The evaluation of Persian Thesauri of Health and Biomedical Sciences Based on ANSI/NISO z39.19 2005 Standard

Nadjla Hariri[1], Maryam Kazerani[1,*], Reza Moghaddam Kiya[2]

[1]Department of Library and Information Science, Science and research Branch, Islamic Azad University, Tehran, IRAN
[2]Faculty of Foreign Languages and Literatures, University of Tehran

*Corresponding Author: email address: kazerani.m@gmail.com (M. Kazerani)

ABSTRACT

Establishing order, integrity, consistency, and accuracy in thesauri as information storage and retrieval systems plays an important role in the organization of information. The deployment of universally accepted standards, therefore, facilitates the accomplishment of this goal. To achieve this goal, the ANSI /NISO Z 39.19 -2005 standard offers certain rules and principles for constructing controlled vocabularies, including the thesauri, with regard to four dimensions of form construction, semantic relationships, displaying controlled vocabularies, and management systems. These dimensions served as a basis in this study to evaluate the accuracy of vocabulary construction in three Persian Thesauri of Health and Biomedical Sciences, namely, the thesaurus of biological sciences, the thesaurus of health promotion and the thesaurus of medicine. The research method utilized in this study was the analytical survey method. A sampling method was also used to establish samples of the terms contained in the latest edition of every thesaurus. The research instrument used in this study was a checklist defined in Microsoft Excel. The samples for the four dimensions were entered in the checklist and were evaluated on the basis of the relevant standards. The results of the study revealed that the thesauri under study had not given due attention to the standards related to the management dimension. The level of observance of standards for this dimension was only 14% in all of the thesauri. Regarding the dimension of displaying controlled vocabularies, the highest level of observance of the standards was 73% which belonged to the thesaurus of medicine, the lowest level was 54.5% which belonged to the thesaurus of biological sciences, and for the thesaurus of health promotion it was 59%. As for the dimensions of the semantic relations and form construction, the results of the study indicated a higher level of observance of the standards in the thesauri under study. Finally, it was concluded that possessing the required expertise, familiarity with the standards of thesauri construction, utilizing appropriate models, taking the needs of the end users into consideration are among the major factors that if observed could lead to the construction of more effective and efficient thesauri.

Keywords: Persian Thesauri; Evaluation; ANSI /NISO Z 39.19 - 2005 standard; Iran; Health and Biomedical Sciences; semantic relations; form construction; displaying controlled vocabularies; management systems

INTRODUCTION

Information systems have been developed in line with the development of knowledge and the media through which it is disseminated. This expansion which was needed to facilitate researchers' quick access to knowledge resources resulted in the emergence of new challenges against setting the standards, methods and approaches to the organization of information. The emergence of computer and modern information technology, which, in turn, led to the development of new ways for the communication of information, entails reassessment of the standards, tools and organizing techniques.

There are two approaches to designing adequate systems for the organization of information to respond to these new developments. First, adjusting the traditional information organization tools according to the new electronic environment; and second, developing new tools and organizing techniques to make the maximum benefit out of the capabilities and potentialities of the new environment.

Applying thesauri in the new electronic information environments is an example of the first approach currently attracting attention because of the efficiency of these tools in supplying the needs of the organizers and users of information systems [1]. On the other hand, constructing thesauri is regarded as a professional activity that requires the collaboration of well-experienced staff and familiarity with the standards of the creation of information resources. In Iran, using information in the area of health sciences has suffered from a number of shortcomings and converting documented records into the language of the information communication system has experienced very serious problems [2]. The researchers and practitioners in the field, however, have tried to tackle this problem by creating thesauri of health sciences. Three of such thesauri have been developed between 2005 and 2006. However, inadequate attention to principles and standards in the construction of thesauri could result in difficulties in information retrieval [3].

Therefore, carrying out a research to evaluate the quality of the thesauri in the field of human knowledge deems to be necessary. To respond to this need, the present study aimed to examine the Persian thesauri of Health and Biomedical Sciences based on ANSI /NISO z39.19 2005 standard [4] considering four key areas of semantic relations, form construction, displaying controlled vocabularies, and management systems.

MATERIALS AND METHODS

The research method utilized in this study is the analytical survey method. Descriptive statistics is also used in the analysis of the information. The research population of the study consists of the latest edition of the following three Persian thesauri of Health and Biomedical Sciences created and published in Iran.

