Comparative analysis of nuclear energy vocabulary

N.V. Maksimov, O.L. Golitsyna, V.M. Kupriyanov, A.S. Pryakhin, G.V. Tikhomirov

National Research Nuclear University MEPhI (NRNU MEPhI), Moscow, Russia

E-mail: NV-Maks@yandex.ru

Abstract

The results of the analysis of terminological subsets of the characteristic vocabulary used for semantic identification of scientific and technical information in the field of fast reactors and light water reactors are discussed. The procedure of automatic formation of thematic dictionaries is considered. The dictionaries were formed through the processing of full-text materials such as dissertations and manuals on operation of nuclear reactors. The characteristic terms (in the form of phrases) are singled out from full texts by the methods of the statistic-linguistic analysis. The IAEA-INIS thesaurus is used.

Key words: comparative analysis of vocabulary; concept text analysis; scientific information; operational documentation.

1. Introduction

Not only the tendency of global migration of knowledge and its processing (and a synthesis of new knowledge as a result), but the availability of sufficient technical conditions for this are typical at present. The design, engineering, scientific, technical, and operational documentation as well as on-line information are already in the digital environment. But the efficiency of its use (in particular, as a basis for the synthesis of new knowledge) is directly determined by the contents identification quality. The identification of contents includes the construction of semantic indices (search images) ensuring the required selectivity of search with an acceptable level of terms generality, which is the necessary condition for their sustainable choice by users while solving thematically similar tasks.

This process is similar and is somehow linked to an explicit/implicit categorization (classification) of the document content. At the upper level, the author provides a "categorization" of meaning, inventing the title of the text. At the next level, he forms a summary of the words, which – in his opinion – are best associated with the knowledge contained in the document. Thereafter, the author structures the text using certain rules and singling out the relevant concepts and entities. The selection of concepts and their relationships (representing knowledge in this subject area) are considered as the most difficultly formalized process for computer analysis of texts. For example, the words “nuclear education” mean "development of

person's competences in the field of nuclear power", while in nuclear medicine they mean “formation of tumors inside the living cells of the organism”.

The objective of this work is studying the degree of generality/specificity of terms used to identify the semantic content of text documents. A comparative analysis of the documents language is conducted for words and expressions associated with the operation of pressurized water reactors and fast reactors at different stages of their life cycles.

2. Approaches and tools

To construct subject area concepts and terminology systems for the analysis of thematic areas, one could use an interactive and iterative technology based on methods and means for the retrieval of scientific, technological, and educational information from various sources. The technology includes the following steps [1]:

(1) Build up an information vision at the level of documents reflecting the subject area condition most completely and objectively.

(2) Build up a terminological framework as glossaries of nuclear terms (descriptors of the thesaurus) along with set phrases expressing the basic meaning in the selected documents. The algorithm also takes into account "extended" descriptors, where words follow each other in a different order, or clarifying words that are not part of the descriptor and are located between its individual words. An empirical threshold value is used to reduce the list.

(3) Construct conceptual terminological networks based on the sequential formation of lexicographical neighborhoods for nuclear terms including nominal noun phrases. The construction of phrases is made in compliance with lexical patterns [2].

The resultant lists of descriptors, treated as conceptual images, are subjected to the fuzzy matching procedure providing an analysis of subject areas. Such lists can be considered as candidates for entering into the expanded thesaurus.

3. Experimental comparison of the vocabulary

An analysis of terms applied for semantic identification of scientific and technical texts was carried out using full-text documents on pressurized water reactors (PWR) and fast reactors (FR). The texts belong to different stages of the life cycle of nuclear knowledge – the research stage and the operational phase. Research texts were submitted by scientific dissertations (ten dissertations over each subject for the period from 2007 to 2012), and operational documentation – reports on irregularities at nuclear power plants (about one hundred reports for pressurized water reactors and ten reports for fast reactors). Frequency dictionaries of word combinations were compiled for each direction. INIS thesaurus descriptors [3] were singled out in each dictionary.

3.1.Analysis of research texts language

Table 1 gives the results of the analysis of descriptors from the INIS thesaurus used in the texts of dissertations. A conceptual image of each dissertation was made by phrases statistically significant within a dissertation – descriptors of the INIS thesaurus.

Table 1

Frequency	*% of common descriptors (OR)*		*% of commmon descriptors (AND)*	
	PWR	*FR*	*PWR*	*FR*
> 1	81.5	82.5	54.6	63.3
> 2	91.5	94.3	48.0	67.2
> 3	95.2	95.8	35.7	62.1
> 5	100	100	32.9	79.3

867 descriptors from dissertations on pressurized water reactors and 1015 descriptors from dissertations on fast reactors were entered in the composite dictionary. 519 descriptors became common (overlap of the descriptors sets) for these subjects.

