International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Synonymy of Therms and Terms and Its Presentation in the Linguistic-Informative System

L. M. Badyorina

03134, Ukraine, Kiev, Zholudeva St. 4g, 128

e-mail: vada@ukr.net

Abstract: ***In the article we offer a model of calculation of a ratio parameter (relevance) of meaningful terms of reference definition in answers. Here is given an example of the calculation of a numerical synonymic parameter of meaningful terms.***

Keywords: information technologies, natural language, multifunctional model, linguistic multifunctional model, semantic coincidence, automated educational systems.

1. Introduction

Thanks to the current stormy development of the automated systems studies, the problem of construction of formal models describing various aspects in the subject field has become extremely important. Among them the leading position is taken by models and modes oriented on the automated evaluation of results of educational process.

It should be noted that if the construction of educational contents and integral systems in the noted area are sufficiently developed, the automation of evaluation processes is, actually, still in the initial stage. It is linked, first of all, with the circumstance, that the results of educational process appear as answers at examinations and because of that they have a naturally linguistic form. Consequently, the evaluation technology in such method gains the character of automatic (automated) comparison of naturally linguistic texts or fragments of texts.

The development of modern information technologies in the sphere of education has created the need for automated monitoring of student learning. Great importance to the educational purpose of automated systems must be devoted to model estimation of answers not in the form selected options but in the form of free text of arbitrary length regarding synonyms. The introduction of progressive forms of education creates the necessity of transition testing of computer students. Assessment of student learning is possible only through a comparative analysis of the reply with the reference text and determination of their relevance. Theoretically set model based on the synonymous terms of the subject field helps to set the correspondence between the reference and the actual definition, presented in the form of text of arbitrary length using words, synonyms [3]. Evaluation of correct answer text is based on the method of absolute coincidence of responses to one of the standards. Since the definition of a therm is formed on the system of basic concepts (terms), each of which has its own definition, it is proposed to calculate the index of task relevance of answers to open-use quantitative terms of synonymous field. Meanwhile, general scientific works devoted to this object are unknown to us; this fact has stipulated the necessity of writing this work.

2. General Structure of the System of Evaluation of Answers and Means of Its Modeling

Taking into account the naturally linguistic specific character of our research, the basic theoretical unit for modeling in the subject field, we have decided to assume the model of lexicographic environment (or computer-integrated lexicographic system), which was developed in a number of works.

Creating our model it is necessary to note formal correlates of linguistic constructions, which represent the essence of the subject industry, while modeling must take place both from the side of the form and from the side of the meaning. Moreover, we must take into account that a linguistic system comprises difficult hierarchy of various level complexes of units, objects and relations.

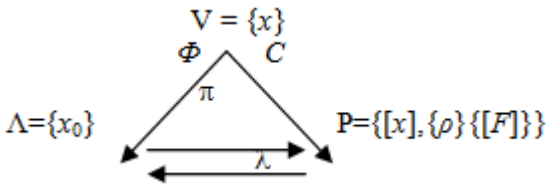
The first step on the way of construction of such model, to our opinion, there must be a modeling in the aggregate of lexical units, which represent the «dictionary» of the subject industry which is the research object, because a lexical subsystem itself plays the central role in a linguistic system in general.

We consider the noted dictionary must contain, above all things, «class of therms», which consists of the grammatically specified aggregate of lexemes of subject field. An adequate model for this purpose is a model of the grammatical L-system (G-system) in the structure of which we can select such structural elements as:

1. Class of elementary informative units V = {x}, that illustrates the class of all words in Ukrainian (in our case it is the class of therms of the subject field) ;
2. Class of initial forms, that illustrates exit forms for changeable parts of speech (for dictionary forms);
3. Class of curricula words: $\pi(x)= \rho(x)*\{\omega_i(x)\}$, and accordingly, factor of constant$\{\rho\}$ and variable

$[F]^k \equiv \{\omega_i(x)\}$ parts for all words (quasiroots and quasiflexes, accordingly);

