USING SYNERGETIC INFORMATION THEORY FOR SPEECH DIAGNOSTICS OF SPECIALISTS BY MEANS OF DIGITAL TECHNOLOGIES

BIKESH OSPANOVA¹, AINUR SEILKHANOVA², AZAT ABDIN³

¹Head of Russian Language and Culture Department, Abylkas Saginov Karaganda Technical University, Karaganda, Kazakhstan
²PhD Student, Department of the Theory and Practice of Foreign Language Training, Faculty of Foreign Languages, Buketov Karaganda University, Karaganda, Kazakhstan
³Master of Engineering, engineer of Russian Language and Culture Department, Abylkas Saginov Karaganda Technical University, Karaganda, Kazakhstan

E-mail: ¹o.b.r@mail.ru, ²seylkhanova.ksu@mail.ru, ³azat_abdin@mail.ru

ABSTRACT

The article deals with the use of entropy criteria to identify language mistakes in spontaneous oral speech. The key problems of synergetics as an interdisciplinary scientific trend are analyzed. Speech in the form of an oral statement (text) is studied as a complex multi-level synergistic system. Its structure is analyzed within the framework of multidimensionality. Some aspects of the experimental approach to calculating entropy in the oral speech of bachelor, master, doctoral students and specialists are outlined, and experimental data demonstrating the results are presented. A linguo-mathematical model for speech diagnostics of specialists is proposed. It is built on the basis of the fundamental law of conservation of the sum of information and entropy using the Shannon formula.

In addition, the article shows the importance of synergetics in linguistics that represents today a new unifying trend, the purpose of which is to identify common ideas, methods, patterns of language transition from one level of organization to another.

It should be noted that present day linguistic research requires the use of digital technologies, computer science, mathematics and other methods of information processing to identify the essence of linguistic phenomena. Today computer technologies play the key role in all the spheres of social life, so the main type of human activity is increasingly the ways and methods of obtaining, storing and expanding knowledge and information.

Thus, the most urgent task of present day linguistic education is the use of digital technologies in scientific research.

Alongside with this, there arise new aspects of discussing and considering the synergetic approach to analyzing a language. At the present stage of language education one of these issues is studying speech diagnostics in the flow of professional communication.

Keywords: Information, Linguistics, Digital Technologies, Diagnostics, Speech.

1. INTRODUCTION

A language as an object of linguo-synergetics is a very complex system of intra-level and inter-level dependences. It should be noted that a language as a non-linear open environment is a self-organizing process therefore, despite the difficulties that arise in any undertaking the synergistic process in linguistics is transformed changing the views on a language as a system, changing a language itself. The introduction of synergetic ideas into the language contributes to its enrichment with new definitions, categories, terms. It will cause its internal transformation forming the linguo-synergetic scientific trend.

Today, our life is synergetics that is not only a new paradigm of scientific knowledge but also a new interdisciplinary trend.

One of the most interesting works in the field of linguistic synergetics is the research of
Kazakh scientist Z. Zhanabayev, who considers the prospects for the relationship of linguistics with synergetics.

According to Z. Zhanabayev, a language as a means of communication between people, of thoughts transmission should be considered as a self-organizing system, since this “stochastic, randomly formed macroscopic system of signs consists of many structural elements: phonemes, words, sentences. Phonetic, grammatical, syntactic rules reflect the emergence of order from chaos, self-organization [1].

In his research, famous Russian linguist, expert in the field of psycholinguistics V.G. Borbotko pays considerable attention to the origin of a language as a dynamic self-developing system drawing on the material of various languages and the ideas of synergetics. According to the scientist, the prospect of exploring a language as a synergistic entity requires deep reflection in terms of dynamic and linguistic aspects.

The author notes that the view of a language as a self-developing system that goes back in its origins to the teachings of W. Humboldt, contrasts with the firmly rooted position, according to which a language has been invented by man, and the relationship between forms and meanings is purely conditional. Indeed, many words of the modern language are invented, often contrary to grammatical rules. And yet, studying the deep dynamics of verbal and cogitative activity speaks in favor of the fact that a language has been formed spontaneously, regardless of the will of man; a person is free to consciously only normalize the language, but is not possible to change either phonetics or grammar in it. As for vocabulary, new words are always built on the basis of already existing language elements.

