AD662574

FTD-HT-66-515

FOREIGN TECHNOLOGY DIVISION

ON THE QUANTITATIVE EVALUATION OF THE TERMINOLOGY OF A VOCABULARY

by

L. G. Kravets

50 YEARS
1917 1967
GOLDEN ANNIVERSARY
FOREIGN TECHNOLOGY DIVISION

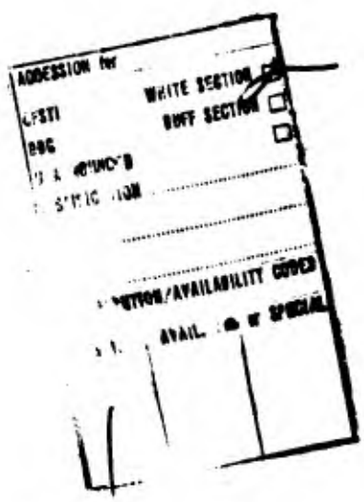
DDC
DEC 19 1967
A

Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce, for sale to the general public.

Reproduced by the
CLEARINGHOUSE
for Federal Scientific & Technical
Information Springfield Va. 22151

19

This translation was made to provide the users with the basic essentials of the original document in the shortest possible time. It has not been edited to refine or improve the grammatical accuracy, syntax or technical terminology.



UNEDITED ROUGH DRAFT TRANSLATION

ON THE QUANTITATIVE EVALUATION OF THE TERMINOLOGY OF A VOCABULARY

By: L. G. Kravets

English pages: 13

SOURCE: Nauchno-Tekhnicheskaya Informatsiya (Scientific Technical Information), No. 2, 1965, pp. 27-29.

Translated by: E. Harter/TDBXT

UR/0315-065-000-002 TP6001661

THIS TRANSLATION IS A RENDITION OF THE ORIGINAL FOREIGN TEXT WITHOUT ANY ANALYTICAL OR EDITORIAL COMMENT. STATEMENTS OR THEORIES ADVOCATED OR IMPLIED ARE THOSE OF THE SOURCE AND DO NOT NECESSARILY REFLECT THE POSITION OR OPINION OF THE FOREIGN TECHNOLOGY DIVISION.

PREPARED BY:

TRANSLATION DIVISION
FOREIGN TECHNOLOGY DIVISION
WP-AFB, OHIO.

ITIS INDEX CONTROL FORM

01 Acc Nr	68 Translation Nr	65 X Ref Acc Nr	76 Reel/Frame Nr
TP6001661	HT66005 15	BD5006831	1654 1074

97 Header Clas	63 Clas	64 Control Markings	94 Expansion	40 Ctry Info
UNCL	UNCL, O	O		UR

Q2 Ctry	03 Ref	04 Yr	05 Vol	06 Iss	07 B. Pg.	45 E. Pg.	10 Date
UR	0315	65	000	002	0027	0029	NONE

Transliterated Title
O KOLICHESTVENNOY OTSENKE TERMINOLOGICHNOSTI LEKSIKI

09 English Title ON THE QUANTITATIVE EVALUATION OF THE TERMINOLOGY OF A VOCABULARY

43 Source
NAUCHNO-TEKHNICHESKAYA INFORMATSIYA (RUSSIAN)

42 Author
KRAVETS, L. G.

98 Document Location

16 Co-Author
NONE

47 Subject Codes
05, 09

16 Co-Author
NONE

16 Co-Author
NONE

39 Topic Tags: computational linguistics, character recognition, machine translation, computer language, natural language

