ Simple\ \ Descriptors$

$b = the\ \ Total\ \ Number\ \ of\ \ Descriptors$

According to thesauri standards, thesaurus builders should construct simple or single descriptors. Accordingly, the number of simple or single descriptors usage in the structure of descriptors is an important factor in quantitative evaluation of simplicity in descriptors structure in order to comply with thesauri construction standards. Therefore, the results of simplicity in descriptors structure will change if the number of the simple or single descriptors increases or decreases in vocabulary. Consequently, the amount of the single descriptors as a factor plays a tremendous role in changing the results of simplicity in descriptor structure of thesauri. Table 2 demonstrates the results of the simplicity ratio in CAB, Chemical and ASFA thesauri.

Table 2 shows that the simplicity result in the CAB thesaurus is equal to 0.3881 (i.e., 4319 / 11127 = 0.3881) which is the best result of simplicity in the thesauri. Simplicity of chemical thesaurus is equal to 0.2007 and ASFA is 0.2178. The result of simplicity in the thesauri implies that about 40 per cent of descriptors in the CAB, 20 percent in the Chemical and 22 per cent in the ASFA thesauri have simple structures. In spite of standards which emphasize the use of single, simple and unitary words in the construction of descriptors, the result in this research shows that the amount of compound descriptors are more than simple ones in the thesauri.

Table 2: Simplicity Ratio Results in the Thesauri

Thesauri / Indicators	CAB	Chemical	ASFA
The number of simple descriptors	4319	316	789
The total number of descriptors	11127	1574	3622
Ratio results	0.3881	0.2007	0.2178

Factoring Ratio

The constructing and proposing idea of the Factoring Ratio is derived from standards (i.e., BS 5723, ISO 2788, ISO 25964 and ANSI/NISOZ39.19) which emphasize the compound words splitting. In this matter, the complex concept which is usually conveyed by a multi-word term should split into simpler concepts (ISO: 25964 2011) or separate components (ISO: 2877 1986) to become a more easily understandable option (ISO: 25964 2011). The individual components should be assigned as separate indexing terms (TESE 2006) with the exception of the conditions which affect the users' understanding (Aitchinson 1990). For example, "human resource management" could be conveyed by "human resources + management" or "people + resource management" or even "people +resources + management" (ISO: 25964 2011). Breaking down complex terms into several unitary concepts to move toward simplicity is known as splitting or factoring. Hence, the amount of descriptors with fewer words should be more than the descriptors with more words in their structures.

The Factoring Ratio is proposed to achieve three main objectives: first, to propose a simplicity ratio's complement to interpret and explain simplicity results; second, to analyze the movement from complexity toward simplicity in the structure of descriptors; and third, to identify effective factors on increasing or decreasing simplicity results in the structure of descriptors. The Factoring Ratio analyzes the simplicity between two kinds of descriptors as a specific domain, which have sequenced value in the number of words in their structures (e.g. simplicity analysis between two-word and three-word descriptors). In other words, Factoring Ratio measures the proportion of the size of two kinds of descriptors, which have sequenced value in the number of words in their structures (e.g., one-word and two-word descriptors) in relation to the total number of these descriptors in the whole of the descriptors in a thesaurus. The indicators of the factoring ratio are the number of n-word descriptors and the number of (n+1)-word descriptors.

$$FR = \frac{a}{a+b}$$

$a = the\ Number\ \ of\ \ n-word\ \ Descriptors$

$b = the\ \ Number\ \ of\ \ (n+1)-word\ \ Descriptors$

As stated previously, Factoring Ratio analyzes simplicity in specific domains of the descriptors structure which have various numbers of words in their structure. When the intention is to analyze the Factoring Ratio between two-word and three-word descriptors in a thesaurus, we should calculate the amount of the two-word and three-word descriptors. The number of two-word descriptors, for instance, is equal to 60, and the number of three-word descriptors is

equal to 40. Therefore, the Factoring Ratio for these values is FR=60 / (60 + 40) = 0.6. In conclusion, this result means that our thesaurus has 6 two-word descriptors in every 10 descriptors, which have two and three words in their structure.

Factoring Ratio Results

Table 3 shows the results of the Factoring Ratio between descriptors which have various words in their structure in the CAB, Chemical and ASFA thesauri.

Table 3: The Factoring Ratio Results in the Thesauri

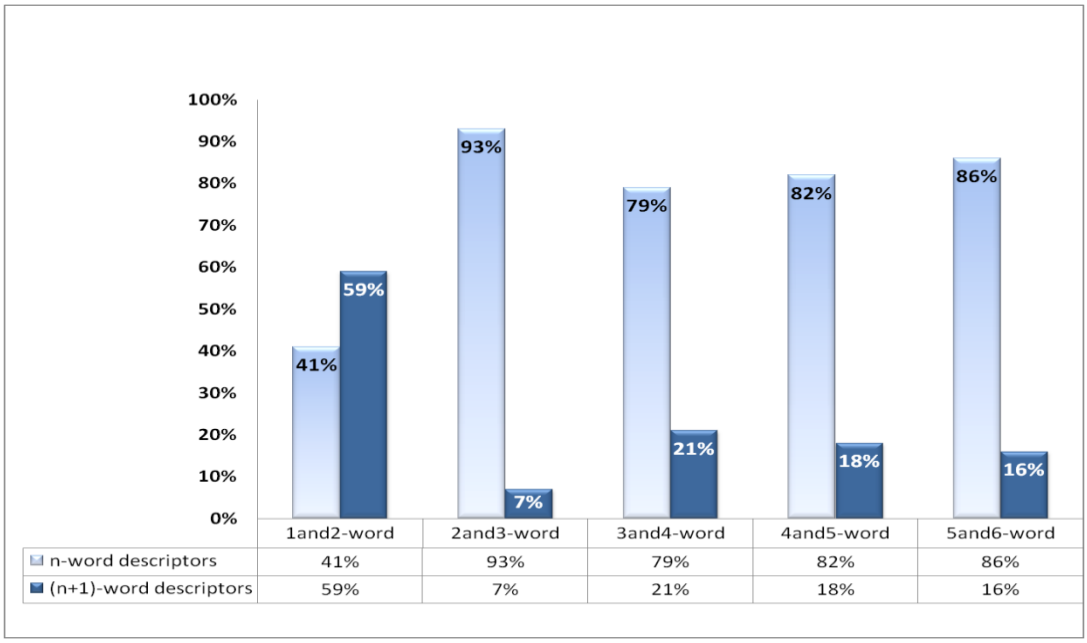
Thesaurus n-word descriptors	**CAB**		**Chemical**		**ASFA**	
	No.	Ratio	No.	Ratio	No.	Ratio
1-word	4319	0.41	316	0.25	789	0.26
2-word	6198	0.93	926	0.77	2201	0.79
3-word	460	0.79	275	0.85	569	0.9
4-word	120	0.82	49	0.89	61	0.97
5-word	26	0.87	6	0.75	2	
6-word	4		2			