According to the linguist, a language as a symbolic system is capable of self-complication, self-organization and self-enrichment [2].

Due to constant interaction and data exchange with other information systems and external environments, “it is in the state of fluctuation, balancing between synchrony and diachrony, preservation and change, organization and entropy, “order” and “chaos”, well-known Soviet linguist R.G. Piotrovsky writes about a language. “... for all its external mobility, language and VCA (verbal and cogitative activity) retain their coherence due to the action of self-organization and self-regulation mechanisms [3].

Thus, our speech as a self-organizing system is closely connected with the external and internal environment, with the processes taking place in society itself, as well as within it, associated with the speech system, which indicates the openness of the language as a speech activity, as a system.

In linguo-synergetics, the interaction of a language with internal and external processes leads to oscillations, that is, “fluctuations”, which makes it possible to speak of the instability, non-linearity of the system.

Studying the synergetic ideas in linguistics seems to be quite promising. At this, the synergistic approach to the consideration of a language (speech) is based on "self-organization, the procedural nature of verbal and cogitative activity" [4].

Undoubtedly, if to talk about verbal and cogitative activity, it is necessary to take into account the parameters of communication: the situation, purpose, subject of speech, etc., as well as its communicative environment that is not an external factor but a participant in the communicative process. In other words, linguo-synergetics is based on the approach to a language (text) as a self-organizing system that is closely interconnected with the external environment (consciousness of communicants and the general language system), whose influence, alongside with the internal processes of the speech system, leads to functional fluctuations.

In turn, studying verbal activity as a non-equilibrium, unstable system that changes under the influence of fluctuations allows considering an oral statement (text) as an object of synergetics, as a complex multi-level natural object. Exploring the text as a synergistic process, many well-known linguists analyze the text in their works taking it as an open non-linear self-organizing system.

2. MATERIALS AND METHODS

The success of the work is largely determined by selecting research methods that allow solving the tasks and achieving the goal. The methods that have been used in obtaining the most important scientific results in the framework of the study can be characterized as follows. During the study, linguo-statistical, linguo-stylistic methods, computer modeling, mathematical methods of linguistic analysis, general scientific methods of system analysis, quantitative analysis method, information analysis, as well as the calculation of speech mistakes in the respondents' oral speech within the framework of the Shannon formula approach have been used.

The proposed methodology is based on the systematic, multi-level approach to building a complex hierarchical language system.
In addition, within the framework of the ongoing project, qualitative and quantitative assessments of the perception of the speech competence level have been applied. There has also been shown the significance of the work conditioned by the overdue trends in the interdisciplinary relationship of present day innovative technologies with computational linguistics, mathematical and computer modeling, and information theory.

3. LITERATURE REVIEW

The basis of studies dealing with the methodology of entropy and information, in the authors’ opinion, can serve the works of such scientists as S. Angrist and L. Hepler [5], C. Shannon [6], R. Arneheim R. [7], L. White L. [8], R. Narashimha [9], R. Karnap [10], and many others.

Many different works can be cited on this topic. The concept of entropy was introduced by Clausius in the 19th century as a measure of the degree of disorder. With the help of entropy, it became possible to evaluate such important concepts as order and disorder. For example, S. Angrist and L. Hepler give the following definition of entropy: "... entropy is defined as a quantitative measure of disorder in a system..." [5].

According to Doctor of Philology M.Yu. Oleshkov, "...entropy is understood as a measure of uncertainty (unpredictability) of the text development in the discursive process characterized by the possibility of selecting a certain stage from a number of options as a subsequent stage. The entropy index quantitatively characterizes the level of information orderliness of a text as a system: the higher it is, the less ordered the system (=text), the greater its deviation from the “ideal” development. Thus, entropy is a state function; any state of the system can be given a well-defined value of entropy” [11].

Considering the language by the methods of information theory has become a promising scientific trend that studies complex systems from the point of view of the processes of self-organization taking place in them. Within the framework of this trend, a language is modeled as a complex, dynamic, self-organizing system from the disordered state to the ordered one.

So, for example, domestic scientists, such as R.G. Piotrovsky [12], R.K. Potapova [13], V.V. Boguslavskaya [14], I.V. Baevsky [15], V.G. Sadur [16], B.Yu. Gorodetsky [17] and many others were engaged in theoretical research in that area.