ABSTRACT : A quantitative method for selecting key words to be used in information retrieval language is described. The method is closely associated with the traditional problem of singling out terms. A term is defined as a unit in a vocabulary system used to express a system of concepts in a given branch of science. The process by which words become terms can be broken into two stages: entrance of the vocabulary unit into the sphere of limited function, and maintenance or acquisition by a certain part of the vocabulary of certain characteristics which distinguish the vocabulary of terms from the general vocabulary. There are various degrees in the process by which words become terms, and these degrees can be quantitatively analyzed. Four groups of terms may be distinguished, depending on the place which the terminological meaning occupies in the semantic structure of the word: 1) words all of whose meanings pertain exclusively to a given branch of science; 2) words whose primary meanings pertain to a given terminology; 3) words whose primary meanings do not pertain to a given terminology; and 4) words which are used in texts on a particular subject in one or several of their general meanings, which are to some extent specialized through combination with words that are terms, e.g., "accident" in "nuclear accident" and "start-up accident." The degree to which a word is used in its terminological meaning or meanings is called the weight of terminological significance in the semantic structure of the word. This is the ratio of the probability of the appearance of a word in its terminological meaning to the total probability of its appearance in the text being analyzed. This ratio is given in a scale from 1 (100 percent) to 0. Word combinations also must be quantitatively evaluated in scientific texts. The terminological valance of a word is defined as its potential capacity for forming combinations with words that are terms, and this is expressed as the ratio between the number of cases of combination of a given word with terms and the total number of cases of its appearance in combinations. Another aspect of terminological valance is the likelihood that a word will combine with other words which

A

TP6001661 HT6600515

are not terms to form a terminological combination. This is the ratio of the incidence of a word in terminological combinations to its total incidence in combinations. A direct connection has been discovered between the degree to which a certain category of words have become terms and their distribution characteristics in multi-component and substantive combinations in narrowly specialized texts. The distribution is analyzed as follows. Distribution models are set up on this basis: with respect to the nucleus or primary substantive word in a combination (N), the functions of the words in combination with this nucleus are defined as: 1) defining substantive, 2) attributive adjective, 3) adjective indicating nationality, 4) participle, and 5) qualitative adjective or adjectival pronoun. A table is given of the probability of occurrence of terminological combination according to the various models. The table is based on analysis of 3,300 three-component word combinations. The pattern revealed in the table makes it possible to get an idea of the probability that terminological combinations of a certain distribution pattern will occur. English translation: 12 pages.

U. S. BOARD ON GEOGRAPHIC NAMES TRANSLITERATION SYSTEM

Block	Italic	Transliteration	Block	Italic	Transliteration
А а	*А а*	A, a	Р р	*Р р*	R, r
Б б	*Б б*	B, b	С с	*С с*	S, s
В в	*В в*	V, v	Т т	*Т т*	T, t
Г г	*Г г*	G, g	У у	*У у*	U, u
Д д	*Д д*	D, d	Ф ф	*Ф ф*	F, f
Е е	*Е е*	Ye, ye; E, e*	Х х	*Х х*	Kh, kh
Ж ж	*Ж ж*	Zh, zh	Ц ц	*Ц ц*	Ts, ts
З з	*З з*	Z, z	Ч ч	*Ч ч*	Ch, ch
И и	*И и*	I, i	Ш ш	*Ш ш*	Sh, sh
Й й	*Й й*	Y, y	Щ щ	*Щ щ*	Shch, shch
К к	*К к*	K, k	Ъ ъ	*Ъ ъ*	"
Л л	*Л л*	L, l	Ы ы	*Ы ы*	Y, y
М м	*М м*	M, m	Ь ь	*Ь ь*	'
Н н	*Н н*	N, n	Э э	*Э э*	E, e
О о	*О о*	O, o	Ю ю	*Ю ю*	Yu, yu
П п	*П п*	P, p	Я я	*Я я*	Ya, ya

* ye initially, after vowels, and after ъ, ь; e elsewhere. When written as ё in Russian, transliterate as yë or ë. The use of diacritical marks is preferred, but such marks may be omitted when expediency dictates.

