Int. J. Metrol. Qual. Eng. **8**, 6 (2017)

DOI: 10.1051/ijmqe/2017001

International Journal of Metrology and Quality Engineering

Available online at:
www.metrology-journal.org

RESEARCH ARTICLE OPEN ACCESS

Evaluation of uncertainty in the measurement of sense of natural language constructions

Oleg V. Bisikalo[1] and Oleksandr M. Vasilevskyi[2,*]

[1] Dean of the Faculty of Computer Systems and Automation, Vinnytsya National Technical University, 95 Khmelnitskoye Shose, Vinnitsya 21021, Ukraine
[2] Department of Metrology and Industrial Automation, Vinnytsya National Technical University, 95 Khmelnitskoye Shose, Vinnitsya 21021, Ukraine

Received: 4 March 2016 / Accepted: 4 January 2017

Abstract. The task of evaluating uncertainty in the measurement of sense in natural language constructions (NLCs) was researched through formalization of the notions of the language image, formalization of artificial cognitive systems (ACSs) and the formalization of units of meaning. The method for measuring the sense of natural language constructions incorporated fuzzy relations of meaning, which ensures that information about the links between lemmas of the text is taken into account, permitting the evaluation of two types of measurement uncertainty of sense characteristics. Using developed applications programs, experiments were conducted to investigate the proposed method to tackle the identification of informative characteristics of text. The experiments resulted in dependencies of parameters being obtained in order to utilise the Pareto distribution law to define relations between lemmas, analysis of which permits the identification of exponents of an average number of connections of the language image as the most informative characteristics of text.

Keywords: sense / uncertainty / text / natural language constructions / artificial cognitive systems / language image / lemma

1 Introduction

The complexity of the tasks of semantic analysis of text information is considered to be one of the main barriers to building artificial intelligence in general, and to resolving with appropriate levels of quality a considerable range of problems relating to computer linguistics in particular. Ontogeny is intrinsic to how a person learns and acquires new knowledge all their life, therefore each natural intelligence is a unique and dynamic phenomenon capable of improving and embodying a good understanding of their own kind. Therefore, construction of linguistic knowledge bases should be based on such principles, and the problems in obtaining new formal methods of semantic analysis of natural language constructions, based upon knowledge bases, are quintessential. Formal approaches to the study of artificial cognitive systems need to be determined. Such systems should be able to simulate human activity in the processes of understanding, refining meaning, and the effective use of input text information.

In [1,2], it was proposed and justified that the introduction of a measurement unit of imaginative sense 1 with syntactic associative weighting (SAW) to solve problems of computer linguistics related to the creative thinking of humans. But in the process of such modelling is necessary to take into account the dynamic nature and subjective cognitive ontogenesis, including speech activity. Formally, this can be done in various ways, one of which is to assess the uncertainty of the measurement result of the sense of separate natural language constructions (NLC), the texts, and artificial cognitive systems (ACS) in general, at a given time. It is known [3] that the uncertainty of measurement is a parameter associated with measurement results, characterized by the dispersion of values that can be quite reasonably attributed to the measured value. But it is important that the value that is directly used to express uncertainty should be internally consistent, directly derived from that components that comprise it, and should not be dependent on the grouping of these components and their subdivision into sub-components [4]. In source references known to us, which consider standard uncertainty of measurement types A and B, the concept of uncertainty was not applied as well as the basic requirements needed to solve problems of semantic text analysis.

The subject chosen to be studied is the process of building knowledge bases for linguistic cognitive systems, with the focus of the research on assessment of the uncertainty of sense of NLC formal characteristics. The

* Corresponding author: o.vasilevskyi@gmail.com

purpose is to obtain values of measurement uncertainty of the sense of NLCs, as components of an ACS. To achieve this goal it is necessary to formally define the concept of an ACS, justify the method used to measure NLC sense based on fuzzy relationships, and obtain and interpret formal assessment of the uncertainty of the measurement results of the sense of the NLC.