At first glance, Table 3 shows that the number of two-word descriptors is more than the others. Factoring Ratio results of one-word descriptors, for instance, is 0.41 (i.e., 4319/ (4319+6198)) in the CAB thesaurus. This means that there are 41 one-word descriptors in every 100 of the one and two-word descriptors. Meanwhile, the best result of the Factoring Ratio belongs to the CAB thesaurus in the relation of the one and two-word descriptors. Furthermore, the number of two-word descriptors specifically is more than the one-word descriptors in the thesauri. However, this relation is conversed in the other relations between descriptors which have various numbers of words in their structure in the thesauri. This implies that in comparing the relation between one and two-word descriptors, the number of three-word descriptors is less than the two-word descriptors, the number of four-word descriptors is less than the three-word descriptors and the like. In the CAB thesaurus, for instance, the Factoring Ratio result of the two-word descriptors is 0.93 (i.e., there are 93 two-word descriptors in every 100 of the two and three-word descriptors).

Visual display of Factoring ratio results*:* Figures 1 to 3 demonstrate Factoring Ratio results in the CAB, Chemical and ASFA thesauri. The Factoring Ratio results in the CAB, Chemical and ASFA thesauri are almost alike. Figures 2 and 3 show that the number of one-word descriptors is less than two-word descriptors in the Chemical and ASFA thesauri. However, there is a conversed relationship between descriptors. For instance, the number of two-word descriptors is more than the number of three-word descriptors, the number of three-word descriptors is more than the number of four-word descriptors and so on.

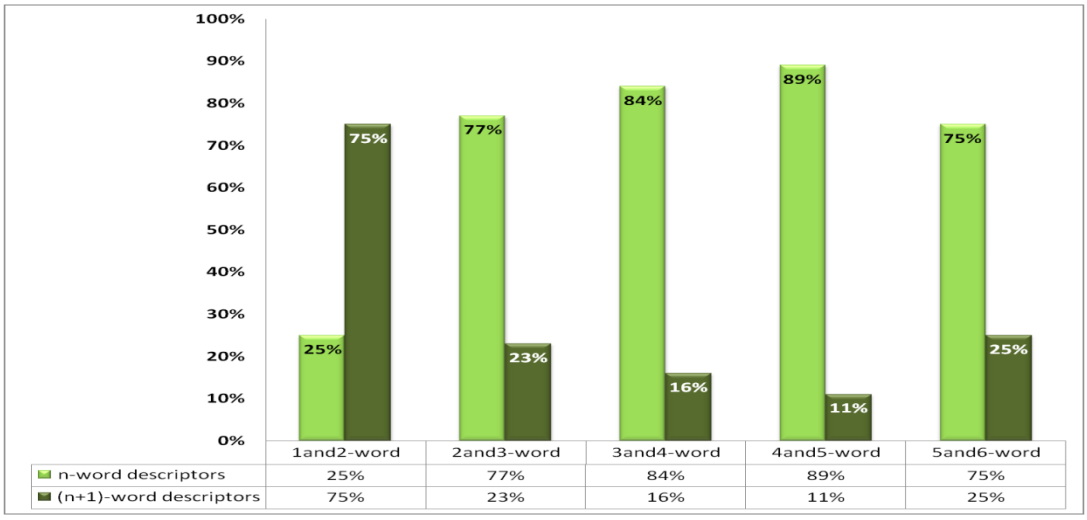
Figure 1 shows that the number of one-word descriptors (i.e. 41 per cent) is less than two-word descriptors (i.e. 59 per cent) in the CAB thesaurus. However, this relation is conversed in other relations between descriptors which have various numbers of words in their structures. In this way, the number of two-word descriptors is more than three-word descriptors (i.e. 93 per cent to 7 per cent), the number of three-word descriptors is more than the number of

four-word descriptors (i.e. 79 per cent to 21 per cent) and the like. Meanwhile, the results of the Factoring Ratio in the Chemical and ASFA thesauri are illustrated in Figures 2 and 3.



	1and2-word	2and3-word	3and4-word	4and5-word	5and6-word
n-word descriptors	41%	93%	79%	82%	86%
(n+1)-word descriptors	59%	7%	21%	18%	16%

Figure 1: Factoring Ratio Results of Descriptors Structure In CAB Thesaurus



	1and2-word	2and3-word	3and4-word	4and5-word	5and6-word
n-word descriptors	25%	77%	84%	89%	75%
(n+1)-word descriptors	75%	23%	16%	11%	25%