4. Eventual factor of word-changeable (paradigm) classes: $\cup t_i/\pi_i$;
5. Operator of paradigmatizm π, what puts in accordance to every word of x its complete word-changeable paradigm [x];
6. Operator of lemmatization λ, which gives accordance to any word $\xi \in [x]$ its initial form x 0 . Schematically the structure of the G-system appears by such method:



(1)

Designating R operator **R** = π_ρ° Φ, where π_ρ – limit π on ρ, so for each ξ∈[x] we shall get

$$\mathbf{R}\xi = \rho(\xi). \quad (2)$$

Operator R will be used for the construction of the system of answers analysis. Consequently, the grammatical system has such structure:

$$\Gamma=\{V=\{x\};\ \Lambda=\{x_0\}=\Phi V;\ P=CV=\pi\ \Lambda=\{[x],\ \{\rho\}\ \{[F]\};\ \lambda x=x_0;\mathbf{R}\} \quad (3)$$

Thus, in accordance with general determination of L-system, the following condition is true: π°Φ = C та λ°C = Φ, where the composition of reflections is marked by character «0».

The functionally designated G-system allows us to select words (units of lexical level) from any text, make their grammatical identification, to set paradigm classes which they belong to, select the quasiroot and quasiflexity for every word – that is, to present in a form being adapted for subsequent analysis.

3. Statement of the Problem

To estimate the relevance degree of standard reference definition and the answer of a trainee it is necessary:

- To establish mutual monosymantic synonymic conformity with terms of standard reference definition and answer;
- To calculate the value of relevance parameter of standard reference definition and answer.

Statements are considered as a set of terms. Thus, the standard reference definition should be considered as a set of base terms, and the answer should be considered as a set of terms t, for each of them it is necessary to find a corresponding base term *e*. The search of conformity of a base term and an answer term proposes the definition of function $\mathring{a} = \varphi(t)$ and the calculation of the size of synonymic conformity $k = \theta(e,t)$. Thus, the pair $< e,k >$ will allow characterizing a term t in relation to a term-standard e. It means conformity of answer terms with base terms.

Let A be a set of standard-term definition, B - a set of answer terms.

Then the description of standard definition and answer is as follows:

$$A = \{e_1, e_2, ..., e_i, 1 \le i \le N\},$$
$$B = \{t_1, t_2, ..., t_i, 1 \le i \le M\}, \quad (3)$$

N - quantity of terms of standard definition;
M - quantity of answer-terms.

To calculate the conformity with terms of standard definition and answer it is necessary to characterize terms t according to terms-standards e. We are going to define synonymic conformity of answer-terms with standard definition.

4. Comparative Analysis of Terms

Knowledge - oriented approach to automation of assessment of learners' knowledge through the text answers provides the availability of modes in submissions of standard model in natural language, and learners' answer, to the formalized representation in a model of knowledge of subject areas. Each model of learners' answers is compared according to the reference model. A language structure as it has been said can contain a variety of logical and semantic relations between concepts, each of which set the degree of conformity. Let's make the following restrictions to the conformity assessment of answers to the standard sample:

- Consider as a response definitions (interpretations) of terms and concepts from a particular academic discipline;
- On the set of relationships being defined for terms and concepts from the relevant academic discipline we have selected only gender-specific attitude and relationship in synonymy.
- Definition in its broadest sense is a logical operation, in which it is possible to depict the content of the concept. There are seven rules being studied in formal logics about basic standard definitions of terms and concepts [2]:
- The concept is defined through generic and specific distinctions;
- Definition must have the same dimension, namely the amount of meanings of the determined notion, and the concept with which the definition must be consistent with each other;
- Specific difference should be a sign or group of characteristics inherent in this concept only, and no other terms which belong to the same generic concept;

- Definition should not contain a circle, i.e. The notion which is determined must not be determined by the concept that becomes clear only through concepts defined;
- The definition should not only be negative, because the objection shows no signs and gives essential features that characterize this notion;
- The definition should not be controversial in terms of formal logics;
- The definition should be clear, accurate and contain no double interpretation.