Note that the use of information theory to an oral speech utterance was developed by foreign scientists, such as R. Jakobson [18] and the staff of the Prague School. The information theory had become a field of activity for linguists who tried to apply certain concepts and methods of the information theory to solving linguistic problems.

For example, scientist E. Janch [19] in his work gives an original concept of synthesizing natural and human sciences. In the work by R. Kehler [20] the language model is reconstructed in the unity of structure and dynamics.

At present, due to the development of digital technologies and specialized software, many scientists continue dealing with the problems of oral speech. So, for example, the studies of K. Zechner [21], T. Nielsen [22], R. Rosenfeld [23] on interpolation weights and decomposition allowed determining the evaluation set and dwell on the morphological analysis of the word. The approaches of H. Strick [24], E. Holliman [25], M. Marcus [26] and their developments in the field of word-formation language models and the data post-processing step helped understand that it was necessary to introduce restrictions for combining morphemes and changing the mistake rate. The works by M. Metir [27], O. Shiohan [28] directed us to studying the newspaper text through peculiar phrases for spontaneous speech, using words outside the vocabulary, which led to the development of the training set of oral speech.

4. RESULTS AND DISCUSSION

Today, any scientific research is carried out using present day digital technologies. Immediately, let’s note that in linguistics, in order to obtain significant speech information for analysis and accurate data, methods and means of computer technology should be involved.

At present, the versatility and complexity of such a phenomenon as oral speech seems to require a systematic approach. Oral spontaneous speech is one of the most relevant trends in the development of computational linguistics and present day information technologies.

In this study, the authors have tried using the computer program developed to carry out speech diagnostics of specialists from the position of information theory, when entropy in oral speech serves as a measure of uncertainty, chaos and disorder.

The object of the study has become speech activity of a specialist, namely, his oral speech in various situations of communication. At this, the
focus is on a new linguo-mathematical model with
the theoretical justification based on the use of the
fundamental law of conservation of the sum of
information and entropy. It has been built using the
Shannon formula. With a general description of the
entropy-information (entropy is a measure of
disorder, and information is a measure of removing
disorder) analysis of texts, the Shannon statistical
formula has been used to determine the perfection,
purity of the text:

\[
H = - \sum_{i=1}^{N} p_i \log_2 p_i
\]

where \( p_i \) is the probability of revealing some unit of
the system in their multitude

\[
N; \sum_{i=1}^{N} p_i = 1, p_i \geq 0, i = 1,2, \ldots , N
\]

In order to determine the degree of the
communicative skills of specialists formation, non-
survey research methods have been used, in
particular, an experiment has been carried out to
identify language mistakes in spontaneous oral
speech, which is an indicator of the general culture,
intellect and intelligence of a person. At this, the
peculiarity of non-survey methods consists in that
the subject behaves naturally. Such a research
principle, in the authors’ opinion, allows making
appropriate measurements of the linguistic literacy
level and detecting speech mistakes.

Experimental approbation of the practical
program of the synergetic theory of information,
which calculates communicative defects and failures
in oral speech, has been carried out at a technical
university.

The main task of the study at this stage has
been considered implementing substantiation of the
model of professional communication reproducing
the general principles of constructing an oral
statement based on the entropy analysis. The entropy
indicator demonstrates a quantitative measurement
of the ordering of information in an oral statement,
namely, the higher the indicator, the less ordered the
statement in the system; from here it follows that
there is more chaos in the system than information.

For implementing the entropy analysis of
the text, it is necessary to clarify the meaning of the
concept of "communicative defect". This concept
should be understood as communicative linguistic
facts that occur outside the norm and represent
mistakes and communicative failures.

A communicative defect is understood as non-standardized communicatively manifested
linguistic facts including actual mistakes and
communicative failures, that is, when the speech
norm of the speaker is not respected. So, for
example, N.L. Shubina notes that a communicative
failure occurs when communication participants use
a different set of codes to transmit and to receive
information [29].

An insignificant level of language
competence, insufficient culture of language
proficiency is perceived as a mistake, while the
concept of a "communicative failure" comes into
conflict with the concept of a "norm". This is caused
by non-compliance with the language (speech)
norm. Scientists have noted that the origin of the
communication failure is explained by the use of a
different set of codes by communicants when
transmitting and receiving information.