2 Formulation of the problem

On entering any system S_i with known quantities nt, a flow $X = \{x_1, x_2, ...\}$ as at time t_L may be defined by a Berge graph $G_Q(V, E)$ with a corresponding adjacency matrix A_Q with dimensions $L \times L$. We also know that in a sparse matrix A_Q the number of non-zero $lj - x$ elements equals m and each of them acquires the value k_{lj}. It is necessary to obtain values for the uncertainty σ of the results of observations k_{lj} of each system S_i and to calculate the standard uncertainty of type A – $u_A(X)$ and type B – $u_B(X)$ for all systems. Given the purpose of the study it is necessary to interpret and analyze the formal results in terms of the domain of computer linguistics.

3 Literature review

Consider the fundamental requirements for the notion of uncertainty of measurement as set out in [4,5]. The ideal method for determining the uncertainty of measurement results should be universal, suitable for all kinds of measurements and for all types of input data used in the measurements. The internal consistency of the values directly used to express uncertainty, allows the direct use of uncertainty of one result as a component to determine the uncertainty of another component, which uses the first result.

The uncertainty of the measurement result generally consists of several components, which can be grouped into two categories, depending on the method of evaluation of their numerical value: type A components that are evaluated by statistical methods, and type B components measured by other methods. Each detailed statement of uncertainty must include the full list of components and each of them show the method used in the preparation of each numerical value.

The components of category A are generally characterized by their estimated variances S_i^2 (or their estimated "standard deviations" S_i) and a number of degrees of freedom. If necessary, their covariance should be indicated. Components of category B should be characterized by values U_j^2, which can be regarded as approximations to the corresponding variances, the existence of which is allowed. U_j^2 values can be viewed as variances and U_j as a standard deviation. If necessary, the covariance should be treated similarly.

The combined uncertainty should be characterized by a numerical value obtained when applying the usual method for mapping variances. The combined uncertainty and its components should be expressed in the form of "standard deviations". If in some cases the total uncertainty is obtained when the combined uncertainty is multiplied by a coefficient, then that factor should always be specified. In general terms, the word uncertainty means doubt, and thus, in the broadest sense "uncertainty of measurement" means doubt in the veracity of uncertainty measuring.

Consequently, the uncertainty of the measurement result does not necessarily show the probability that the measurement result is close to the value of the measured value; it appears only as evaluations of the proximity of a measurement result to the best value that corresponds to the currently available information. The introduction of the concept of the "uncertainty of measurement" is a necessary measure to obtain uniform and simplified assessment of the reliability of the evaluation of measuring authenticity, since its definition is based on obtained measurement results, known conditions of the measurement, and the characteristics of the equipment, and not on the unknown actual value of a measured value [6].

To evaluate the input variable X_i that was not obtained as a result of repeated observations, the estimated variance $u^2(\mathrm{x}_i)$ and the standard uncertainty $u(\mathrm{x}_i)$ associated with it must be determined based on a scientific judgment that relies on all available information about possible variability X_i. That is, the type B standard uncertainty is obtained from the presupposed function of the density probability that is based on a degree of confidence that the event will happen (this probability is often called subjective probability).

Since information that enables the evaluation of measurement uncertainty can comprise the data of previous measurements discussed in [2], our approach enables a measurement process of the NLC sense based on fuzzy measures. Thus [1], the fuzzy binary relationship, set on the same base population of language images (or universe) I, is defined as the fuzzy ratio

$$Q = \{ \langle i_l, i_j \rangle, \mu_Q(\langle i_l, i_j \rangle)\}, \tag{1}$$

where $\mu_Q(\langle i_l, i_j\rangle)$ is the function of dependency of the binary fuzzy ratio, defined as the representation μ_Q: $I \times I \to [0, 1]$. In the expression (1), a sequence of two elements is defined through $\langle i_l, i_j\rangle$, where $i_l \in I$, $i_j \in I$. If the carrier Q_s of the fuzzy relationship Q is finite, then the power of this fuzzy ratio is numerically equal to the number of sequences of its carrier and is defined as $card(Q_s)$.

If binary fuzzy relation (1) is a basic cognitive feature of the ACS, then the functional dependency $\mu_Q(\langle i_l, i_j\rangle)$ should be considered as a natural numerical measure of sense. The value $\mu_Q(\langle i_l, i_j\rangle) = 1$, according to [1], is given the sense value of one SAW unit. In general, the function of the dependency of the fuzzy ration of the sense for a pair of language images at the basic level is defined as:

$$\mu_Q(\langle i_l, i_j \rangle) = f(k_{lj}, t_L), \tag{2}$$

where k_{lj} is the number of fixed ACS connections between the l^{th} and the m^{th} images at the moment of time t_L. The value of k_{lj} is not difficult to obtain, by calculating the number of fixed ACS sequences $\langle i_l, i_j\rangle$, based on the technological capabilities of modern linguistics software packages, which allow, for the first time, the application and justification of the concept of measurement uncertainty of the NLC sense.

4 Materials and methods

4.1 The concept of artificial cognitive systems: formalization and interpretation

Let us consider a system S which henceforth will be called an ACS, Artificial Cognitive System, from the point of view of the process accumulating its knowledge base. Let S have the ability to identify images of infinite population $I=\{i_1, i_2, ..., i_{nt}, ...\}$ and perceive associative links between pairs of images as elements of the population $\omega \in \Omega$, where $\Omega \subseteq I \times I$, space ordered pairs. To determine an image construction, we will apply the notion F – sigma algebra (σ-algebra) of subsets of Ω. Further assume that this subset $\gamma \subseteq \Omega$ is a language construction that has the property $\gamma \in \mathrm{F}$. In accordance with the properties of σ-algebra [7] the populations A, $\mathrm{B} \in \mathrm{F}$, the combination, overlapping and difference between A and B in the theoretical-population sense, also belongs to F.

Suppose that the system S communicates information with the outside world as a black box exclusively as language constructions, of which we differentiate a sequence of incoming events $X=\{x_1, x_2, ...\}$ and a set of image responses of the system $Y=\{y_1, y_2, ...\}$, where $x_i \in \mathrm{F}$, $y_i \in \mathrm{F}$. Figure 1 shows a diagram of an abstract model of cognitive activity, which includes an external "black box" and internal ACS, which receives as an input a continuously set of images of events in the form of an X stream. The ACS output images appear as Y, which is a response of this system to the external situation X according to the modelling approach to human image thinking [2].

Farther will now use the *Ontogenetic Principle* to build an ACS. The cognitive resource Ω of the system S, which determines the sense of its functionality, can be obtained exclusively through successive accumulation of sequential parameters ω from an external "black box" and further self-improvement of the set Ω. Formally, the ontogenetic principle is reflected in the fact that the knowledge base system S is built with $C = \bigcup_{i=1}^{m'} x_i$, where m' is overall number of input image constructions accepted by the system at a given time.

In order to solve applied problems of computer linguistics, let us interpret the components of a derived abstract model of cognitive activity. For an ACS linguistic construct, we will consider image i to be a language image that is approximately defined by a lexeme or a word form [8]. Then the analogous association between pairs of images ω is a phrase, and the image construction γ is a sentence or an utterance – in general an NLC. Accumulated ACS cognitive resources Ω are shown as a processed set of texts, and the result is the building of a linguistic knowledge base C.

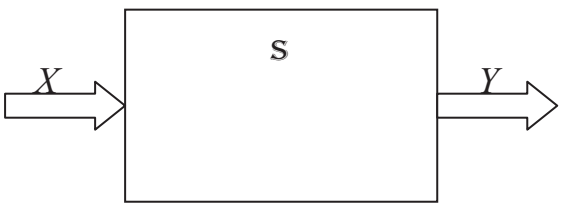

Fig. 1. Diagram of an abstract model of cognitive activity.