1 - Thesaurus of Biological sciences [5]
2 - Thesaurus of Medicine [6]
3 - Thesaurus of Health promotion [7]

Due to the great number of entries in every thesaurus, a sampling method was adopted. Therefore, Morgan's Table was used to determine the number of samples for every thesaurus. If the total number of terms in one of the thesauri was not specified by the authors in the preface, the number was estimated via multiplying the approximate average number of terms in each page by the number of pages indexed alphabetically. It is worth mentioning here that according to Morgan's Table, the number of samples was 344 terms for the Thesaurus of Biological Sciences, 368 terms for the Thesaurus of Medicine, and 294 terms for the Thesaurus of health promotion.

The next stage was to use systematic sampling method to obtain sampling fractions via dividing the number of pages of the thesaurus by the number of samples. For example, if a thesaurus contained 3000 alphabetically indexed pages and if it was required to study 300 terms on 300 pages of it, then one page had to be selected from every 10 pages. And this was done with the help of the simple random number table. What had to be done next was to choose one term among the terms in each one of the selected pages which again was easily done through using a simple random number table. Then the selected terms were entered into a checklist in Excel 2007.

If a term had followed the standards with regard to semantic relations and form construction, it gained the value of one; and if it had not, it obtained a value of zero. In cases where a term lacked the intended feature a value of two was given to it. Finally, the frequency distributions were presented in tables.

Regarding the dimension of displaying controlled vocabularies, if a selected entry had followed the required standard, the value of one; if not, the value of zero; and if it was wrong, the value of 0.5 was given to it. And finally, with regard to management system for every thesaurus, observance and non-observance of standards were once again measured through a zero and one evaluating system. The resulting values were rounded off up to one decimal digit and were again illustrated via tables.

RESULTS

In this section, in three separate tables, the levels of observance of standards for three dimensions of semantic relationships, form construction, displaying controlled vocabularies, and management system for each of the three

Thesauri of Health and Biomedical sciences are presented.

A. Level of observance of standards of semantic relationships in the thesauri of health and biomedical sciences: The level of observance of standards for hierarchical relationships is 96.5%, for equivalent relationships is 100% and for associative relationships is 93%. Therefore, the average level of observance of standards for semantic relationships in the Thesaurus of Biological Sciences is 96.5%.

The level of observance of standards for hierarchical relationships is 80%, for equivalent relationships, is 100% and for associative relationships is 91%. Therefore, the average level of observance of standards for semantic relationships in the Thesaurus of Health improvement is almost 90%.

The level of observance of standards for hierarchical relationships is 95%, for equivalent relationships is 100% and for associative relationships is 100%. Therefore, the average level of observance of standards for semantic relationships in the Thesaurus of Medicine is almost 98%.

Table 1. Frequency distribution of correctness and incorrectness in hierarchical, equivalent, and associative relationships in the Thesaurus of Biological Sciences

Outcome	Type of Relationship	Hierarchical		Equivalent		Associative	
		number of terms	percentage	number of terms	percentage	number of terms	percentage
Observed	correct	326	96.5	120	100	106	93
	incorrect	12	3.5	0	0	3	7
Total		338	98	120	35	109	32
Not observed		6	2	224	65	235	68
Sum total		344	100	344	100	344	100

Table 2. Frequency distribution of correctness and incorrectness in hierarchical, equivalent, and associative relationships in the Thesaurus of Health promotion

Outcome	Type of Relationship	Hierarchical		Equivalent		Associative	
		number of terms	percentage	number of terms	percentage	number of terms	percentage
Observed	correct	185	80	149	100	202	91
	incorrect	46	20	0	0	21	9
Total		231	78.5	149	51	223	76
Not observed		63	21.5	145	49	71	24
Sum total		294	100	294	100	294	100

Table 3. Frequency distribution of correctness and incorrectness in hierarchical, equivalent, and associative relationships in the Thesaurus of Medicine

Outcome	Type of Relationship	Hierarchical		Equivalent		Associative	
		number of entries	percentage	number of entries	percentage	number of entries	percentage
Observed	correct	348	95	194	100	75	100
	incorrect	18	5	0	0	0	0
Total		366	99.5	194	53	75	20
Not observed		2	0.5	174	47	293	80
Sum total		368	100	368	100	368	100

B. Level of observance of standards of form construction in health and biomedical sciences thesauri

The minimum level of observance of standards is 60% for single word versus multiword feature, and the average level of observance of standards in the form construction of the Thesaurus of Biological Sciences is 91%.