Table 1 presents the distribution of descriptors depending on the frequency of mention in the texts. The column "*frequency*" reflects the number of dissertations in which the descriptor was found. The column "*% of common descriptors (OR)*" shows the percentage of common descriptors encountered at the frequency specified in one of the conceptual images and at least once in another. The column "*% of common descriptors (AND)*" shows the percentage of descriptors encountered at the frequency specified in each of the conceptual images.

As can be seen from the table, using thesaurus descriptors rather weakly depends on the subject specifics.

Similar results are presented in Table 2 for phrases that were selected with due account of their morphological and linguistic characteristics. These phrases are not represented in the thesaurus.

Table 2

Frequency	*% of common descriptors (OR)*		*% of common descriptors (AND)*	
	PWR	*FR*	*PWR*	*FR*
> 1	15.0	18.7	13.4	17.8
> 2	23.1	27.8	14.7	20.3
> 3	23.3	27.7	13.5	18.6
> 5	31.3	35.4	14.3	20.2
> 10	42.5	44.2	15.0	21.6
> 15	49.8	46.4	15.1	20.1
> 20	56.6	50.0	15.4	21.0
> 30	60.5	55.8	14.0	23.1
> 40	61.5	56.0	11.5	17.6
> 200	100	100	0	0

It is obvious that the percentage of lexical items overlap for uncontrolled vocabulary is much lower vs. controlled one.

3.2. Lexical analysis of reactors operational documentation

The analysis of texts showed that more than 90% of the thesaurus descriptors, singled out in reports on fast reactors, were entered in the set of descriptors for pressurized water reactors. On

the one hand, this can be explained by the great domination of manuals on pressurized water reactors. On the other hand, the tendency of the thesaurus vocabulary unification is observed.

It should be noted that quite a small number of descriptors were identified for each subject: 237 for pressurized water reactors and 24 for fast reactors.

An analysis of uncontrolled vocabulary (selected phrases) also shows a high level of commonality – even the glossary with a frequency greater than 1 showed that about 50% of phrases built from the texts for fast reactors, had come into the terminological overlap, and the number of phrases that occurred more than 3 times exceeded 90%.

3.3. Revealing of common vocabulary at different stages of life cycle

Table 3 presents the share of common vocabulary (terms with frequency greater than 1) relative to the analyzed dictionaries.

Table 3

Theme	*% of common descriptors*		*% of common uncontrolled phrases*	
	dissertations	*reports*	*dissertations*	*reports*
PWR	12.5	33.9	1.7	5.1
FR	3.0	35.3	0.2	12.8

For obvious reasons, the level of commonality for controlled vocabulary is significantly higher than for uncontrolled one. A higher indicator for the scientific subject of PWR due to the significantly more elaborate than for FR. The obtained inverse relation to manuals due to the relatively small size of the analyzed sample, and the uniqueness of the source of information. This led to some similarity of the texts.

3.4.Qualitative analysis of non-controlled vocabulary

Statistically significant phrases, not included in the conceptual terminological structure (the INIS thesaurus and the alphabetical index of fuel cycle terms [4] were considered as conceptual structures) were analyzed for some basic concepts in the nuclear power industry (category level). Table 4 gives some examples of phrases that are common for the subjects discussed and are specific for each of them.

The table gives Russian terms without word-for-word translation because it would not be authentic. For example, the Russian term "водо-водяной энергетический реактор" (VVER) corresponds to "pressurized water reactor" (PWR), and the term "снятие установки с эксплуатации", to "decommissioning".

An analysis of specific vocabulary, which is not reflected in the thesaurus, shows its capacity for FR significantly (almost 10 times on an average) more than for PWR. Apparently, this is due to the predominance of the latter in the nuclear power industry and, consequently, a higher level of vocabulary standardization.

Table 4

Common vocabulary	*FR*	*PWR*
НУКЛИД		
Стартовый нуклид Сырьевой нуклид Топливный нуклид Тяжелый нуклид	Промежуточный нуклид Трансурановый нуклид	Легкий нуклид Нуклид урана Основной нуклид Резонансный нуклид Фиктивный нуклид
ТОПЛИВО		
Благородные газообразные продукты деления ядерного топлива Выгорание ядерного топлива Изготовление топлива Перегрузка топлива Продукты деления ядерного топлива Распухание топлива	Денатурированное топливо Карбидное топливо Облучение ядерного топлива Обогащение топлива Оксидное топливо Очистка топлива Перемещение топлива Переработка металлического топлива Переработка плотного топлива Перестройка топлива Плотное топливо Плутониевое топливо Радиохимическая переработка ядерного топлива Растворение топлива Растрескивание топлива Расщепление ядерного топлива Рециклирование ядерного топлива Солевые технологиии регенерации топлива Топливная композиция Ториевое топливо Трансурановые нуклидные топливные композиции Фторирование ядерного топлива Циркониевые сплавы топлива Ядерное оксидное топливо	Непрерывная перегрузка ядерного топлива Неравновесная перекристаллизация оксидного топлива Поставка ядерного топлива Циклическая перегрузка ядерного топлива
ТОПЛИВНЫЙ ЦИКЛ		
	Равновесный топливный цикл Ториевый топливный цикл	Конверсионный топливный цикл
ТРАНСМУТАЦИЯ		
Трансмутация долгоживущих продуктов деления	Трансмутация актинидов Трансмутация изотопов Трансмутация малых актиноидов Трансмутация минорных актиноидов Трансмутация младших актинидов Трансмутация нуклидов Трансмутация плутония Трансмутация тяжёлых ядер Трансмутация ядер Цепочка трансмутации ксенона Цепочки трансмутации	