Let S - a set of standard definitions and terms from the relevant academic disciplines represented in the form of natural language text and signed by the above-defined rules. The set S is finite and disordered:

$S = \{s_i : 1 \le i \le n\}$, where - the definition of the term - a whole number/integer.

A set of learners' answers represented as natural language let's define as a set T. This set is a subset of the set S and has all its properties:

$T \subset S$; $T = \{t_i : 1 \le i \le m\}$, where m – integer; $m \le n$.

Each response from the set T may contain terms and concepts related to gender-aspect relations, or relations of synonymy notions of appropriate standard answer of set S. The relationship between terms and concepts in the given subject area (academic discipline) we will represent as a thesaurus. Thesaurus - dictionary that reflects semantic relations between concepts in a particular subject area and is designed to search for a given word in its semantic connections with other words [1].

The structure of Thesaurus typically includes the following ratio:

Concepts: = <gender-aspect> <part-whole> <synonyms> <antonyms> <association>.

The ratio of gender-aspect type allows including in the search box more abstract or concrete concepts. The ratio of part-whole includes the search box part of the whole object. The ratio of synonymy and antonymity allows you to search synonyms and antonyms. The ratio of associations are various and individual in its nature and indicate the dependence of contextual search terms.

Lerner's response is determined with the structure of certain concepts and terms, i.e., AC. According to certain restrictions, each concept in the explanatory part is described by synonyms.

The element e being the object of forming a set of forms (i.e. synonymous line), is given the name a basic term, other elements of the set (words-synonyms) are given the names of dependent terms. You must establish a correspondence between the Term and the Term of standard definition of response based on the notion of synonymous correspondence of terms under which it is substituted in the thesaurus, so you can calculate the ratio of standard definition and relevance of the learner's response. Thus, the standard definition should be regarded as a set of basic terms, and the answer as a set of terms t, each of which must find an appropriate base term e [3].

If A - a set of standard terms determination, V - set of terms of answer, then formalized representation of standard definitions and answers will be as follows:

$A = \{e_1, e_2, ..., e_i, 1 \le i \le N\}$, where N - number of standard terms determination.

$B = \{t_1, t_2, ..., t_i, 1 \le i \le M\}$, where M - number of terms of the answer.

As a result we can get one of these relations between sets A and B.

1. $A = B$ – the learner's response fully coincides with a reference response.
2. $A \subset B$ – the learner's answer contains all of the Terms of standard answers and additional Terms.
3. $B \subset A$ – the learner's response partially meets the standard answer, it lacks some basic Terms.
4. $A \cap B \neq \varnothing$ - the learner's response does not fully correspond to the reference response.
5. $A \cap B = \varnothing$ – the learner's response and the standard response present joint Terms. [3]

Let's outline the above mentioned in the following example. Let us have the standard pattern:

"Program is the description of the algorithm of solving the problem, given in the computer language." [3].

In the standard definition the key Terms are highlighted in bold italics, which correspond to the conditions of necessity and sufficiency of correct answers for learners. Other concepts are complementary. They may have a series of synonyms, but they are not included in the quantitative evaluation of responses of the learner. That is, the correct answer is determined by two necessary and sufficient notions which according to the rules for constructing explanatory of the term "program" shape its unique distinctive features. It is possible to construct a synonymous line for these basic terms of the thesaurus:

Algorithm: = (a set of rules, order of operations, a set of actions);

Language computing machines: = (language, artificial language, machine language, formal language, the language of computers).

Let's denote by A1 the set that defines the number of synonymous to the term "algorithm", and through A2 - a synonymous line to the term "a computer language.