BLANK PAGE

ON THE QUANTITATIVE EVALUATION OF THE TERMINOLOGY OF A VOCABULARY

L. G. Kravets

Resume

The creating of an industrial system of machine translation with the automatic indexing of translatable materials presumes the development of dictionaries which provide the separating out of key words and word combinations, followed by the translation of them into an information-search language of the descriptor type. Three objective signs of the terminization of words are considered. The weight of the terminological significance in the sense structure of the word is expressed by the ratio of the probability of the appearance of a given word in its terminological meaning to the overall probability of its appearance in texts being analyzed. The potential capacity of composing with word-terms, or of introducing into the composition of terminological word combinations, is called terminological valence. The weight of the terminological valence in the first conception (TV_1), is determined as the ratio of the number of instances of composability of a given word with word-terms, to the overall number of instances of its appearing in the makeup of nominal word combinations. The weight of the terminological

significance in the second concept (TV_2), is expressed by the ratio of the number of instances of the appearance of a given word in the makeup of terminological compositions to the overall number of instances of the word's entering into any nominal word combination. The last of the indications considered of the terminological significance of a word are its distributive characteristics (probable positions) in the makeup of the multi-component nominal word combinations. The degree of the terminological significance is higher in proportion as the word is closer to interpositioning with the nucleus of the word combination. The conclusion is reached that all the nouns and words determining them, actively used in scientific-technical texts, are characterized by one or another degree of terminological significance, which fluctuates within the limits of 1 (100%) and 0.

* * *

In the not distant future, machine translation will become one of the links in the automation system of processing information. In this situation, the concept of machine translation takes on ever wider meaning signifying, along with the translation from one natural language to another, any algorithmitized process of conversion of significant units of processable texts.

Such a formulation of the question makes unavoidable the gradual approximation of the separate aspects of the overall problem of automation of information processes. In particular, in the area of patent information, there has already arisen the necessity for creating an industrial system of machine translation which will assure simultaneously with the translation, also automatic indexing of the patent materials, i. e., the translation of a determined part of the

units of the text of the natural language into the information seeking language*.

The indispensable elements of automatic translation and indexing are:

a) a two-language dictionary for some particular line with an aggregate of key words and word combination in it, separately - terms used in the particular line of knowledge;

b) a dictionary-thesaurus which contains a stock of terminological units designatable by descriptors.

The creation of each of the indicated dictionaries is tied up with a whole series of problems, of which, in our opinion, the most important one is the development of objective criteria for the selection of the key words and the subsequent connection of them into groups of equivalent terms.

Below, we deal with one of the quantitative methods of selecting such key words, which is closely coupled with the traditional problem of separating out the terms.

Objective Possibility of Quantitative Evaluation

The term proves to be the unit of the lexical system used for the expression of system of concepts of a given branch of knowledge. The process of setting up a system of concepts goes on continuously. In proportion as science and technology develop, there are constantly developing new concepts, and among the existing concepts, there are arising new connections and relationships. The objectivity

* The idea of creating a complex system of automated processing of the different kinds of foreign-patent literature was expressed by R. P. Vcherashnyy ("Mechanization of patent-information operations," Bulletin ("Information on invention" 1964, No. 12, 3-9)). Some preliminary remarks connected with the development of this problem are contained in the article [1].

of the existing concepts also, as well as the degree of their crystallization, differs. Therefore, by far not all of the concepts and their relationships find at once their places in the system of concepts of a determined branch of knowledge.

A terminological system represents the aggregate of lexical units regularly used for the formation and expression by the media of language of the units of the system of concepts. The new lexical units (words, word combinations) become the terms in a given branch of knowledge only after determining their position in the terminological system. It is natural to assume that to a different degree of conciseness of the formulation of concepts and relationships, there will correspond between them a different degree of terminological significance of the vocabulary (although between the mentioned phenomena one should not even look for unconditioned dependence).

Respectively, the process of terminization of a vocabulary can be developed in two stages:

The inclusion of a lexical unit in the sphere of limited functioning (i. e., placing on a determined part of the vocabulary supplementary systems of limitations).

The obtaining (or conserving) by a determined part of the vocabulary of some peculiarities which distinguish the terminological vocabulary from the one generally used.

Such a setup of the problem gives a basis for assuming the presence of different degrees of terminological significance of the vocabulary and, consequently, opens up the possibility of quantitative evaluation of the lexical units.

The Weight of the Terminological Significance of the Word

The analysis of words actively used in scientific and technical texts enables one to break them down at least into the following four groups, (as depends on the place occupied by the terminological significance in the sense structure of the word)*:

1) words, all the meanings of which refer to a given branch of knowledge:

nuclear yadernyy.

2) words, the first meaning of which relate to a given terminology:

reactor 1) reaktor; 2) El stabilizator.

3) words, the first meanings of which do not refer to a given terminology:

agent 1) deyatel'; 2) agent; predstavitel', postednik; 3) deystyushchaya sila, faktor; 4) khimicheskoye veshchesvo, reaktiv; 5) fizicheskoye telo;

4) words used in texts of given subjects in one (or several) of the general meanings, which are somewhat "specialized" being formed with word terms:

accident avariya, polomka;

nuclear a. avariya yadernoy ustanovki, sluchaynyy yadernyy vsryv;

start-up a. puskovaya avariya, etc.

The degree of usability of the word in its terminological meaning (meanings) let us call the weight of the terminological significance (meaning) depending on the sense structure of the word.

* In the given case there is used the "dictionary" criterion in breaking down the words into groups [1, 3].

In the analysis of the texts, the weight of the terminological significance can be represented with sufficient objectivity in the form of the ratio of the probability of the appearance of the given word in its terminological significance to the overall probability of the appearance of the word in the texts being analyzed. In this situation the weight of the terminological significance takes on a completely determined numerical expression within the limts of from 1 (100%) to 0. It is understood that in the case of the word nuclear having a single value "yadernyy", the weight of the terminological significance will be equal to 1. In the case of the words reactor and agent, the weight of the terminological significance proves to be less than one, since it is really possible to use the word in senses relating both to the given branch of knowledge and to bordering branches (or in general use senses). The weight of the terminological significance in the case of the word accident will be still less since its specialized meaning appears either in conjunction with word-terms or in determined context conditions.

Terminological Valence

One of the objectives of the use of the numerical evaluations is the joinability of words in scientific-technical texts.

It is noted that different words possess different potential capacity for joining up with word-terms. In this situation, the indicated capacity is greater in proportion to the weight of the terminological significance of one or another word, i. e., it can also be disconnected as an objective phenomenon of the degree of terminological significance of the word.

The potential capacity of uniting with word-terms is called

terminological valence*. The terminological valence can have different weight, which is determined as the ratio of the number of cases of joinability of the given word with word-terms to the overall number of cases of its appearance in the make-up of word combinations.

For example, it is possible to place side by side, the terminological significance of the words nuclear and heavy, of which the first is characterized by greater significance of weight of its terminological meaning.

In the texts examined (about 120,000 words), there were fixed 558 two-component combinations with the word nuclear, and in this in 509 cases, nuclear was connected with word-terms (n. absorption - yadernoye pogloshcheniye, n. weapons - yadernoye oruzhiye and others), and in 49 cases with non-terms, (n. age - yadernyy vek, n. parity - ravenstvo yadernykh sil and others). Hence the weight of the terminological valence of the word nuclear amounts to 509/558 = 0.91. The word heavy was met with in 68 two-component combinations. In this situation, in 43 instances it was joined with word-terms (h. nucleus - tyazheloye yadro, h. hydrogen - tyazhelyy vodorod, and others), in 25 cases with non-terms (h. equipment - tyazheloye oborudovaniye, h. tonnage - bolshoy tonnazh, and others). Consequently, the weight of the terminological valence of the word heavy amounts to 43/68 = 0.63.

The comparison of a sufficiently great number of two-component terminological and non-terminological word combinations shows that word-terms in being joined one with another, as a rule, form

* As to the essence of the concept of valence in linguistics, see for example, the articles [4] and [5]. The possibility of the use or probability evaluations of the joinability of words in a text is given particular treatment by the authors in the articles [6] and [7].

terminological combinations (plutonium fission - deleniye yader plutoniya, thermonuclear bomb - termoyadernaya bomba, and others). On the other hand, word-non-terms most often form non-terminological combinations (successful accomplishment - udachnoye osushchestvleniye, large amounts - bol'shiye kilichestva, and others). If only one of the components proves to be a term, then the probability of meeting with a terminological combination can be quite different. However, one observes a directly proportional dependence between the weight weight of the terminological values of the words and the probability of their entering into terminological combinations. On the average, in more than 50% of the cases as a result of the joinability of the terms and non-terms, there is formed a terminological combination (primary beam - pervichnyy puchok, nuclear accident, and others). In the remaining cases there are formed non-terminological combinations (dangerous radiation - opasnaya radiatsiya, atomic era - atomnaya era, and others).

In this connection, there is proposed still another interpretation of terminological valence - capacity for entering into terminological combinations with other words. Terminological valence in the first (dealt with above) and second concepts is given, respectively, the symbols TV_1 and TV_2. The weight TV_2 is expressed by the ratio of the number of cases of the appearance of a given word in the make-up of terminological combinations to the overall number of cases and its appearance in the make-up of any two-component word combination.

The comparison of the values TV_1 (capacity for joining with word-terms) and TV_2 (capacity for forming terminological combinations) indicates on the whole a directly proportional dependence between

these values. In connection with this fact, in the case of words with clearly expressed terminological significance, there is observed a tendency towards some increase in the value TV_2 as compared with the values for the TV_1. This tendency reveals itself in the fact that in the case of the majority of the indisputable terms TV_2 is greater than TV_1. Thus, in the case of the word nucleus, $TV_1 = 0.81$ (out of 91 cases of entering into two-component combinations, it was combined with word-terms, 74 times) and $TV_2 = 0.93$ (since out of 91 combinations, 85 proved to be terminological). The increase of TV_2 as compared with TV_1, is explained by the fact that in a number of cases these words form terminological combinations, joining also in word-non-terms. For example:

daughter nucleus - docherneye yadro;

heavy nucleus - tyazheloye yadro;

magic nucleus - yadro s magicheskim chislom neytronov;

odd-even nucleus - nechetno-chetnoye yadro, etc.

This fact reflects the process of the influence of the terms on the generally required vocabulary. The terms impart to the vocabulary the specific of the given area of knowledge, in short, they terminologize it.

Distributive Characteristics of Terminological Units

An analysis of nominal word combinations in scientific-technical texts enables one to establish a definite connection between the terminological quality of the word and its distributive characteristics (or probability position) in the make-up of multi-component word combinations.

The morphological analysis of the determining words which characterize in a varying degree, the terminological quality show

that the more numerous part of the terminologically "strong" determining words is expressed by nouns and relative adjectives, and among the terminologically "weak" words, there predominate qualitative and pronominal adjectives, as well as participles.

Together with the fact, that in the course of the distributive analysis, it was revealed [8] that in the position adjacent to the nucleus (N), most often there are fixed nouns in a determined function (determinants of class 1). Then there follow relative adjectives (class 2), adjectives signifying nationality (class 3), participles (class 4), and finally qualitative and pronominal adjectives (class 5). For example, new cascade theory - novaya kaskadnaya teoriya; anticipated nuclear burst - predpologayemyy yadernyy vzryv; advanced Swedish nuclear reactor - usovershenstvovanyy shvedskiy yadernyy reaktor.

There is noted a direct connection between the degree of terminologically determined category of words and its distributive characteristics in the make-up of multicomponent nominal word combinations from narrowly specialized texts.

The noted conformity to rule, enables one to put together a general presentation about the probability of the appearance of terminological combinations with one or another distributive model*. For example, in the analysis of the three-component combinations, the highest percentage of terminological combinations is noted in the distributive model 1 1 N - 89% (electron capture decay - raspad s zakhvatom electrona). Then there follow the distributive models

* The distributive model represents a sequence of indices of classes of determining words to which refer the components of a given combination, including the index N, which designates the nucleus of the word combination, for example, anticipated nuclear bursts - 4 2 N.

1 2 N - 71% (radiation chemical reaction - raditsionno-khimicheskaya reaktsiya), 2 1 N - 69% (nuclear chain reactor - yadernyy reaktor), etc.

The overall picture of the probability of the appearance of a terminological combination with one or another distributive model can be presented with the aid of the following table, which was composed on the material of the analysis of 3,300 three-component word combinations fixed in the analysis of narrowly specialized texts*.

Table

1 1 N 80%	2 1 N 69%	3 1 N 16%	4 1 N 29%	5 1 N 25%
1 2 N 71%	2 2 N 67%	3 2 N 12%	4 2 N 30%	5 2 N 12%
1 3 N не зафиксировано	2 3 N 0%	3 3 N 0%	4 3 N 0%	5 3 N 0%
1 4 N 65%	2 4 N 50%	3 4 N 0%	4 4 N 32%	5 4 N 0%
1 5 N 62%	2 5 N 21%	3 5 N 0%	4 5 N 0%	5 5 N 0%

Words in table: not determined.

It is not difficult to note that the displacement, both in accordance with the vertical and in the horizontal row, lead to a lowering of the probability of the appearing of terminological combinations. The greater the indices of the determining components, i.e., the farther from the nucleus the probability position of the representatives of the given glasses of determining words is, the lower the probability of the appearance of the terminological combination will be.

Thus, to the above listed objective signs of the terminological quality of a word, one can add also its distributive characteristics in the make-up of a multi-component combination of words. The degree

* The sequence of indices, for example, 1 1 N, 2 4 N and so on, indicates the morphological make-up of the word combination. The adduced percentages express the ratio of the number of terminological combinations of the overall number of combinations with the given model.

of the terminological quality of a word is greater in proportion as the relative placing of this word with regard to the nucleus of the original combination, is lower.

On the basis of the analysis made, one can draw the conclusion that all the nouns and the words determining them actively used in scientific-technical texts are characterized by one or another degree of terminological quality, which fluctuates within the limits of from 1 (100%) and 0. The possibility of quantitative evaluation of the indications under consideration makes possible a more objective division of the vocabulary of scientific-technical texts into categories of lexical units with a different degree of terminological quality. This, in turn, enables one to compute the permissibility of including lexical units in the list of key words.

Literature

1. Kravets. L. G.: Machine Translation in the System of Patent Information, Informatsiya po izobretatel'stvu (information on inventions), 1964, No. 12, 15-18.

2. The Short Oxford English Dictionary, 2nd Ed., 1936.

3. Myuller, V. K.: Anglo-Russian Dictionary, Moscow, 1960.

4. Andreyev, N. D., Berkov, V. P., and Zasorina, L. P.: Investigation of the Valence in the Transition from the Input Language to the Intermediary Language, Reports from the Conference on Processing Information, Machine Translation, and Automatic Introduction of the Text, Issue 5, Moscow, Academy of Sciences of the USSR, Institute of Scientific Information, 1961.

5. Leykina, B. M.: Some Aspects of the Characteristics of Valence, Ibid., Issue 2.

6. Sherry, M. E.: Syntactic Problem in Sepantic Analysis, Proc. of the IFIP Congress, 1962, Amsterdam, 1963.

7. Neethan, A. R.: Probabilistic Pairs and Groups of Words in a Text, Language and Speech, 1964, 7, No. 2, 98-106.

8. Kravets, L. G.: Analysis of the Structure of Word Combinations in English Scientific-Technical Texts, NTI (Scientific-Technical Institute), 1963.