For a specific test of entropy analysis, there
have been obtained the data based on oral speech.
Due to the high costs involved in such a study,
testing has be initially been limited to one language.
For reasons of greater accessibility, Russian has
been selected as the first language for testing. As a
material for the experiment, spontaneous oral speech
of bachelor, master, doctoral students and specialists
has been used. Oral speech has been processed into
a textual material, and with the help of a computer
program, the information entropy in the speech of
the subjects has been calculated, i.e. there have been
detected communicative defects in speech.

Spontaneity is referred to the quality of
speech manifested in its unpreparedness, which is
expressed at the lexical, morphological and semantic
levels. In addition, a characteristic feature of
spontaneous speech, both monologue and dialogic,
is that a thought arises in the speaker and is verbally
formed at the very moment of utterance. It is
characterized by such properties as initiative,
unpreparedness, improvisation, natural tempo, free
use of language forms. Spontaneous speech is
always conditioned by the real situation of
communication. It assumes that skills and abilities of
using a language material must be brought to such a
level that speech communication can be performed.
Without any doubt, the skills of spontaneous,
unprepared speech are not formed by themselves but
are developed in the process of careful, consistent
preparation.

The category of spontaneity in linguistics
was first identified by S. Balli in describing
dialogical speech. He defined spontaneity as
unplanned, thoughtless speech limited by lack of
time [30]. However, there is a difficulty in
interpreting the term spontaneity, which lies in its
different understanding:
- spontaneous speech is such speech that is born on the initiative of the speaker, without the influence of any external motives;
- the concept of spontaneity is not polar to the concept of preparedness;
- spontaneous speech is interpreted as ill-conceived speech, not prepared in advance, which simultaneously reflects the processes of thinking and speaking.

So, scientist-linguist D.V. Zhabin who was engaged in studying spontaneous speech, noted the difficulties of its scientific study. According to the researcher, “…the complexity of studying spontaneous speech lies in the fact that, with unprepared speech reproduction, its structure differs from other types of oral and written speech at the level of phonetics, vocabulary and syntax. This in turn requires the development of other methods plus, the recording and accuracy of transferring all the features of this type of speech is another difficulty for identifying the mechanism by which it is possible to describe a system containing the volume and specificity of the speaker's spontaneous utterance" [31].

To identify the computation of the level of entropy and the information volume of the text on the example of analyzing a graphically fixed speech fragment, spontaneous oral speech of bachelor, master, doctoral students and specialists has been recorded and processed. The speech is presented without editorial corrections. When analyzing the text, the authors have encountered some difficulties. The fact is that in the course of oral presentation it is very difficult to isolate the structure of the sentence itself, since the boundaries of a single statement are lost when translating an oral text into a written form.

In addition, the issue of the text size has arisen, since its volume affects the quality of the analysis and causes difficulties. In the authors’ opinion, a more optimal unit would be the statement as the smallest part of the text, which is distinguished by relative communicative independence. In accordance with this, it has been decided to select the counting of the number of propositions in a segment of the statement, thanks to which it becomes possible to establish the speech mistakes of the material under study. Of course, this option does not give the most accurate and ideal indicators, however, it is one of the most reliable one among those available.

For a more visual and meaningful presentation, the authors have settled on a text volume of 300 words. Of course, this is not an ideal option but in general it allows quantifying the information volume of the analyzed material and detecting language mistakes. Numerical data contained in the text have been written in cursive. Oral statements of the respondents have been used and processed as the experimental material.

In the course of the experiment, a computer program has been developed to calculate the entropy in the oral speech of specialists and to identify the indicator of speech diagnostics in various communicative situations. The object-oriented programming language CSharp or C# (C Sharp) has been selected as the programming language. The program allows analyzing textual information based on the computational search for entropy in the oral speech of the respondents.

This program performs the analysis of the entered text by the following points:
- a message length;
- a number of alphabet characters;
- message entropy;
- average entropy of a symbol;
- a character length with uniform coding (bits);
- a message length with uniform coding (bit);
- absolute redundancy in the presentation of the message;
- average absolute redundancy in character representation.

In addition, the program encoder uses Shannon-Fano compression algorithms with variable-length codes: a frequently occurring character is encoded with a code of a smaller length, and a rarely occurring character is encoded with a code of a larger length. Each character entropy and its excess in the text have been determined.