Figure 2: Factoring Ratio Results of Descriptors Structure in Chemical Thesaurus

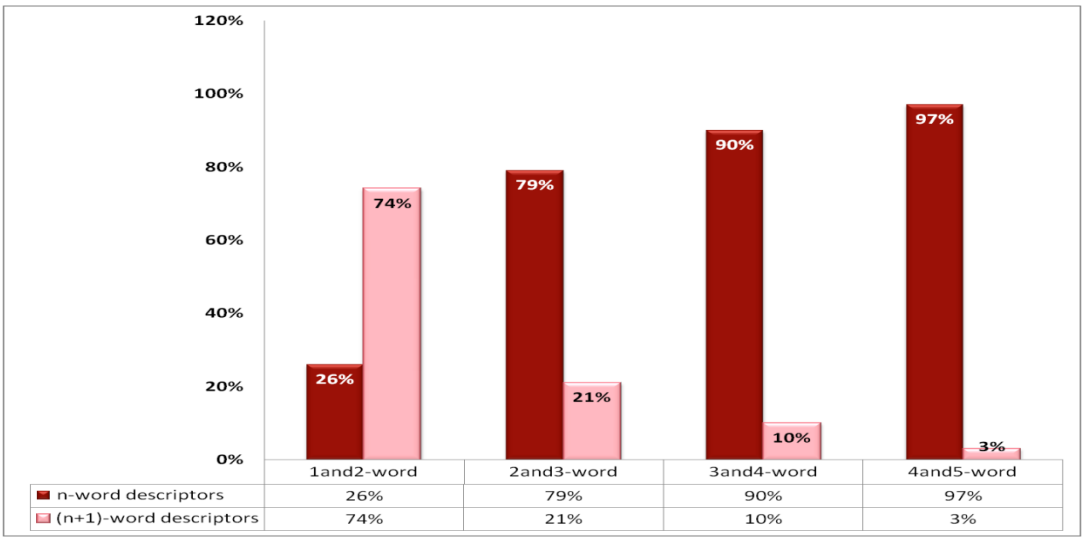


Figure 3: Factoring Ratio Results of Descriptors Structure in ASFA Thesaurus

DISCUSSIONS

In this section, the role of simple word and compound word criteria in identifying the simplicity and complexity of the descriptors' structure are explained. As stated previously, thesaurus construction standards emphasized the use of single, simple and unitary words as a descriptor. It implies that in this ground, in most of the cases, only a single word can be considered as a criterion to evaluate simplicity in the structure of descriptors. Thus, compound words have to be factored (i.e. split) into simple elements so that the structure of descriptors moves from a complex to a simpler form. On the other hand, since simplicity is developed in the structure of descriptors, whenever complex structure of compound terms is factored into separate components, a compound word is an appropriate criterion to identify the complexity of the descriptors structure. Therefore, in this section, we consider single word as the criterion to identify simplicity and the compound word as the criterion to identify the complexity in the structure of descriptors.

Condition of Simplicity in the Descriptors' Structure

The simplicity results in Table 2 showed that thesauri builders did not properly move towards simple and unitary word usage, especially in the Chemical and ASFA thesauri. Here, we intend to show the movement from complexity towards simplicity by analyzing the Simplicity Ratio result in the CAB thesaurus.

Figure 4 shows the complexity and simplicity result on a percentage basis. If we intend to analyze the movement from complexity toward simplicity in the Simplicity Ratio's results, we should start from complexity towards simplicity in the above axis and find the result of simplicity in the structure of the descriptors. This implies that whenever the result of the Simplicity Ratio increases, the simplicity of the descriptors' structure increases as well in the vocabulary as a whole. As Figure 4 demonstrates, the simplicity result of the CAB thesaurus is equal to 0.39, that is: the domain of simple, single and unitary descriptors usage is less than

the equivalent point (i.e., 50 per cent) between simple and compound descriptors usage. Ultimately, there is no balance between simple and compound descriptors' usage.

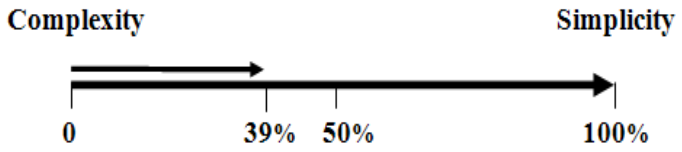


Figure 4: The Movement from Complexity toward Simplicity in the Simplicity Ratio Result of the CAB Thesaurus

Condition of Simplicity between Descriptors which have Various Numbers of Words in their Structure

The Factoring Ratio results in Table 3 and Figures 1 to 3 show that the simplicity results in the structure of descriptors which have various numbers of words in their structures were almost the same in the thesauri. Results show that the number of two-word descriptors is more than the other descriptors. Specifically, the number of two-word descriptors is more than one-word descriptors. In spite of the upward trend of two-word descriptors in comparison with one-word descriptors, this relation is conversed in the other relations. This means that the number of two-word descriptors is more than three-word descriptors, the number of three-word descriptors is more than four-word descriptors and so on. In general, the amount of the descriptors with fewer words is more than the descriptors with more words in their structures with the exception of the relation between one-word and two-word descriptors. Subsequently, these relations between descriptors which have various numbers of words in their structure are reasons for increasing simplicity in the descriptors' structure with the exception of the relation between one-word and two-word descriptors.

Types of Relations between Simplicity and Factoring Ratios' Results

The research results show that the number of compound terms and the number of words in the structure of descriptors are the criteria to analyze the movement from complexity toward simplicity in the descriptors' structure. In this matter, if the number of compound terms and the number of words which are used in the structure of descriptors decrease through breaking down compound descriptors into several unitary descriptors, the complexity of descriptors' structure will decrease while the simplicity will increase and vice-versa. This idea can be clearly explained by presenting an example which is demonstrated as follows: "*Integrated Information Systems*" as a compound descriptor (i.e., complex structure) should be factored or split into simple elements and several unitary descriptors, such as "*Integrated Information" and "Systems*" or "*Integration*", "*Information*" and "*Systems*". Then simplicity can be increased in the structure of descriptors. All in all, the relation between the number of compound terms and the words used in the structure of descriptors with complexity is a direct relation while it has an indirect one with simplicity. Consequently, the effective factor to increase simplicity and decrease complexity is the structure of descriptors in terms of their simple and compound structure.

Movement from Complexity toward Simplicity in the Simplicity and Factoring Ratios' Results

Figure 5 illustrates the movement from complexity toward simplicity between the descriptors which have various numbers of words in their structure in the CAB thesaurus.

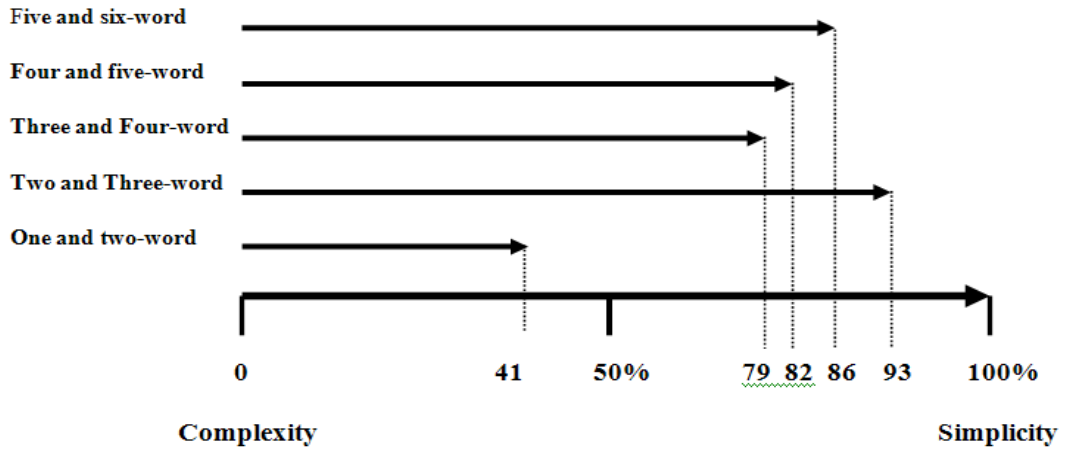


Figure 5: The Movement from Complexity toward Simplicity in the Factoring Ratio Results in CAB Thesaurus

Figure 5 shows that the Factoring Ratio result between one-word and two-word descriptors is less than 50 per cent (i.e. 41 per cent one-word and 59 per cent two-word), but the relation between the results of two-word descriptors and three-word descriptors, three and four-word descriptors, etc. are conversed (e.g. 79 per cent three-word and 21 per cent four-word descriptors in the relation between two and three-word). Subsequently, these results are more than the equivalent point (i.e. 50 per cent), and the movement from complexity toward simplicity was fulfilled in the construction of these kinds of descriptors with the exception of the relation between one and two-word descriptors.

Here, we intend to compare the result of the Simplicity Ratio with the results of the Factoring Ratio in the relation between one-word and two-word descriptors, which is drawn in Figure 6.

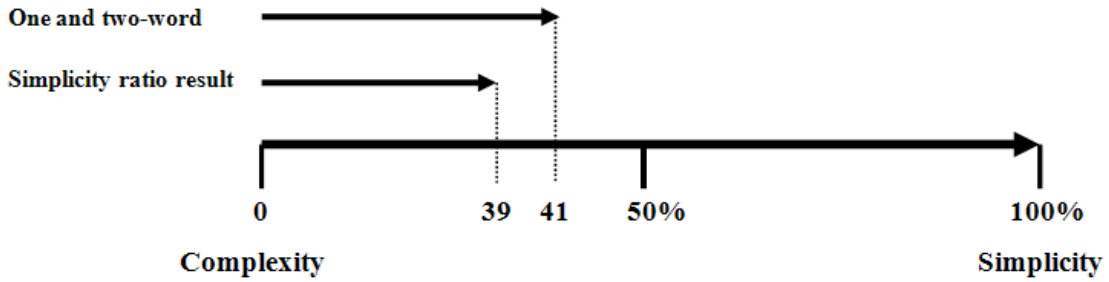


Figure 6: Comparison of the Simplicity Result and the Relation between One-word and Two-word Descriptors in the CAB Thesaurus

Figure 6 demonstrates that the Simplicity Ratio result and the Factoring Ratio result between one-word and two-word descriptors are almost similar in the CAB thesaurus. The Simplicity Ratio measures the simplicity concept in the whole of the vocabulary (i.e., a macro viewpoint), and the Factoring Ratio result between one and two-word descriptors measures the simplicity concept between the one and two-word descriptors (i.e., a micro viewpoint). The similar results in Figure 6 can be analyzed into two perspectives. In the former, the results of the relations between the two and three-word descriptors, three and four-word descriptors, and so on which could not have a major effect in changing the simplicity result, were more than the equivalent point (i.e. 50 per cent) in Figure 5. Later, the cause of decreasing the simplicity result can be found in the relation between one and two-word descriptors because of the large number of two-word descriptors in the CAB thesaurus (see Table 3).

Importance of Effective Factor in the Results of Simplicity

Simplicity results show that the thesauri builders have not properly moved to simplicity in the construction of descriptors. This means that the domain of the single descriptors' usage is less than compound ones. If we only rely on the Simplicity Ratio results, we can conclude that the results in the thesauri are not acceptable. However, in a comparison between Simplicity and Factoring Ratio results, we can analyze, describe and interpret the Simplicity Ratio results. In general, the results of the Factoring Ratio showed that the thesaurus builders move from complexity to simplicity in descriptors construction with the exception of the relation between one-word and two-word descriptors. This relation shows that the number of compound words (i.e. two-word descriptors) is more than simple ones. Furthermore, the number of two-word descriptors is more than the other descriptors which have various numbers of words in their structures. As a consequence, the effective factor to explain the result of Simplicity Ratio and an increase of the simplicity results in the thesauri is the two-word descriptors.

Significant Roles of Research Idea in the Information Retrieval Performance

The focal point of the research idea is the usage of simple and single words as descriptors and the split of compound terms into their constituents. In this case, the number of words in the structure of descriptors should be decreased to construct simpler descriptors. This idea was the main source to propose Simplicity and Factoring Ratios in measuring the simplicity concept in descriptors structure. The simplicity analysis method can operate in analysing the structure of descriptors in various kinds of thesaurus. The research results, in fact, reported the condition of the descriptors structure in considered thesauri in terms of simplicity and complexity concepts. Moreover, the research idea comprises the tremendous roles in information retrieval performance. In this section, we intend to discuss on the functions of the research idea in improving the information retrieval performance in terms of recall, precision and also the number of semantic relations and descriptors in vocabulary.

Recall Rate

Increasing simple or single descriptors results in increasing recall rate. In this matter, simple and single descriptors consist of generic concepts to develop the rate of recall because a single-word descriptor which is not specified for a specific subject field can be linked to numerous descriptors. For instance, "*Information*" as a generic concept consists of a capacity to link with several narrower and related terms such as "*Communication Information*", "*Cultural Information*", "*Educational Information*", "*Scientific Information*", "*Social Science*

Information" , "*Information Transfer*", "*Information Users*", "*Knowledge*" and so on. On the other hand, a specific descriptor such as "*Information System Evaluation*" very often has not a chance to link with several descriptors. Consequently, generic descriptors in the form of single-word descriptors guarantee the recall rate to improve the performance of information retrieval. In contrast, a specific descriptor is not a requirement for increasing recall.

Precision Rate

Compound descriptors and concepts play the role for increasing precision in process of the information retrieval. In this matter, a key question is: How can we take the advantage of compound descriptors in increasing precision while moving toward simple, single and unitary word usage? Regarding this, the modern ontological relations have a spectacular capacity to expand new ontological relations to link single descriptors or concepts for the sake of increasing precision. Ontological relations, in general, consist of the three main elements which are subject, object and property to make relation between concepts. Figure 7 shows these elements in making relation between two concepts.

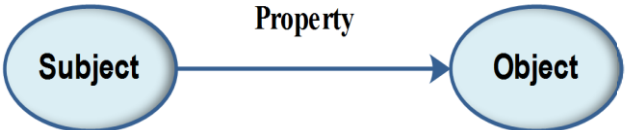


Figure 7: The Main Elements to make Relation between Concepts in Ontological relations

The mechanism of the increasing precision while moving toward single descriptors usage consists of three main steps: In the first step, compound terms split into their constituent parts to construct simple elements (i.e., single descriptors) in indexing time. For instance, "*Information System Evaluation*" splits into single-word descriptors such as "*Information*", "*Systems*" and "*Evaluation*". The second step relies on recognizing the subject and object in constituent parts of compound terms. For example, "*Systems*" plays a role as object for its subject (i.e., *Information*) and also "*System*" can be a subject for "*Evaluation*" as an object. Object, in fact, is a main factor in identifying an exact compound term while it was factored into its simple elements. Finally, a proposed novel property which we called "Co-occurrence relation" (i.e., an ontological property in making linkage between concepts) uses to make relations between the single-word descriptors in simulating virtual compound descriptors. Figure 8 demonstrates the method of virtual compound descriptors generation by co-occurrence relations.

Modern ontological relations comprise the capability to develop new ontological relations for automatic generating of virtual compound descriptors in indexing time and demonstrating them in retrieval process, while the real descriptors are simple and single ones. The construction method of the virtual compound descriptors relies on a syntactical operation to link single-word descriptors via co-occurrence relations on the basis of a simple automatic process in indexing and retrieving virtual compound descriptors. Therefore, the simulation process of the virtual compound descriptors plays an extraordinary role in achieving the advantage of compound terms in increase precision in information retrieval while single descriptors are used in the real environment of semantic relations in knowledge organizations.

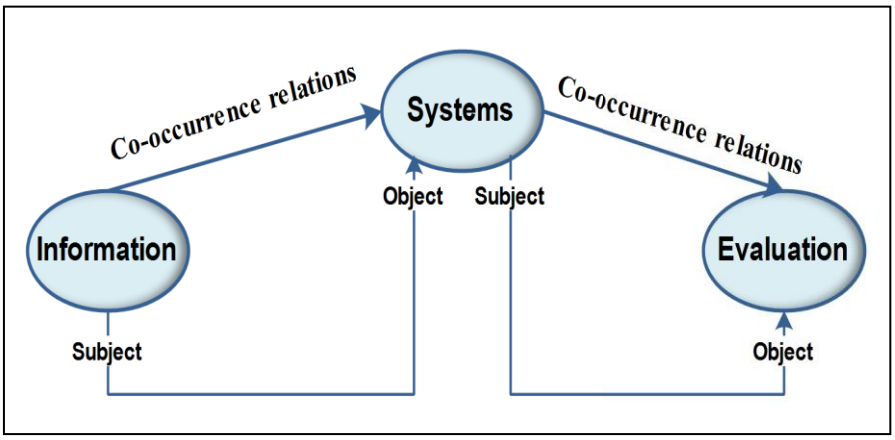


Figure 8: The Method of Generating the Virtual Compound Descriptors by Co-occurrence Relations

The generation of the virtual compound descriptors consists of two main purposes: The first purpose is the answer of this logical question: If the multi-word descriptors are split up into single words, how can retrieval system prepare the proper and precise results in searching a compound term while the exact compound term would be of interest? As stated previously, virtual compound descriptors result in finding the precise results in searching for the factored compound terms. The identification method of the old compound terms can operate through recognizing their subject and object in ontological relations. Object is a main factor in identifying an exact compound term which would be of interest in retrieval time. For instance, "*System*" is an object in determining a virtual compound term such as "*Information System*" and "*Evaluation*" is an object for "*System*" (i.e., a subject) in identifying "*Information System Evaluation*". Therefore, the last object in the string of the subject-object in compound terms is the main factor to appear virtual compound descriptor in finding the precise results in retrieval time.

The second purpose of generating virtual compound descriptors is to avoid mismatching single descriptors to increase precision in information retrieval. Mismatches which result in retrieving unrelated information are the possible conditions to link single descriptors incorrectly in information retrieval. For example, "*Information*" and "*Management*" can be linked with each other in the form of the "*Information Management*" and "*Management Information*". If user needs the related information about "*Information Management*", mismatch will happen when the retrieved information is related to "*Management Information*". In this manner, determining the subject and object of compound terms solve the mismatching problem through identifying "*Information*" as a subject and "*Management*" as a object. Consequently, the identification method of the virtual compound descriptors solves the mismatching problem through recognizing the role of subject and object in compound terms to increase precision in retrieval time.

Number of Semantic Relations

Single-word usage as descriptors results in increasing the number of semantic relations in thesauri. In this case, increasing the semantic relations can be fulfilled into three ways in thesauri. Firstly, in comparison between compound and single terms, single words consist of several semantic relations to link with descriptors in various subject fields. As an example, "*Information*" includes these semantic relations: "*Communication Information*", "*Cultural Information*", "*Educational Information*", "*Scientific Information*", "*Social Science Information*" and the like. Secondly, splitting compound terms into simpler constituent is the method to increase semantic relations. For instance, the structure of "*Information system*" is simpler than "*Information System Evaluation*". In this manner, "*Information system*" has a capacity to have more semantic relations in its specific area, such as "*Bibliographic Services*", "*Data Centres*", "*Archives*", "*Information Processing*" and so on. Finally, the composition method of single-word descriptors in retrieval time is the main factor to increase and develop semantic relations in various subject fields. For example, "*Information*", "*Management*" and "*Resources*" can be connected with each other in several ways to develop semantic relations in different subject fields, such as "*Information Resources*", "*Information Management*", "*Management Information*", "*Resources Management*", etc.

Number of Descriptors

Simplicity and complexity in the structure of descriptors play a tremendous role to increase and decrease the number of descriptors in vocabulary. For example, the compound terms such as "*Integrated Information Systems*", "*Educational Information Systems*", "*Information Systems Evaluation*", "*Geographical Information Systems*" and the like are the composition of the similar single words such as "*Information*" and "*Systems*". If these kinds of compound descriptors construct through the syntactic linkage between similar single words, the similar single-words usage increases in the structure of compound terms while the number of descriptors increases in vocabulary due to the capacity for constructing other compound terms. For instance, "*Integrated Systems*", "*Information Integrated Systems*", "*Integrated Systems Information*" and so on. In brief, while the number of compound terms increases, the number of descriptors will increase as well. On the other hand, an increase in the number of single descriptors reasons for decreasing the number of descriptors in vocabulary.

Conceptual Model of Single Descriptors Efficiency in Linking with one another.

The significant roles of simple and single descriptors in the information storage and retrieval can be clearly explained in Figure 9 which displays a conceptual model of the simple and single descriptors' efficiency in linking with one another.

Storage and retrieval through five possible conditions: First, fewer descriptors have the capacity to develop several complex concepts in the form of compound terms in retrieval time. For example, "*Information*", "*Retrieval*", "*Systems*", and "*Management*" can be combined as follows: "*Information Management Systems*", "*Management Information*", "*Information Retrieval Systems*", "*Retrieval Management*" and so on. The second condition is the potency of single descriptors in network connections via 'send' and 'receive' of several links, for instance, "*Management*" 'received' six links from the other descriptors and 'sent' two links to the other ones. The third condition is the generic essence of single descriptors which are not specified for specific subject fields (e.g., *Information Management Systems*) resulting in developing the search scale (i.e., recall rate).

In the fourth condition, the linkage possibility of single-word descriptors causes the increase in the semantic relations in various subject fields in retrieval time. For example, "*Information*", "*Management*" and "*Evaluation*" consist of the capacity to link with one another to create compound terms which belong to different subject fields in terms of increasing semantic relations, such as "*Information Management*", "*Management Information*", "*Evaluation Information*" and the like. Finally, the co-occurrence relations between single descriptors (see Figure 8) that can generate virtual compound concepts increase the precision rate. Subsequently, the movement from complexity to simplicity in descriptors structure follows the vital purpose to decrease the number of words in the structure of descriptors in terms of syntactic relations in improving information retrieval performances.

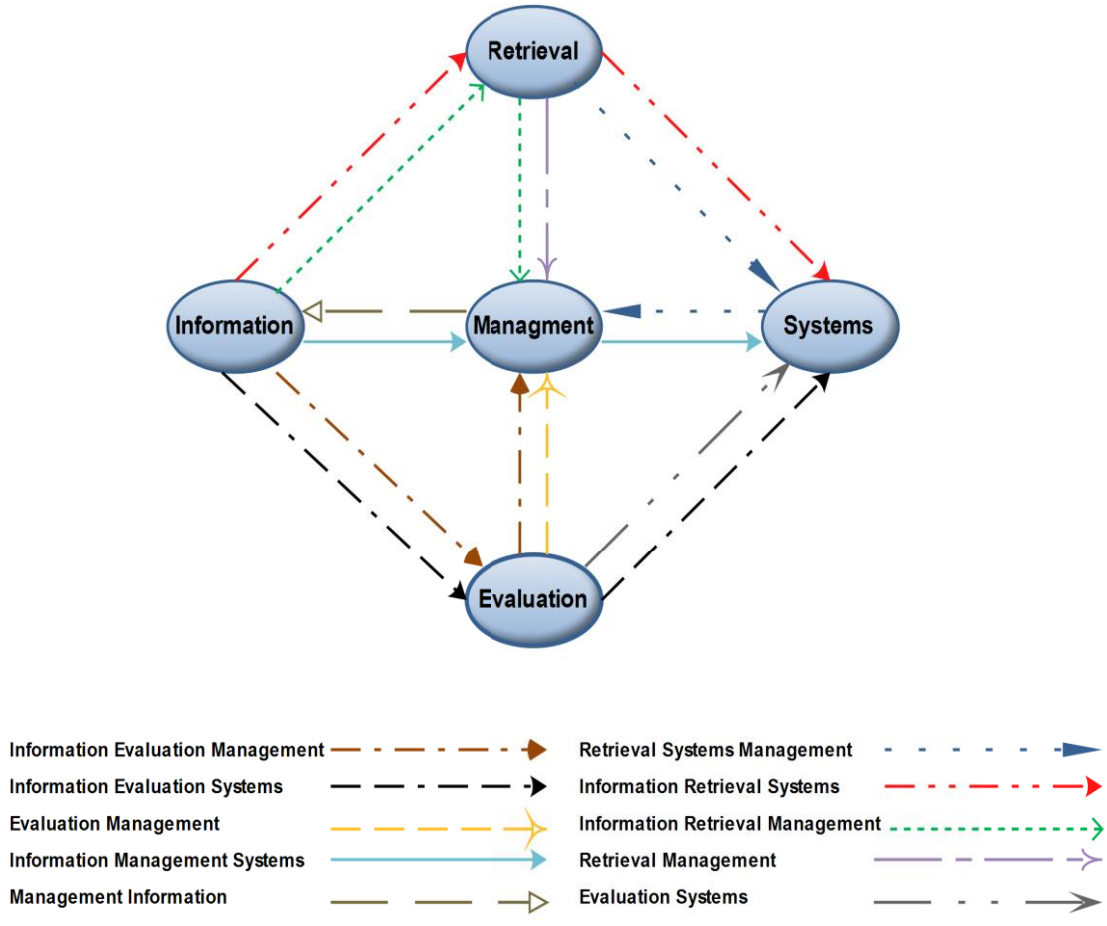


Figure 9: A Conceptual Model of the Single Descriptors Efficiency in Linking with One Another

Position of Proportional Analyses in the field of ontology evaluation

In 1983, a model to describe the relation between terms and concepts was proposed to identify the domains of terms and concepts (De Saussure 1983). Terms and concepts have specific domains which are connected and closely related. In this theory, term is identified as a

signifier and a concept as a signified. As mentioned before, simplicity and factoring ratios focus on the measurement of the simplicity in the structure of a descriptor or a term as a signifier. Therefore, when the measures were proposed to evaluate term structure and relations, the concepts and their relations cannot be ignored. Hence, our method in quantitative evaluation of the terms or the descriptors' structure can be used in the measurement of the structure of a concept as a signified in the field of ontology evaluation.

Our method on evaluating the structure of thesauri descriptors can be related to a major evaluation approach in ontology evaluation. This point of view on ontology evaluation which is done by humans who try to assess how well the ontology meets a set of predefined criteria, standards, requirements, etc. was proposed in 2004 (Brank, Grobelnik, and Mladenić 2005). Additionally, our methodology is similar to two influential ontology evaluation methodologies (i.e., OntoClean and OntoMetric) (Yu, Thom, and Tam 2009). Our methodology, like OntoClean, focuses on meta-properties or philosophical notions (Guarino and Welty 2006). Simplicity as a meta-property plays a great position in evaluating the structure of descriptors in our methodology. Although our methodology in identifying the ratio's indicators is similar to the method used in OntoMetric (Lozano-Tello and Gómez-Pérez 2004), our method in data gathering and analyzing it involves statistical and proportional analysis.

CONCLUSION

Simplicity Ratio results showed that the domain of compound descriptors is wider than simple ones in the thesauri. The movement from complexity toward simplicity in the structure of compound descriptors was confirmed by the factoring ratio results. Despite the Factoring Ratio results, these had not been effective in the simplicity results in the whole of the thesauri vocabularies. The comparison between ratio results demonstrated that the wide range of the two-word descriptors' usage is an effective factor on the simplicity decrease in the structure of thesauri descriptors. Therefore, thesaurus builders should decrease the amount of these descriptors in the factoring process to fulfil simplicity purposes to increase recall, precision and also the number of semantic relations, while there is a decrease in the number of descriptors in vocabulary to improve IR.

The inquiry results demonstrated that there are two kinds of relations in terms of the structure of descriptors. First, the relation between the number of words in the structure of descriptors and the number of compound terms with simplicity consists of an indirect relation, and second, the relation between the number of words in the structure of descriptors and the number of compound descriptors with complexity shows a direct relation. Besides two-word descriptors which are the effective factors on simplicity, thesaurus builders should move towards the use of fewer words in the descriptors' construction to develop the movement from complexity toward simplicity in the descriptors' structure. Consequently, simplicity increasing depends on decreasing the number of compound descriptors, especially two-word descriptors and the number of words in the structure of descriptors.

In this research, we presented the quantitative evaluation of the simplicity concept in the structure of descriptors which is a new method in ontology evaluation. As stated previously, this method can be developed in the field of ontology evaluation by using detailed statistics, new ratios and well-defined metrics through predefined criteria, standards, and requirements.

We are continuing our research on the evaluations of simplicity and unity notions in the structure of concepts and their semantic relations (Amirhosseini 2011) in an agricultural ontology (i.e., VocBench).

In addition to research on the evaluations of simplicity in the structure of concepts and semantic relations, this paper could be used as a starting point for a comparative in-depth-study on the statistics of knowledge organizations in different areas and languages. In this manner, consideration of the descriptors characteristics which may affect the expression of the concepts in the subject, can take into account in the splitting process for the sake of simplicity in future investigations. In this case, the novel splitting methods regarding thesauri construction standards should propose to consider descriptors characteristics in increasing simplicity in knowledge organizations. In this matter, novel ratios and well-define metrics should be proposed to develop simplicity evaluation in the form of quantitative evaluation. Furthermore, semantic complexity of descriptors should not be ignored in the form of quantitative and qualitative evaluations.

ACKNOWLEDGMENT

Special thanks go to Professor Dr. Frederick Wilfrid Lancaster for his beneficent and helpful comments in the simplicity concept for proposing Simplicity Ratio. He passed away in August 25, 2013. The authors express their appreciation to Mrs. Peggy Stubbs who edited this manuscript for language. We would like to thank the Centre for Graduate Management, Universiti Kebangsaan Malaysia (National University of Malaysia) for providing the research grant to conduct this research.

REFERENCES

Airio, E. 2006. Word normalization and decompounding in mono- and bilingual IR. *Information Retrieval*, Vol. 9, no. 3: 249–271.

Aitchison, J., Gilchrist, A., and Bawden, D. 2000. Thesaurus construction and use: A practical manual. 4th edition. London: Aslib.

Amirhosseini, M. 2007. Qualitative and quantitative evaluation of effective factors in information storage and retrieval in Persian thesaurus. Ph.D Diss., Shiraz University.

Amirhosseini, M. 2010.Theoretical base of quantitative evaluation of unity in a thesaurus term network based on Kant's epistemology. *Knowledge Organization*, Vol. 37, no. 3: 185-203.

Amirhosseini, M., and Salim, J. 2010. Quantitative evaluation of simplicity invisible domain in Islamic knowledge organizations. *2010 International Conference on Information Retrieval and Knowledge Management*: CAMP 10, exploring the invisible word, 17-18 March, 2010, Shah Alam, Malaysia. Malaysia: IEEE.

Amirhosseini, M., and Salim, J. 2011. OntoAbsolute as an ontology evaluation methodology in analysis of the structural domains in upper, middle and lower level ontologies. *STAIR'11: International Conference on Semantic Technology and Information Retrieval,* 2011, Putrajaya, Kuala Lumpur, Malaysia. Malaysia: IEEE.