Then the formalized representation of standard answers will be as follows:

Program: = description (submitted)
A1 ⊂ (algorithm; aggregate + rules; point + operations + total action) calculation (address; calculation)
problems
∧
set (submitted; description)
to A2 ⊂ computer language (+ machines + programming languages, machine + languages + formal languages, computer languages +).

In this example, the concept, through which interpretation is presented as a search pattern through a "+" combines the words which are terms for a given academic discipline, logical operation ∧ indicates compulsory presence of two basic terms. Other relations are missed because of restrictions imposed earlier.

This representation is the basis for comparison with the current responses of those who study. Let's consider Terms $\grave{a}$ (with standard response) and b (learner's response) coincide completely, if for b it is possible to find at least one search image with a synonymic line of thermo $\grave{a}$. That is for two terms a and b we can determine the function **f(a,b)**, which characterizes the rate of completeness interpretation of the term through the notion that it describes, in relation to synonymy and it takes the value 1 if **b** ⊂ **A,** and – 0, if **A** ∩ **b**=∅ assessment.

Let for a certain reference sample, we have the following current learner's response:

"The program is a sequence of operations on data necessary for processing the given algorithm.

This response is driven to the formalized representation. At this a search image of each word is included in the explanatory portion of the term, and compared with the elements of set A1, which determines the number of synonymous for Thermo "algorithm", and A2, which determines the synonymous line for Thermo "language of computers". Other words can also be checked with synonymous line of terms that are not key to the interpretation of the term "program", but they are not considered during the evaluation of learner's response. After filling the necessary transformations the formalized learner's response will have the following:

Program: =
A1 + (point) operations
data
necessity
treatment / treatments
information
implementations
A1 (algorithm).

From the given example it is possible to understand that the explanatory concept of the term "program" matches only with the set A1 of the reference sample. And there are 2 equivalents found in the answer because it takes the value 1of the formula (1) if we can find at least one match, so collapse all returned matches from one set gives a value of 1, **f(a_1,b)=1, f(a_2,b)=0**. Quantitative assessment is calculated by formula (4).

$$K = \frac{1}{2} \quad (4)$$

Synonymy of quasiroot and quasiflexity (terms) and subject area concepts (terms)

5. Synonymy of Therms and Terms and Its Presentation in the Lexicographic System

Working through naturally linguistic objects, and especially at comparing, it is possible to use a number of linguistic facts, relations etc. with the help of which the closeness between linguistic constructionsA and B is built.

The simplest model which can be applied in this case to our opinion is a model of lexical synonymy. We start from that supposition, that relation of synonymy between linguistic units of x and y, which is set as the condition of closeness of their semantic state $|c(x) - c(y)| < \varepsilon$, xSy, can be estimated by number $\delta = 1 - \varepsilon$, as a degree of synonymy between the members of one sin set can be different. For comfort we will mark the degree of synonymy as a magnitude $\delta = 1 - \varepsilon$, but we will consider thus, that maximally possible magnitude ε equals 1, and δ = 0 (when x and y are not synonyms), and the minimum possible magnitude ε equals 0, and δ = 1 (when x = y, or x and y are absolute synonyms).

Executing the quantitative estimation of degree of synonymy with the help of our method, we get as a result a synonymous matrix K(x y), x,y∈W, which on formal level is determined as a function from Cartesian work of W× W at the segment [0, 1]. The elements of synonymous matrix K(x y) are determined in the following method:

$$\begin{cases} K(x\ x)=1; \\ 0< K(x\ y)\le 1;\ x\neq y;\ xSy; \\ K(xy)=0 x\neq Sy. \end{cases} \quad (13)$$

Let's designate character $K^{R}(x\ y)$ of the matrix which appears with the help of K(x y) in applications to x and y and in procedures of R. Let us put on determination:

$K^{R}(x\ y)=K(\mathbf{R}x\ \mathbf{R}y)$. (14)

It means that we spread the set expert estimation of synonymous closeness from the therms onto their quasiroots.

6. Relevancy of Terms and their Definitions

Let's spread a concept of synonymy from separate therms, which in linguistic sense are elements of a lexical system, onto constituents of thesaurus of subject field Σ[**Z**].There are two aspects in this task – formal and semantic.

From the formal point of view the task consists in the establishment of semantic closeness, analogical to synonymous property, not in the set of separate therms, but in the set of chains $x_1\Delta_1\ x_2\Delta_2\ \dots\ \Delta_{q-1}x_q$, q=1, 2, … ., determined by formula (30), on condition that elements of $x_1, x_2,\dots, x_q$, get to the range of function definition of K(x y). A semantic aspect foresees the establishment of relation of semantic closeness, analogical to synonymous property, on the set of terms definitions:

$C^{\Sigma}(\mathbf{Z}) = \{C^{\Sigma}(z)\,|_{\forall z\in\Sigma(z)}\} = \{\{C^{\Sigma}_1(z);\ C^{\Sigma}_2(z);\ \dots;\ C^{\Sigma}_{l(z)}(z)\}\,|_{\forall z\in\Sigma(z)}.\}.$ (15)

A=z_M та B= z_N (length M and N, accordingly), designating:

As the concept of synonymy in linguistics is correctly determined only for the lexical system, to establish semantic closeness of elements from $C^{\Sigma}(z)$ we propose the name of relation of relevancy, which will be marked as REL. Let's define the quantitative measure of relevancy of two chains of A=zm and V= of zn (long M and N, accordingly), which will be marked as:

REL(A,B). (16)

Thus, the reflection of **REL** is determined: $C^{\Sigma}(\mathbf{Z}) \times C^{\Sigma}(\mathbf{Z}) \to \Delta$, where Δ – a certain subset of set of inalienable numbers. Let's consider that chain B is a relevant chain A, meaning that A **REL** B, only in case, when the value of function of **REL(**A,B) is not less than some certain $\delta\in\Delta$: **REL(**A,B)≥ δ, the choice of which depends on the specific of subject field and concrete tasks of the research and evaluation.

To find the obvious type of measure of relevancy, let's use such analytical model:

REL(A,B)= $\omega\eta$, (17)

where:

ω – certain function depending on numeric value of therm synonymy factor, re-entrant to A and to B, that is, it is a certain function of matrix elements of synonymy matrix K (x, y);

η – certain function of chain lengths A and B (meaning from M and N).

Function η sets dependence of relevancy level **REL(**A,B) on the quantity of therms in chains A and B, meaning integer numbers M and N. It is obvious, that only when the chains cross, they become maximum relevant and in that case **REL(**A,B) reaches its maximum value. The first property η appears here as follows:

(1). $\eta = \eta_{\max}$ only in that case when $M = N$.

It is obvious that function η is symmetric to variables M and N, and thus it is symmetric to its maximum value. The second property η comes here as follows:

(2). η (M, N) = η (N, M) i η is symmetric to value η_{max}.

The simplest function with such values is function $|M - N|$.

The next property of function η is linked with its behaviour at relevantly big differences of $|M - N|$. It is obvious that if chains A and B have big difference in lengths(meaning the quantity of valuable therms), they can not be relevant as each therm being absent in one chain changes the semantics of other chain, and with each therm the difference in semantics becomes bigger, and their interrelevancy becomes smaller. Here comes the third property of functionη:

(3). If $|M - N| \to \infty$ (or $M - N \to \pm\infty$), then $\eta \to 0$.

Conditions (1) – (3) set a certain functional equation with one decision: functionη (M, N) having the following:

$$\eta(M, N) = l(h)e^{-h|M-N|^2}, h > 0.\ (18)$$

Parameter h and function l(h) is experimentally set up and they can vary by user's condition; they can be defined by laying certain conditions of regulation.

Function ω depends on three variables: ratio of therms synonymy K (A, B) in chains A and B, quantity of valuable therms in chain A (that is from N) and the quantity of valuable therms in chain B (that is from M):
ω =ω (K (A, B); N; M).

It is obvious that quantity of therms in chain A equals the yield of set $\grave{A}$. The quantity of therms in chain $\hat{A}$ equals the yield of set $\hat{A}$ before the processing:

$$N = \|A\|, M = \|B\|.$$

If $N = const$, function ω has the following properties:

1. At increasing K (A, B) value ω is increasing.
2. At increasing the quantity of answer therms the value ω is decreasing.
3. If $\frac{\sum_{e\in B} k_t}{M} = 1$, then $\omega = \omega_{\max}$. In the latter formula values k_t are certain matrix elements of synonymy for the

so-called regulated (normalized) chains, which will be designated in some special work.

Generalized parameter K (A, B) equals the sum of maximum ratios of therms relevancy in chains A and B:

$$K(A, B) = \sum_{e \in B} k_e .$$

So, after analyzing the properties we get the following formulae to represent ω:

$$\omega = \frac{\sum_{e \in B} k_t}{M} . \quad (19)$$

Thus, we have come to the conclusion that the level of relevancy of chains A and B, being the sum of functions ω and η, shown in formulae (62), is designated by the following formulae:

$$\mathbf{REL}(A,B) = \omega\eta = \frac{\sum_{e \in B} k_t}{M} . l(h) e^{-h|M-N|^2} , \; h > 0 . \quad (20)$$

The necessities of practice are predetermined by direction and rate of evolution in computer informative systems. Stormy, often out of control and unpredictable progress of web-space has given certain characters to the noted development which already allow characterizing the fourth generation of informative systems. We relate such informative systems to the first generation which were created before appearance of control system (or management) by data bases.

The second generation is characterized by active application of classic data bases and by creation and application of different kinds of data. The apotheosis in this direction has become the formulation of relevant model of data, the development of numerous relevant and creation of developed and standardized language of queries as SQL and its varieties.

The third generation being post relevant informative systems is characterized by the combination of relevant models with the object-oriented approach to data modeling, programming, and application of agent technologies. Finally, the fourth generation of the informative systems being lingo-informative systems we bind it with application of human language mechanisms. The lexicographic systems and their generalizations serve as formal basis of this approach (lexicographic environments, lexicographic calculations and linguistic systems)

In fact, development of social knowledge persistently requires the systems of «content management». The noted problem, to our opinion, will determine the progress in communicative information technologies. This will be the key tasks for applied linguistics and linguistic technology for nearest decades.

7. Conclusion

The developed method of a trainee's answer analysis on the task of the open type allows establishing conformity with the terms of a subject area, used in standard definition and answer of the trainee. The result of conformity is the ratio between sets. The analysis of the results of synonymic conformity of terms of standard definition and answer give the possibility to account the numerical parameter of relevance.

References

[1] Hopfield J.J. Neural networks and physical systems with emergent collective computational abilities [Text] J.J. Hopfield // Proc. Natl. Acad. Sci.79, 1982.- p. 2554-2558.
[2] Електронний ресурс. – Режим доступу: http://ifets.ieee.org
[3] Badorina L. N. Method of the relevance degree estimation of the text answer in computer training systems [Text] / L. N. Badorina // Вісник Національного авіаційного університету. – 2007. – № 1. – С. 80–84.
[4] Кириличев Б.В., Широков Л.А. Системный анализ проблемы создания интеллектуальных компьютерных обучающих комплексов. Сборник научных трудов МГИУ [Текст] / Б.В. Кириличев, Л.А. Широков, П.Д. Рабинович. – М.: МГИУ. – 1996. – с. 166-171.
[5] Программированное обучение и кибернетические обучающие машины: Сборник статей под ред. Шестакова А.И. – М.: Сов. Радио, 1963. – 247 с.
[6] Ретинская И.В. Системы и методы поддержки принятия решений по оценке качества и выбору компьютерных средств учебного назначения /[Текст] И.В. Ретинская // Информационные технологии, 1997.- № 6. – с. 42-44