The minimum level of observance of standards is 29% for homograph feature, and the average level of observance of standards in the form construction of the Thesaurus of health improvement is 83%.

The minimum level of observance of standards is 99% for single word versus multiword feature, and the average level of observance of standards in the form construction of the Thesaurus of Biological Sciences is almost 100%.

Table 4. Frequency distribution of correctness and incorrectness of form construction in the Thesaurus of Biological Sciences

Features		Homograph		Scope notes		Single word vs. multiword terms		Nouns & noun phrases		adjectives		Initial articles		Compound terms		Spelling	
Outcome		terms	percentage	terms	percentage	terms	percentage	terms	percentage	terms	percentage	terms	percentage	terms	percentage	terms	percentage
Observed	correct	12	92	4	100	207	60	344	100	83	99	3	75	105	100	344	100
	incorrect	1	8	0	0	137	40	0	0	1	1	1	25	0	0	0	0
Total		13	4	4	1	344	100	344	100	84	24	4	1	105	30.5	344	100
Not observed		331	96	340	99	----	----	----	----	260	76	340	99	239	69.5	----	----
Sum total		344	100	344	100	344	100	344	100	344	100	344	100	344	100	344	100

Table 5. Frequency distribution of correctness and incorrectness of features related to form construction in the Thesaurus of Health promotion

Features		Homograph		Scope notes		Single word vs. multiword terms		Nouns & noun phrases		adjectives		Initial articles		Compound terms		Spelling	
Outcome		terms	percentage	terms	percentage	terms	percentage	terms	percentage	terms	percentage	terms	percentage	terms	percentage	terms	percentage
Observed	correct	2	29	133	100	197	67	287	98	49	87.5	----	----	101	100	294	100
	incorrect	5	71	0	0	97	33	7	2	7	12.5	----	----	0	0	0	0
Total		7	2	133	45	294	100	294	100	56	19	----	----	101	34	294	100
Not observed		287	98	161	55	----	----	----	----	238	81	294	100	193	66	----	----
Sum total		294	100	294	100	294	100	294	100	294	100	294	100	294	100	294	100

Table 6. Frequency distribution of correctness and incorrectness of form construction in the thesaurus of Medicine

Features		Homograph		Scope notes		Single word vs. multiword terms		Nouns & noun phrases		adjectives		Initial articles		Compound terms		Spelling	
Outcome		terms	percentage	terms	percentage	terms	percentage	terms	percentage	Terms`	percentage	terms	percentage	terms	percentage	terms	percentage
Observed	correct	2	100	----	----	364	99	368	100	74	100	2	100	232	100	368	100
	incorrect	0	0	----	----	4	1	0	0	0	0	0	0	0	0	0	0
Total		2	0.5	----	----	368	100	368	100	74	20	2	0.5	232	63	368	100
Not observed		366	99.5	368	100	----	-----	----	----	294	80	366	99.5	136	37	----	----
Sum total		368	100	368	100	368	100	368	100	368	100	368	100	368	100	368	100

C. Level of observance of standards of displaying controlled vocabularies in health and biomedical sciences thesauri

As table seven indicates, the scores for observing features related to the displaying controlled vocabularies in the three Thesauri of Health and Biomedical Sciences are 12, 13, and 16 out of 22, and the percentage of the level of observance of these standards in the three thesauri are 54.5%, 59%, and 73% respectively.

D. Level of observance of standards of management system in Health and Biomedical Sciences Thesauri

As table eight indicates, all three thesauri of Health and Biomedical Sciences acquired the same score of one out of seven for following the features related to the management system. This score accounts for 14% of observing the standards of the management system in these thesauri.

DISCUSSION

Production, storage and the use of information constitute three important processes in information management systems, which are positively or negatively affected by any improvement or any disruption in these processes. In our country the use of information, involves many difficulties and shortcomings. These problems partly relate to the production of information and partly to the way it is stored. Those who need information visit libraries and other data bases to obtain the information they need. However, the information they receive is far different from what they should actually be able to obtain. Fortunately, the creation of three thesauri in the field of Health and Biomedical Sciences in a short period of time has been an attempt toward solving these problems. Whether or not these thesauri follow the standards, however, is a quite different matter which needs thorough investigation.

In a research carried out by Kazerani (1999) on evaluating the Persian language macro-thesauri on the basis of ISO 2788, the most serious shortcomings of thesauri in Iran were shown to be the non-observance of the standards in semantic relations and displaying controlled vocabularies [8]. Another qualitative research was carried out by Hosseini Zadeh (2004) in which she interviewed 6 people who were responsible for the construction of 12 thesauri constructed and published in Iran up to the year 2003. Her study aimed to identify the problems involved with the job of constructing thesauri in Iran from the producers' perspective. She reported that most of the interviewees had experienced problems which were managerial and executive in nature [9]. Also, almost all the producers referred to the lack of interaction between responsible institutions, at national and organizational levels, as the main problem of thesauri construction in Iran.

In line with the findings of previous studies, the results of the present study show that the most serious problem of the thesauri under study is their failure in observing the standards in their management system.

The study further provides evidence for the fact that, putting technical features aside, the problems of constructing standard thesauri are mostly managerial and executive in nature.

Table 7. Frequency distribution of the features related to the displaying controlled vocabularies in   the Thesauri of health and biomedical sciences

Thesauri / Features	Thesaurus of Biological sciences			Thesaurus of Health promotion			Thesaurus of Medicine		
	observed		not observed	observed		not observed	observed		not observed
	correct	incorrect		correct	incorrect		correct	incorrect	
Alphabetical displays	1			1			1		
Graphic displays			0			0			0
Permuted displays	1			1			1		
Term detail displays	1			1			1		
Hierarchical displays	1			1			1		
Faceted displays			0			0			0
Needs of thesaurus maintenance personnel			0			0	1		
Needs of indexers & expert searchers			0			0	1		0
Needs of end users			0			0	1		0
Element to address (Presentation , Type of displays , format & documentation)			0			0			0
Displaying the equivalence relationship	1			1			1		
Displaying associative relationship	1			1			1		
Displaying hierarchical relationship	1			1			1		
Usage (literary warrant, organizational warrant& user warrant)			0			0	1		
Indention	1			1			1		
Typography	1			1			1		
Filling and sorting		0.5		1			1		
Print format	1			1			1		
Screen format			0			0			0
Web format	1			1			1		
Documentation			0			0			0
Elements to address		0.5		1					0
Scores	12			13			16		
Percentages	54.5			59			73		

Table 8. Frequency distribution of factors related to the management system of the Thesauri of Basic Sciences

Features	Observed	Not observed
Avoid duplicating existing vocabularies	1	
Determine the structure & display formats		0
Testing and evaluation		0
Updating the vocabulary		0
Error checking		0
Candidate terms		0
Term deletion		0
Scores	1	
Percentage	14	

The low observance (14%) of the standards in the management system of the thesauri examined in this study could be due to financial and supplying problems experienced by the creators of these thesauri in publishing the new editions of the thesauri [10]. In fact, because of such problems, they are not sure about whether the new edition of the thesaurus will be published, and thus do not take the measures necessary for publishing the next editions of the thesaurus.

According to the ANSI/NISO z39.19 2005 standard, avoiding duplication, continuous evaluation, staying up-to-date, and using an appropriate mechanism for entering a term into or omitting it from the thesaurus are among the standards that if they are observed in the management system of thesauri, will indicate that thesauri constructors take the needs of the end users into consideration in the processes of the storage and retrieval of information. Moreover, if thesauri, as Aitchison et al (2000) claim, are living and growing organisms[11], then adopting clear policies in the management system as well as mechanisms for controlling the system, could lead to the producers, end users and information indexers' trust in the thesauri.

Displaying controlled vocabularies is another problematic area in the thesauri examined in this study. As the findings of the present study show the conformity to the standards of the presentation method in the thesauri of Basic sciences fluctuates between 54.5 to 73%. In the same way, if certain concepts like subject coverage, number of terms, aims, punctuation, abbreviations and symbols, the rule of choosing selective terms and their relations, elaboration of the rules of the page layout, date of updating and updating policy, and instructions on how to use the thesaurus are not clearly stated, observed or described in a thesaurus, it cannot effectively satisfy the needs of the users.

A thesaurus is supposed to be useful, at least, for two groups of addressees, that is, the indexers and the ordinary users. Therefore, if it lacks order, clarity and the fundamental information, it will not act as a guide for the users or as a tool for organizing information [3]. Constructing a thesaurus in which the selection of terms is not based on literary warrant, organizational warrant, and user warrant, is very much like sewing a piece of clothing for someone whose size is not known to us [12].

Since literary warrant and user warrant are two important factors in the construction of a thesaurus, it is, therefore, necessary that the specialized centers that have the required warrants in a specific subject area and that intend to create thesauri be advised about how they can do it in a correct and scientific manner [10].

As for the dimensions of semantic relations and form construction, according to the results, there were fewer problems compared to the management system and displaying controlled vocabularies; that are the average level of observing the standards fluctuated between 96.5% and 99% for semantic relations and between 83 and 99% for form construction.

This high degree of observing the standards in these two areas implies that the producers of these thesauri have been aware of the results of the previous researches.

Today, at the time of the emergence of the World Wide Web, expansion of Media, and variation in information storage and retrieval tools, not only the need for constructing thesauri is not abolished, but is felt more urgent than before, especially for the purpose of information architecture on the Web [13], and despite the emergence of more advanced tools for the storage and retrieval of information, thesauri have kept their importance even more than before and they now make an indispensible part of the biggest system of information storage and retrieval, i.e. the Web [14]. However, achieving higher standards in creating thesauri requires that the people involved in this job in Iran be recommended to:

1. Adopt standard approaches in the process of creating thesauri;

2. Deepen their knowledge, skill, and specialization on how they can create standard thesauri;

3. Adopt an appropriate model for creating thesauri and remain loyal to universal rules and principles of the field;

4. Be always attentive of the needs of the end users of the thesauri.

ACKNOWLEDGEMENT

REFERENCES

1. Taheri, S. (2007). Reflections on the national conference on thesauri and the application of thesauri in the electronic environment, retrieved from: http://thecon.islamicdoc.com/index.asp, October 2011.

2. Rahadust, F. et al. (1998).The thesaurus of Medicine. National Library and Archives of Islamic Republic of Iran.

3. Kazerani, M. et al. 2000. Evaluation of Persian Thesauri on the basis of ISO Standards: a step toward improving the storage and retrieval of information. Proceedings of the 5th conference on Library and Medical information Sciences. Ketabdar Publication, Tehran.

4. ANSI/NISO Z 39.19 (2005). Guidelines for the construction, format and development of monolingual thesauri. Bethesda, Maryland, NISO press

5. Akbari, E. et al. (2005). The Thesaurus of Biological Sciences. Iranian Research Institute for Information Science and Technology.

6. Rahadust, F. et   al. 2005. The Thesaurus of Medicine. 2nd.ed. National Library and Archives of Islamic Republic of Iran.

7. Dehsarai, Z. (2005). The Thesaurus of Health Promotion. Iranian Research Institute for Information Science and technology.

8. Kazerani, M. (1999). Evaluation of Persian Thesauri on the basis of ISO 2788 standard. M. A Thesis. Iran University of Medical Sciences.

9. Hosseinizadeh, A. (2004).Problems involved with constructing thesauri in Iran from the producers' Perspective. M.A Thesis. Iran University of Medical Sciences.

10. Mohammadi, M. (2007). Adjusting Persian Thesauri to the information storage and retrieval needs of the third millennium and a plan for integrating the existing thesauri using "term-based" and "conceptual" models. Presented at the national conference on Thesauri and their application in the electronic environment, Qom, center for Islamic information and documents

11. Aitchison et al. (2000) Thesaurus construction and use .4th ed. London ASLIB.

12. Rahadust, F.(2007). A review of the thesaurus of Physics. Ketab e mah e koliyat [monthly review of books]. No. 118: 62-65.

13. Rosenfeld, L. and Merville, P. (2002). Information architecture for the World Wide Web. 2nd ed. Sebastopol, CA: O'Reilly

14. Shiri, A. and Revie, C.(2000). Thesauri on the web: current developments and trends. Online Information Review, Vol. 24, no. 4: 273-279.