For a more visual and meaningful presentation, there is shown starting the program, which is performed by opening the file "Program based on the computational search for entropy according to Shannon and analysis of textual information.exe".

After opening (Figure 1) there appears the program interface that consists of the following items:
1. Text entry field.
2. Modeler.
3. Encoder.
Considering the forms of information representation, let’s note the fact that the analogue form is natural for the human senses, but the discrete form of information representation using a certain set of signs should still be considered universal. In particular, it is in this way that the information presented is processed by a computer, transmitted via computer and some other communication lines.

As for the experiment, the result is simple in solution and in the technique of proving. The proposed approach to calculating the level of entropy in an oral statement based on the Shannon formula has its own significance.

In the course of studying, in the program developed there has been carried out the analysis of oral speech of bachelor, master, doctoral students and specialists (Figures 2-5).
Figure 3: Analysis of oral speech of master students

Figure 4: Analysis of oral speech of doctoral students
Figure 5: Analysis of oral speech of specialists

Table 1: Results of computation search for entropy according to Shannon and analysis of the text information

<table>
<thead>
<tr>
<th>Oral speech</th>
<th>Statement length (characters)</th>
<th>Character mean entropy</th>
<th>Character length with uniform coding (bit)</th>
<th>Mean absolute redundancy in representing a character</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bachelor students</td>
<td>1995</td>
<td>4.76043</td>
<td>6</td>
<td>1.23957</td>
</tr>
<tr>
<td>Master students</td>
<td>2018</td>
<td>4.64635</td>
<td>6</td>
<td>1.35365</td>
</tr>
<tr>
<td>Doctoral students</td>
<td>2508</td>
<td>4.68652</td>
<td>7</td>
<td>2.31348</td>
</tr>
<tr>
<td>Specialists</td>
<td>2456</td>
<td>4.67081</td>
<td>6</td>
<td>1.32919</td>
</tr>
</tbody>
</table>

Thus, the results of studying and analyzing Shannon entropy in oral speech show that the mean redundancy in representing a character is high for all the testees. It is noteworthy that when computing the entropy according to Shannon, the program does not take into account the features of a particular language, which in turn does not reflect the meaning of the transmitted information.
5. CONCLUSION

The results of the experiment show that the analyzed texts have different levels of entropy. In the authors’ opinion, the oral statements of bachelor, master and doctoral students have high entropy, i.e. there is a significant number of structural-semantic, communicative defects. They violated the integrity and coherence of sentences, there is a lot of chaos and disorder. The theoretical and practical results obtained within the framework of the project show that the highest concentration of entropy is registered among the students, which confirms their special “language competence” as a result of the communicative process. In addition, in the course of the experiment, the participants revealed obvious violations of the lexical, semantic-syntactic and stylistic norms of the Russian language both on the part of specialists and on the part of the bachelor, master, doctoral students. As a result of processing the material, there has been such a classification of speech mistakes according to linguistic aspects and levels:
- lexical;
- syntactic;
- morphological;
- phonological.

At the same time, it should be emphasized that the following quantitative parameters were taken into account when calculating the entropy level of the text:
- the volume of the analyzed text at the sentence level. The sentence is transformed into a statement, while acquiring logical-semantic meanings;
- the number of language failures (actual mistakes and communication defects);
- the number of language mistakes that cause a fluctuation that characterizes any fluctuation or any periodic change.

Thus, the experimental data show that, based on the synergetic information theory for the analysis of speech diagnostics, it becomes possible to analyze the communication features of oral interaction participants, their competence, their ability to competently build communication, as well as to predict the consequences of speech interaction and to anticipate the possibility of communicative successes and failures.

In addition, the developed computer program can be used as an elective course in the educational process based on the interdisciplinary relationship of present day digital technologies with computational linguistics, mathematical and computer modeling, information theory in training bachelor, master, doctoral students as future specialists.

In the authors’ opinion, using the analysis of oral speech, it is possible to determine the future business competence and professional suitability of specialists who work primarily in the educational environment. The effectiveness of this study will depend on the course of the experiments and the quality of the entropy analysis of the texts under study.

The work is funded by the Science Committee of the Ministry of Education and Science of the republic of Kazakhstan (grant No. AP08856918).

REFERENCES: