UDC 510.62

Ryabova N., Lysenko I., Shubkina O.
Computer Science Department,
Kharkiv National University of Radio Electronics,
Lenin Av., 14, Kharkiv, 61166, Ukraine,
E-mail: ryabova@rambler.ru, olga_shubkina@rambler.ru, lysenko_i@mail.ru

CONTENT – ANALYSIS SYSTEM FOR ADVERTISING SPECIALISMS TEXTS

Methods of the phonosemantic analysis of texts and the analysis of texts on the basis of allocated linguistic categories are considered. The phonosemantic analysis assesses emotional tone of the text on the basis of sound-color associations. The method of the so-called semantic differential uses 6 scales submitted by pairs Russian antonymous adjectives for the estimation of tonal tinge. The analysis and comparison of two assigned subjects texts is carried out on the basis of allocated linguistic categories.

Keywords –content-analysis, phonosemantic, category, linguistic categories, self-descriptiveness.

1. Introduction

Appearance of the content - analysis was a reaction to the arisen need for creation of objective methods of texts analysis. Problem area is that manual content - analysis is impossible when it is necessary to estimate big arrays of texts submitted continuously. A way out from the situation is the development of computer aided methods of content - analysis. Thus we exclude carelessness, as well as ambiguity in case criteria are accepted. Labor intensity in this case is solved due to speed. The present paper is devoted to computer methods of content - analysis of texts.

The initial stage of the content - analysis of texts is the phonosemantic analysis. Due to C.Osgud's works on studying synesthesia by means of a method of semantic differential (SD) one of the ways of experimental research of phonetic characteristics of language has appeared [1]. Thus, the word meaning can be represented as some point on a scale, set by two polar terms (for example, excellent - bad). For research several such antonymous adjectives are taken on the basis of which tonal tinge of the text is analyzed. Integral part of a mental image of a word and the text are sound-color associations caused by it. For this purpose it is important to define a number of vowel letters which are united by themselves in the certain classes.

Figure 1 gives phonosemantic analysis of the text.

Characteristics or elements of contents in relation to which procedure of calculation in the content - analysis is applied, may be words, word-combinations, sentences, paragraphs, texts. Individual words as elements of the contents, are a special case. In the content - analysis it is refered to as a category [2].

The category is a set of words incorporated together by different attributes. The present research uses texts sample of advertizing leaflets of different specialisms of the Kharkiv National University of Radioelectronics. After the analysis of text information the following categories of self-descriptiveness have been allocated: contacts, disciplines, places of work of graduates, opportunities of students, accomodation.study.

Figure 2 gives the example of the program operation for the analysis of self-descriptiveness of texts on the basis of linguistic categories.

2. Objectives

The purpose of the given work is the research of the content - analysis of texts using of allocated categories of self-descriptiveness, analysis and revealing «excess words» in texts of certain subjects, presence of missing data. Also the sound-color analysis of texts which should be to taken into account when choosing a word or construction of the text bearing a complete image, definition of influence of texts of assigned subjects on audience (the phonosemantic analysis), comparison of texts self-descriptiveness on the basis of the given categories and development of module of content - analysis.

3. Phonosemantic analysis

Phonosemantic analysis of the text, as well as of the word, consists in an estimation of sounding irrespective of the contents. Phonosemantic characteristics of the text are connected to deviations of frequency of sounds from the norm. For example, according to laboratory of suggestive linguistics "Vedium", nine of ten native speakers intuitively correctly imagine normal frequency of sounds and beforehand "expect" to meet each sound certain (usual, average) number of times in the text. If the share of any sounds in the text is within the limits of norm these sounds do not bear special semantic and expressive loading, their symbolics remains latent, hardly perceptible.

(However certain tendency is nevertheless present in most cases). The appreciable deviation of amount of sounds from norm sharply raises their self-descriptiveness, the corresponding symbolics is exhibited in subconsciousness of the listener (reader), coloring phonetic value of the text [3].

At the phonosemantic analysis of the text the list of qualities with which the given text associates with the indication of a degree of association in standard units, as well as verbal comment is given out. The theory of an experimental research of phonetic characteristics of language of C.Osgud on studying synesthesia by means of a method of semantic differential (SD) is put in a basis of phonosemantic module of the system.

The words understood as stimulus, cause various reactions distinguished from each other in two parameters: quality and intensity of its display (compare excellent – good – satisfactory – bad). The word meaning can be submitted as some point on a scale, set by two polar terms (for example, bright - dim). Semantic differential (SD) is a method of quantitative and qualitative indexing of value with the help of the bipolar scales set in pair of antonymous adjectives between which gradation of a degree of ocurrence of one or another word in the given quality are given [2]. To estimate phonosemantic influence Russian scientist A.P.Zhuravlyov offered to use 24 scales for the Russian language. These scales represent pairs of antonymous adjectives (for example, good - bad). The present research defines the tonality of the text on the basis of several scales, such as: active - passive, loud - silent, bright - dim, strong - weak, gentle - rough, good - bad [3]. It is based on the following algorithm:

For each attribute (a scale of evaluation)
{ For each letter
{ To compare rate of letter F (t) in the text with normal rate in informal conversation N (t);
The obtained difference to be multiplied by an estimation of the phonetic importance of the letter for a corresponding attribute (a scale of evaluation);
Summing up this value;
}
}

In the examined algorithm the sum is an estimation of the phonetic importance.

The sound-color analysis is a part of the phonosemantic analysis of texts. An inherent part of a mental image of a word and the text are sound-color associations caused by it [V.Turner, P.A.Florensky, etc.]. After carrying out some experiments, it has been established, that vowel sounds of speech in our perception quite definitely and basically for all are equally colored, though we do not realize it [professor A.P.Zhuravlyov]. The sound "*И*" (Russian) among vowels specifies blue color, "Ы" - black,"А", "*Я*" - red. When creating the text intended for reproduction in a certain color context (the poster, a publicity board, etc.) it is important to take it into account when choosing a word or construction of the text carrying a complete image, especially when it is accompanied by a video series. For example, excess of normal frequency of letter. The account of sound-color association of the text is especially important, as it should be in harmony with the common color scale of the text message [5].

Algorithm of determining emotional tone of the text (the sound-color analysis):

1. Determining amount of vowels letter in the text (count). We will count, vowels having divided them into groups as follows: "Е" together with "Э", "Ё" with "О", "Ю" with "У", "*Я*" with "А" (Russian).

2. Determining of total amount of letters in the text (total amount).

3. Calculation of a share of everyone letter in the text under the formula:

$$D(c)_i = \frac{\text{count}_i}{\text{total amount}} \qquad (3.1)$$

4. According to Table I calculation of the ratio of shares of every letter in the text to norm (Ratio_i):

$$\text{Ratio}_i = \frac{N(c)_i}{D(c)_i} \qquad (3.2)$$

5. Definition of a place of letter by their prevalence over norm. According to it is an allocation of the most typical color tones of the text.

TABLE I

NORMAL SHARES FOR LETTERS IN THE TEXT

Letter	Normal rate for letter, N(c)
Э+Е	0,085
О+Ё	0,109
Ы	0,018
У+Ю	0,035
И	0,056
А+Я	0,117

On the basis of the sound-color analysis it is possible to define the audience the given text information is aimed at. According to research of psychologists on preferences of certain colors by modern Russian native speakers, the following situation turns out: the articulated choice of the color appeared very much inequal in different sexual, age and social groups with a different educational level and political orientations [4]. On the basis of these data tables for definition of a text orientation are made.

Figure 1 gives the example of phonosemantic analysis of the text.

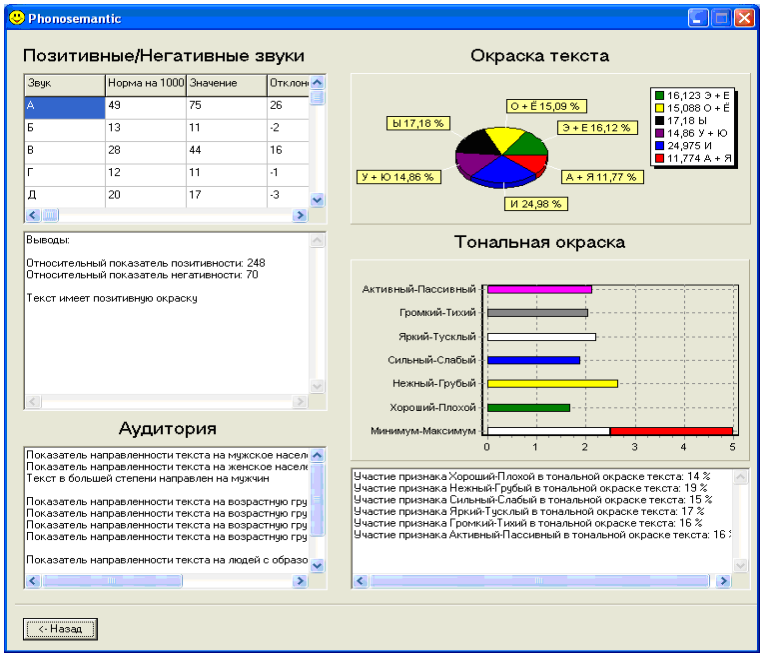

Fig. 1. An example of phonosemantic analysis of the text

4. Determining of self-descriptiveness of the assigned subject text information

The simplest by contents and at the same time the most fundamental in the content - analysis are simple estimations of frequencies. Let us set the following designation f (c, t) as frequency of occurrence of characteristic c in the text t [1]. Separate words as elements of the contents, are a special case of what in content - analysis is refered to as a category. The category is a set of words incorporated together by this or that attribute. For example, as a category Accomodation group of synonyms may stand {a den, the house, dwelling, habitation, a den, a dwelling, a tenement}. Other examples can be categories of aggressively colored vocabulary Aggression = {to beat, storm, threaten, spite, to overcome, pogrom to growl, …} and positively colored vocabulary the Positive = {gratitude, vigorous, tasty, kindness, gentle, nurse, warm, a joke, humour, clear, …}. Frequency of mentioning in the text of some category is counted up as the sum of frequencies of words included in it, i.e. if K is a category,

$$f(K,t) = \sum_{w \in K} f(c,t) \qquad (4.1)$$

Logic operation, underlying creation of a category, is a definition through abstraction. It is not at all compulsory that the category should be set by means of beforehand fixed list of words. It is sometimes much more convenient to set it operationally. An example of such category can be a category of verbs of past tense. Definition of belonging to it will consist not in comparison to the fixed list of words, but in a recognition of grammatic attributes of a verb of past tense. More difficult are the categories not simply consisting from separate words, but from complete word-combinations. For example, the category Sea = {Black Sea, Mediterranean Sea, Red Sea, Baltic Sea, …} [4]. Content - analysis with the use of categories allows estimating texts at higher abstract level. The results obtained with their help are qualitatively richer.

After the analysis of advertising information the following categories of self-descriptiveness have been allocated: “Contacts”: address, phone number, contact, mail; “Disciplines”: teaching students, training of experts, curriculum, directions, knowledge acquisition; “Places of work of graduates”: places of work, work of graduates;

“Opportunities of students”: traineeship, cooperation, predegree practice; “Accomodation. Tuition”: a hostel, laboratory, computer classroom, post-degree education, form of training, the state-guaranteed order, military-training.

The example of defining self-descriptiveness on the basis of categories is shown in Figure 2:

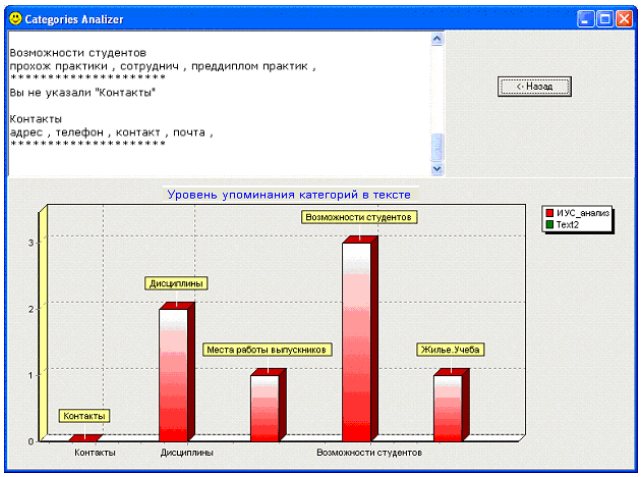

Fig. 2. Example of allocation of categories

5. Results and Business

On the basis of the developed software product (Content analyser) the analysis of advertizing leaflets of various specialisms or faculties of Kharkiv National University of Radioelectronics (KhNURE) has been made. As a result of the research of advertising texts of specialisms of university, the following regularities have been determined. "Excessive words" have been found in a number of texts, namely: for advertising a specialism Information Managing Systems and Technologies - the word "it", Computer Engineering and Management - "it", Economic Cybernetics - "so", Applied Mathematical Methods - "it", "that", Electronic Engineering - "only". Basically for the analysed array of the text information black color (Russian letter "Ы" corresponds to it) is one of the three dominating ones. It is a negative characteristic because a great number of sibiliant, guttural sounds generate a rought text violent for subconsciousness. Absolutely all texts of advertising of specialisms or faculties are directed at male population. The parameter of an orientation at a male population is most strongly pronounced for Information Managing Systems and Technologies, Household Electronic Equipment and Intellectual Decision Support Systems. The age directivity parameter for almost all advertising texts prevails for people of up to 25. As a whole, approximately for half of booklets this factor does not differ strongly for age of 25-35 and 36-55. I.e. the information causes the greatest appeal for people of the above age (first of all for applicants and then for parents).

For advertizing leaflets of almost all specialisms (except for one) the category "Contacts" is allocated - exact coordinates are specified in the text, for example, address, phone number or e-mail. It should be noted, that all analysed set of text information contains a category of "Disciplines" so, the applicant is informed, on what directions training on this or that specialism would include so that he could make a proper choice while applying. The positive tendency is that for all advertising texts, except for Household Electronic Equipment, the category of "Employment" is allocated, having rather identical factor which values vary from 1 to 3. The majority of graduates in the future have an opportunity to be employed according to their specialism and it is clearly indicated for the applicant. The category of "The Opportunities of students" is badly expressed. The majority of brochures does not give the information on the international cooperation, traineeship. One third of advertising texts does not contain the information for the last category including accommodation in a hostel, information on laboratories, the computer classrooms, post degree education, form of training, order, the state-guaranteed, military training.

6. Conclusions and future work

Nowadays large state and commercial organizations have more and more difficulties with keeping up with dynamics of change in the information field. Regular acquaintance with publications for any serious activity is necessary, but it is not always enough. The arrays of information should be exposed to qualitative analysis to

evaluate a state of affairs in the field and to forecast the development of a situation (competitive intelligence). According to statistics, about 80% of the information for computer intelligence can be extracted from open sources of text information [7].

Convenient tools for work with the information are the programs, helping to collect and analyze text materials. The first stage of computer investigation is a content-analysis of texts, i.e. the analysis of self-descriptiveness of texts and the analysis of influence of the information to its perception in human subconsciousness.

The basic result of application of a system of computer estimation of the contents of the text information is not calculation of a "good/bad" ratio or allocation of self-descriptiveness, but formation of a frequency portrait of all positive and negative events connected in the text.

Developed phonosemantic module of a «Content analyser» is universal. It is capable of carrying out the analysis of texts of various subjects, scope and contents. The module of defining self-descriptiveness carries out estimation of the importance of texts according to linguistic categories, as well as it provides comparative characteristic on the basis of these categories.

A new module will be added to application in the future. It will allow to check up to what theme the text refers. The first phase of this checking will be a phase of training. In this phase for set of texts of the set subjects norm will be derived. Next step is a verification of the text. In this phase conditional frequency by each theme will be calculated.

$$\left| pr(c,t_i) - nr(c,T) \right| \times L(t_i) \qquad (6.1)$$

Where pr(c,ti) - conditional probability of theme in the current text, nr(c,T) – the norm for texts of this theme, obtained on the first step, L(ti) – length of the current text (number of words in the current text).

Using norm and conditional probability, probability of deviation of conditional frequency of text content will be calculated by the formula 7.1. If the result value is greater than 0.05 then deviation of norm is not accidental and we cannot relate text to this theme.

References

[1] Психосемантика слова и лингвостатистика текста: Методические рекомендации к спецкурсу. / Сост. А.П. Варфоломеев; Калинингр. ун-т. – Калининград, 2000

[2] Шалак В.И. Контент-мониторинг текстовой информации. Элементы математических методов компьютерного контент-анализа текстов, Научно-практическая конференция "Проблемы обработки больших массивов неструктурированных текстовых документов", М., 2004

[3] Журавлев А.П.Звук и смысл – М., 1991

[4] Шалак В.И. Современный контент-анализ – М., 2003

[5] Ермаков А.Е., Киселев С.Л. – Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’2005. – М., Наука, 2005

[6] Федотова Л.Н. "Анализ содержания - социологический метод изучения средств массовой коммуникации". - М.: Институт социологии РАН, 2001.

[7] Опарин А. "Системы мониторинга и анализа СМИ." — "PCWeek" 16-22 декабря 2003 №47(413) М.