Brank, J., Grobelnik, M., and Mladenić, D. A. 2005. Survey of ontology evaluation technique. *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD).* Available at: http://kt.ijs.si/dunja/sikdd2005/Papers/BrankEvaluationSiKDD2005.pdf .

Braschler, M., and Ripplinger, B. 2004. How effective is stemming and decompounding for german text retrieval? *Journal of Information Retrieval*, Vol.7, no. 34: 291-316.

Burton-Jones, A., Storey, V. C., Sugumaran, V., and Ahluwalia, P. 2003. Assessing the effectiveness of the DAML ontologies for the semantic. *Proceedings of the 8th International Conference on Applications of Natural Language to Information Systems,* Burg (Spreewald), Germany, 2003. 56-69. Available at: http://mis.commerce.ubc.ca/members/burton-jones/PDFs/ABJ-VS-VS-PA-NLDB-03.pdf

CAB International (CABI). 1995. *Commonwealth Agricultural Bureaux (CAB) Thesaurus: 1995 Edition*. Wallingford: CAB International. Available at: http://www.cabi.org/cabthesaurus/

De Saussure, F. 1983. *Course in general linguistics*. Translated by Roy Harris. London: Duckworth.

De Vries, A.P. 2001. A poor man's approach to CLEF. In: Cross language information retrieval and evaluation, *Lecture Notes in Computer Science* Vol. 2069: 149-155.

Gell-Mann, M. 2002. Plectics: The study of simplicity and complexity. *Europhysics News*, Vol. 33, no 1. Available at: http://www.europhysicsnews.com/full/13/article5/article5.html.

Guarino, N. and Welty, C. A. 2004. An overview of Ontoclean. In Staab, S. and Studer, R. (eds.), Handbook on ontology, Heidelberg and Berlin: Springer-Verlag, 151-171

His, I. 2005. *Analyzing the conceptual integrity of computing applications through ontological excavation and analysis*. Ph.D. Thesis, Georgia Institute of Technology. 217p. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.85.3889&rep=rep1&type=pdf

Hollink, V., Kamps, J., Monz, C. and de Rijke, M. 2004. Monolingual document retrieval for European languages. *Information Retrieval,* Vol. 7, no. 12: 33-52.

Holm, S. 2007. *Guidelines for constructing a museum object name thesaurus.* Available at: http://www.mda.org.uk/spectrum-terminology/holm.htm.

International Organization for Standardization (ISO). 1986. *ISO 2788: Guidelines for the establishment and development of monolingual thesauri.* Second edition. Geneva: International Organization for Standardization (ISO).

International Organization for Standardization (ISO). 2011*. ISO/FDIS 25964-1: Information and documentation thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval.* Geneva: International Organization for Standardization (ISO); Final Draft circulated April 2011.

Kang, D., Xu, B., Lu, J. and Chu, W. 2004. A complexity measure for ontology based on UML. *Proceedings of the 10th IEEE International Workshop on Future Trends of Distributed Computing Systems* , FTDCS'04, Suzhou, China: 222-228.

Kochen, M. and Tagliacozzo, R. 1968. A study of cross-referencing. *Journal of Documentation*, Vol. 24: 173-191.

Krippendorff, K. 2004. *Content analysis: an introduction to its methodology.* California: Sage Pub.

Lancaster, F. W. 1986. *Vocabulary control for information retrieval.* Virginia: Information Resource Press.

Lazarinis, F., Vilares, J., Tait, J. and Efthimiadis, E.N. 2009. Current research issues and trends in non-English web searching. *Information Retrieval*, Vol. 12, no. 3: 230-250.

Leveling, J., Magdy, W. and Jones, G.J.F. 2011. An investigation of decompounding for cross-language patent search. *Proceedings of the 34th International ACM SIGIR Conference on Research and development in Information,* July 24-28, Beijing, China: 1169-1170.

Lozano-Tello, A. and Gómez-Pérez, A. 2004. ONTOMETRIC: A method to choose the appropriate ontology. *Journal of Database Management*, Vol. 15, no. 2: 1-18.

Mungall, C. 2005. Increased complexity in the GO. Available at: http://www.fruitfly.org/~cjm/obol/doc/go- complexity.html

Muñoz, A. 1997. Compound key word generation from document databases using a hierarchical clustering art model. Working paper (Universidad Carlos III de Madrid. Departamento de Estadística y Econometría), Vol. 96, no. 76.

National Information Standards Organization (NISO). 2005.*Guidelines for the construction, format, and management of monolingual controlled vocabularies: ANSI/NISO Z39.19-2005*. Bethesda Md.: NISO Press.

National Library of Iran.2001. *Persian cultural thesaurus (ASFA).* Tehran: National Library of Iran.

Pohlmann, R. and Kraaij, W. 1997. The effect of syntactic phrase indexing on retrieval performance for Dutch texts. *Proceedings of RIAO 97*: 176–187.

Scientific Documents and Information Center (SDIC). 2004. *Chemical thesaurus*. Tehran: Scientific Documents and Information Centre (SDIC).

TESE - Thesaurus for Education Systems in Europe. 2006. Available at: http://www.eurydice.org/ressources/Eurydice/TESE/pdf/TESEEN_002_intro.pdf.

UNESCO Thesaurus Online. 2004. Available at: http://databases.unesco.org/thesaurus/

Word. 2012. Oxford Dictionaries Online. Available at: http://oxforddictionaries.com/ definition/word?q=word

Yang, Z., Zhang, D. and Ye, C. 2006. Ontology analysis on complexity and evolution based on conceptual model. U. Leser, F. Naumann, and B. Eckman (eds.): DILS 2006, LNBI 4075: 216-223.

Yu, J., Thom, J.A. and Tam, A. 2009. Requirements-oriented methodology for evaluating ontologies. *Information Systems*, Vol. 34: 766–791.

Zhang, D., Ye, C. and Yang, Z. 2006. An evaluation method for ontology complexity analysis in ontology evolution. In S. Staab and V. Svatek (eds.). EKAW 2006, LNAI 4248, *International Conference on Knowledge Engineering and Knowledge Management,* No. 15, Podebrady, Tcheque, Republique. 4248: 214-221.