As concerns steady occurrence in the texts of statistically significant terms that are not present in thesauruses, it can be explained as follows:

- inconsistency (in the form of presentation) of the term use, and its fixation in the conceptual structure;
- the fact that the scientific community has a strong tradition of special usage associated with the academic schools (scientific councils, educational programs etc). Hence the absence of the uniform terminological base in the publishing practice and a high level of jargon. In the past years, it was science editors that ensured the unification of terminology. Currently, such institution is non-existent. Responsible reviewing, which also contributed to the unification of scientific terminology, is absent as well.

4. Conclusions

The results obtained show the necessity of taking into account statistical features of the vocabulary used depending on a subject area and on a document type because representation of information is largely determined by the corresponding stage of the object life cycle[1]. This would allow a more adequate compilation of vocabularies of terms used for semantic identification documents.

In general, this would enable as follows: (1) computer-aided formation of properly arranged repositories for textual materials at nuclear enterprises; (2) preservation of factual knowledge accumulated at the previous stages of nuclear power development in digital warehouses providing young professionals' access to the knowledge created by the previous generations of scientists. An example of this approach is the Fast Reactor Knowledge Organization System, established by the IAEA experts [5]; (3) experts' access to scientific and technical texts as it will allow efficient high-accuracy thematic search of materials in digital warehouses.

An analysis of the specificity of terms, characterized by the dynamics of their common use, could be also of interest. For example, the indicative role of the TRANSMUTATION term for the closed fuel cycle is shown based on the use frequency analysis in a stream of publications on various aspects [1].

5. Recommendations

The recent analysis of terminology (even on such limited arrays) allows us to formulate the following recommendations for further work on automated formation and maintenance of linguistic support.

(1) Special work is required to form thematic glossaries, where terms should become high level concepts in subject areas. This primarily refers to the fast reactor technology. The composition of these glossaries should be minimized by experts familiar with the

[1] Therefore, the INIS IAEA Thesaurus working group in forming the terms of an INIS multilingual thesaurus practically does not use the formal analysis of full texts. The bulk of terms were determined by experts through consultations with professionals – native speakers.

conceptual space of the subject area. Professionals from the IAEA Nuclear Knowledge Management Section followed this way when developing the Fast Reactor Knowledge Organization System.

(2) SC Rosatom's Directorate of Scientific Programmes could ask scientific experts at leading specialized enterprises to compile thematic glossaries. These glossaries should become an integral part of regulatory documents making obligatory the use of standardized terms during documents execution. Glossaries terminology should comply with the relevant INIS IAEA glossaries.

(3) General coordination of work associated with the formation of a linguistic resource for nuclear terminology could be provided by professionals not involved in specific subjects. These experts should use knowledge management tools in ongoing activity, first of all, in education, since terminology can be put in practice only through the real educational process. This could be most effectively performed through the National Research Nuclear University MEPhI as a university maintaining close contacts with the IAEA INIS, in particularly, in the framework of the Russian National INIS Center.

REFERENCES

[1] GOLITSYNA, O.L., MAKSIMOV, N.V., STROGONOV, V.I., TIKHOMIROV, G.V. 2011, Sistemy Upr. Inf. Tekhnol. 44 (1.1), pp. 126-134

[2] GOLITSYNA, O.L., MAKSIMOV, N.V., Comparative structural and statistical analysis of vocabulary and communication of information search thesauruses, Nauchn.-Tekhn. Inform., Ser. 2, 2015, No. 6, pp. 14–28.

[3] INIS THESAURUS 2016 IAEA-INIS-01 (2016/10) (Vienna),

[4] Vneshniy yaderny tsikl. Terminy i opredeleniya (External nuclear fuel cycle. Terms and definitions), OCT 95 10563-2002

[5] FAST REACTOR KNOWLEDGE ORGANIZATION SYSTEM.