Unlike the existing models of knowledge in computer linguistics, where the vocabulary of word forms is combined with a multitude of morphological, syntactic and semantic rules, in our case the basis for the knowledge base C is formed exclusively with associative knowledge about the combinability of language images i. This gives grounds for unified evaluation of the unit of sense and the quantity of sense of the NLC.

4.2 Measurement method for NLC sense based on fuzzy relationship

Under the proposed approach [9] we will detail the dependency function that generates a binary fuzzy relationship of sense (1) for the following 3 successive levels, built on the basic level (2):

1. The level of probabilistic forecasting – to standardise the dependency functions in the range [0, 1] provide for the calculation of the statistical evaluation λ (mathematical expectation), if known for nt for the given ACS at the time t_L image $k_\Sigma = \sum_{l=1}^{nt}\sum_{j=1}^{nt} k_{lj}$, and m is the number of all non-zero sequences $\langle i_l, i_j \rangle$, then $\lambda = k_\Sigma / m$ where in this case we apply the known sigmoid function [10]

$$\mu_Q(\langle i_l, i_j \rangle) = f_1(k_{lj}, \lambda) = 1/(1+e^{-k_{lj}+\lambda}). \qquad (3)$$

As a result of the standardisation there appears a characteristic property of the dependency function which is obtained by the proposed method with average value $\overline{\mu_Q} = \frac{1}{m}\sum_{j=1}^{m} \mu_{Qj} = 0.5$.

2. The level of incorporation of emotional state. Introduce the opportunity to incorporate a binary model of emotion for the ACS [9] with the help of the indicator $\mu = \{..., -2, -1, 1, 2, ...\}$, where

$$\mu_Q(\langle i_l, i_j \rangle) = f_2(k_{lj}, \lambda, \mu) = 1/(1+e^{-\frac{k_{lj}-\lambda}{|\mu|}}). \qquad (4)$$

In the case of $\mu = -1 \vee 1$, emotions do not affect sense in the functioning of the ACS, and the dependency function (4) regresses to the function (3). The increase in the indicator μ symmetrically smoothes the sigmoid function as shown in Figure 2.

3. The level of incorporation of motivation components based on image centre of needs. It is proposed that the consideration of the image centres of needs j' be undertaken as a model of ACS motive at a given time t_L, as well as calculating the variance and mean-square differentiability of the results of observations k_{lj} as

$$D = \frac{1}{m}\sum_{l=1}^{nt}\sum_{j=1}^{nt}(k_{lj}-\lambda)^2 | k_{lj} > 0, \quad \sigma = \sqrt{D}. \qquad (5)$$

The obtained value σ will now be considered as the uncertainty that is conditional on the imprecision of the ACS motive model. The uncertainty is characterized in particular by the imperfection of basic dependency (3), on

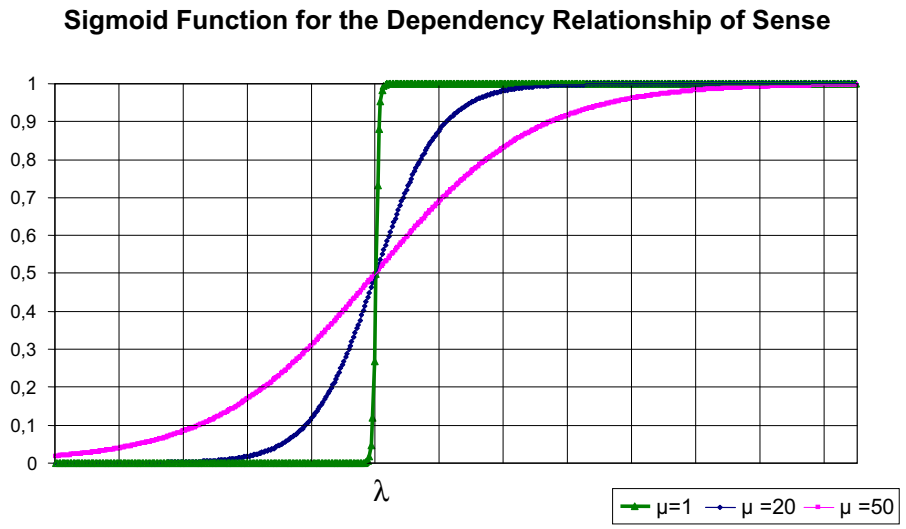

Fig. 2. Impact of indicator μ on dependency function (4).

the basis of which it is proposed to take into account the motivational component based on the image centres of needs.

Depending on the degree of approximation r to the pair of images $\langle i_l, i_j\rangle$, function (4) can shift to the left along the x-axis by reducing the mathematical expectation for the pair $\lambda_{lj} = \lambda - r\cdot\sigma$, where $r = \{0, 1, 2, 3\}$ which results in:

$$\mu_Q(\langle i_l, i_j\rangle) = f_3(k_{lj}, \lambda_{lj}, \sigma, \mu, i') = 1/(1 + e^{-\frac{k_{lj}-\lambda_{lj}}{|\mu|}}). \quad (6)$$

The issue of constructing a separate algorithm to determine the degree of proximity r of the pair $\langle i_l, i_j\rangle$ to the image-needs j' and the introduction of additional level of consideration of reflexes and results of the external tuition is considered in [9]. Note that, unlike (3) and (4), the dependency function related to sense (6) resulting from local shifts in mathematical expectation, the property $\overline{\mu_Q} = 0.5$ disappears. The authors consider this to be evidence of proper formal interpretation of the known facts of psychology and physiology on contradictions between generally accepted (statistically average) sense and actions influenced by strong motives.

4.3 Uncertainty of measurement results of NLC sense

The approach to the measurement of sense corresponds to the linguistic knowledge base of one ACS, the output data of which can be either separate text or a unique set of texts. It should be understood that every text reflects a unique worldview of an author, depicted in their language. To solve the problem of identifying informative text attributes it is important to define the reliability of the knowledge base in general and the meaning of a pair of images $\mu_Q(\langle i_l, i_j\rangle)$ as a basic component of the knowledge base in particular. In as much as this actually refers to the measurement of sense, it is proposed that in order to assess reliability will apply the concept of uncertainty of results of multiple measurements of NLC sense.

In the first approximation, assume that a subjective estimate of the amount of sense of one pair of language images is embodied in a number of statistical arrays of numerical values N for different ACSs. Thus, for an arbitrary sequence $\langle i_l, i_j\rangle$ the value $Y = \mu_Q(\langle i_l, i_j\rangle)$ as measured according to (3), is functionally dependent on the results of repeated measurements X_1, X_2, ..., X_N for different ACSs and, in general, is as follows:

$$Y = f(X_1, X_2, ..., X_N). \quad (7)$$

The evaluation of the measured value Y indicated henceforth as y, is obtained from the general equation (7) using input values x_1, x_2, ..., x_N for N numerical values X_1, X_2, ..., X_N. Thus, the output assessment y, which is the result of a measurement, is expressed as follows:

$$y = f(x_1, x_2, ..., x_N).$$

The baseline assessment of mathematical expectation or expected value μ_Q of value q, that is randomly changing, is the arithmetic mean or average value $\overline{q}$ of n observations

$$\overline{q} = \frac{1}{n}\sum_{k=1}^{n} q_k. \quad (8)$$

The experimental standard deviation characterizing the variability values of q_k, or more specifically, their dispersion σ^2 about the mean values $\overline{q}$ is calculated by formula [6]

$$u_{\mathrm{A}}(q_k) = \sqrt{\frac{\sum_{k=1}^{n}(q_k - \overline{q})^2}{n-1}}. \quad (9)$$

As the average value $\overline{q}$ is taken as the result of multiple measurements, it is important to determine the dispersion. The best estimate $\sigma^2(\overline{q}) = \sigma^2/n$ of the dispersion of the

mean value $u_{\mathrm{A}}^2(\overline{q})$ may be expressed as:

$$u_{\mathrm{A}}^2(\overline{q}) = \frac{u_{\mathrm{A}}^2(q_k)}{n}. \tag{10}$$

Experimental dispersion average $u_{\mathrm{A}}^2(\overline{q})$ and the experimental standard deviation of the mean value $u_{\mathrm{A}}(\overline{q})$, equal to the positive square root of the dispersion value $u_{\mathrm{A}}^2(\overline{q})$, quantitatively determine how well $\overline{q}$ determines the expectations μ_k of the value q. Given the expressions (9) and (10) the experimental standard deviation of the average value $u_{\mathrm{A}}(\overline{q})$ is calculated by formula [6]

$$u_{\mathrm{A}}(\overline{q}) = \sqrt{\frac{\sum_{k=1}^{n}(q_k - \overline{q})^2}{n(n-1)}}. \tag{11}$$

For a deeper consideration of the subjective nature of the measured sense of the sequences in function (7) applied components of standard uncertainty type B, which are usually determined on the basis of information on the upper and lower boundaries $[\alpha_-; \alpha_+]$ predictable (specified a priori) of the distribution law or with interval U, which has given a given confidence level p.

To determine the type B standard uncertainty, need to take the positive square root of the product of the confidence level of each value and the square of the deviation of this value and all products of this type should be added. As a result, a general view of the formula for calculating standard uncertainty of type B in the case of discrete data is of the form:

$$u_{\mathrm{B}}(X) = \sqrt{\sum_{i=1}^{n}\left(x_i - \sum_{i=1}^{n} x_i p_i\right)^2 p_i} = \sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 p_i}. \tag{12}$$

As we can determine the upper and lower limits $[\alpha_-; \alpha_+]$ for value X_i, then the type B standard uncertainty in assumptions about the possible shape of the distribution law can be determined by formulas [4–6]

(a) for the triangular distribution law

$$u_{\mathrm{B}}(X_i) = \frac{\alpha_+ - \alpha_-}{\sqrt{24}}; \tag{13}$$

(b) for the exponential distribution law

$$u_{\mathrm{B}}(X_i) = \sqrt{\frac{(\alpha_+ - x)(x - \alpha_-) - (\alpha_+ - 2x + \alpha_-)}{\lambda}}, \tag{14}$$

where x is the expected value, and λ is the distribution parameter;

(c) for the Pareto distribution law

$$u_{\mathrm{B}}(X_i) = \frac{x_m}{k-1}\sqrt{\frac{k}{k-2}}, \tag{15}$$

where x_m is the initial value, and k the distribution parameter (the density for x_m);

(d) for the uniform distribution law

$$u_{\mathrm{B}}(X_i) = \frac{\alpha_+ - \alpha_-}{\sqrt{12}}. \tag{16}$$

For given intervals U_p with a known level of confidence p where the standard distribution law is assumed, the type B uncertainty is given by the formula:

$$u_{\mathrm{B}}(X_i) = \frac{U_p}{k_p},$$

where k_p is the coverage coefficient, which for the standard distribution law is equal to 1.64; 1.96; 2.58 and 3 for confidence levels 0.9; 0.95; 0.99 and 0.9973 [11].

In the absence of information about the usability of laws (13)–(16) for the distribution of the input value X_i for symmetrical boundaries $\pm\alpha_i$, standard uncertainty of type B is determined by the formula:

$$u_{\mathrm{B}}(X_i) = \frac{2\alpha_i}{\sqrt{12}} = \frac{\alpha_i}{\sqrt{3}}, \tag{17}$$

which can be applied at an early stage of experimental research into the ACS.

5 Experiments

The leading linguistic package DKPro Core, which is based on the platform of Apache UIMA framework [12], was used in order to verify by experiment the results of the evaluation of measurement uncertainty of the NLC sense as a component of ACSs, using the proposed method. To implement this series of experiments an additional Java application program was developed, which not only uses but also improves the collection of software components to process natural language by DKPro Core [13]. A feature of the program as developed that focuses on Java/Maven/Eclipse technology, is the definition of the list of the Lemmas of a text and their complex dependencies, as described in [14], between these lemmas as a list of m links.

As an experimental basis, three famous open source literary works from the Project Gutenberg [15], were selected, namely English copyright versions of 4 texts of different volumes: “Alice in Wonderland” (Lewis Carroll – one excerpt of 4204 words, and a second, being the full version of 26690 words). The third text was “White Fang” by Jack London comprising 48907 words, and the fourth being “Three Men in a Boat (To Say Nothing of the Dog)” by Jerome K. Jerome of 67328 words. The purpose of the series of experiments was to study basic characteristics of uncertainty of each of the 4 texts, and to obtain values of uncertainty of the set of pairs of language images $\langle i_l, i_j\rangle$ common to all four texts, according to the proposed method.

6 Results

The results of the research formalized and interpreted for the subject area of computer linguistics the notion of artificial cognitive systems, incorporating the basic

Table 1. Principal results of processing the 4 English-language texts.

Title of text		m	k_Σ	λ	σ	%	Number of lemmas	Mean number of links
1	*Carroll_part*	2360	2812	1.19	0.78	65.4%	762	3.0971
2	*Carroll_full*	12,156	17,786	1.46	2.25	153.5%	2121	5.7313
3	*London*	25,244	31,234	1.24	1.26	101.8%	5702	4.4272
4	*Jerome*	33,316	47,091	1.41	2.04	144.7%	6048	5.5086

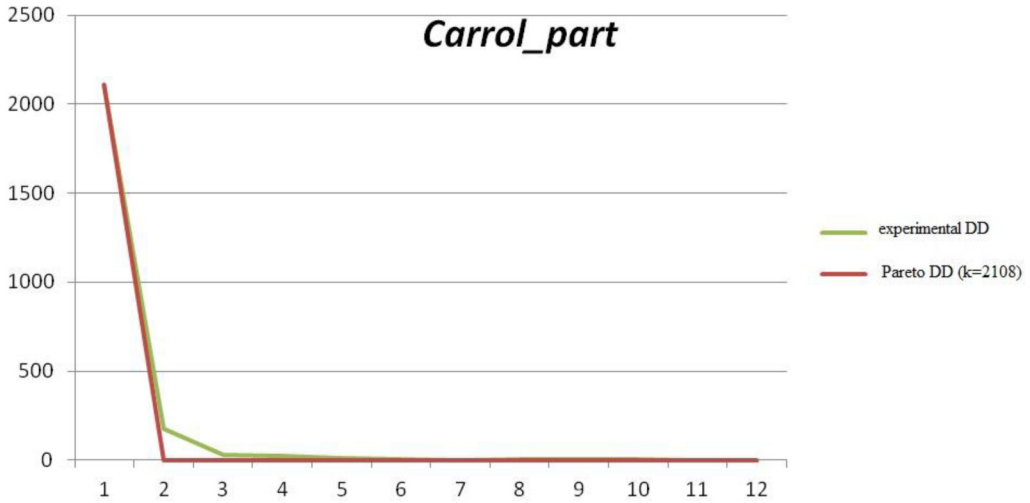

Fig. 3. Analysis of the experimental distribution density (DD) law for text 1.

ontogenetic principle of constructing an ACS. Formal characteristics of the method of creating binary fuzzy relationships of the image sense Q of the ACS S_Q were obtained by modelling the notions of motivational goals and emotional state. The principles of successive multilevel construction of the dependency function $\mu_Q(\langle i_l, i_j\rangle)$ that generate a fuzzy relationship Q were proposed, and a characteristic feature $\overline{\mu_Q} = 0.5$ of the method of measuring the NLC sense was defined.

In accordance with this, the task of identifying informative features of the text resulted in formal theoretical values of uncertainty σ of the results of observations k_{lj} for each ACS S_i were obtained, in addition to calculation of standard uncertainty of type A $u_\mathrm{A}(X)$ and type B $u_\mathrm{B}(X)$ for all ACSs.

With the help of the DKPro Core-based package, the software program developed in [13] produced results by processing the four chosen English texts, which may be interpreted as being four different ACSs. The basic results of processing as defined in (5) are presented in Table 1, where the last three columns contain the following data:

– percentage σ of the mean square deviation of the mathematical expectation λ;
– the number of lemmas in the text identified by DKPro Core;
– the mean number of different links for one lemma in the text.

The resulting histogram of experimental density distribution laws showed a significant resemblance to a Pareto distribution law, which is shown by the example of a comparison of experimental results for text 1 (*Carroll_part*) with the theoretical Pareto density distribution with a value parameter $k = 2108$ (Fig. 3).

Analysis of the language-pair images $\langle i_l, i_j\rangle$, sorted as a descending list k_{lj}, revealed four common pairs at the top of the list, output data and assessment results $\overline{q}$ according to (8) and uncertainty of types A and B, in accordance with (11) and (12), are presented in Table 2.

7 Discussion

The results obtained from the experiments of numerical values of uncertainty of the measurement results of the sense of language-pair images yielded new information about the texts analyzed. Presentation of each text as a separate ACS shows that the experimental density distribution law for the characteristics k_{lj} of the pairs of language images is very similar to Pareto distribution. However, this conclusion does not correspond to the mathematical expectation values λ, which should have been diminishing and moving closer to 1 ($\lambda_{Pareto} = (k \cdot x_m)/(k-1)$) with an increase on the number of pairs [16], as well as the mean square deviation σ which is too large for a Pareto distribution. For example, text 1 according to (5),

Table 2. Results of uncertainty assessment of the 4 selected language-pair images.

Title of texts, assessment parameters		*go-back*	*say-I*	*know-I*	*see-I*
1	*Carrol_part*	0.8591774	0.9431321	0.9431321	0.9431321
2	*Carrol_full*	0.9985531	0.9999995	0.3862398	0.9985531
3	*London*	0.9999999	1.0000000	1.0000000	0.9999999
4	*Jerom*	0.9999995	1.0000000	1.0000000	1.0000000
	$\overline{q}$	0.9288652	0.9715658	0.6646859	0.9708426
	$u_{\mathrm{A}}(X)$	0.0284499	0.0116080	0.1136752	0.0113128
	$u_{\mathrm{B}}(X)$	0.036641	0.0149501	0.1464039	0.0145699

$\sigma = 0.7788$, which represents 65.36% of λ. Similar values in accordance with dependencies (15) and the Pareto distribution (17) for the general case of small value $\alpha_i = \pm 0.01$: $\sigma_1 = 0.0004748$ (0.04%) and $\sigma_2 = 0.58$ (0.48%).

However, analysis of the data in Table 1 provides a formal basis for advancing the hypothesis – the most informative characteristics of an ACS lie in the average number of links for a single Lemma (language image). The justification for this is the Pearson correlation coefficient for columns containing λ and the 'number of Lemma' for all 4 ACSs which equals 0.198, but pairs of columns λ and the 'average number of connections' equalling 0.945.

Simultaneously for pairs of columns σ and "average number of links", the correlation coefficient is 0.984, and pairs of columns, namely "%" and "average number of links" equals 0.996. This suggests that the distribution law is only Pareto-like, but the uncertainty of the sense of ACSs (parameter σ) is directly proportional to the mean number of links. Further advancement of the hypothesis requires further large-scale experimental verification and clarification.

The data in Table 2 shows a high degree of sense similarity in accordance with the approach put forward for 4 selected pairs of language images, which is used by 3 different authors. The general trend is that the values of type A uncertainty $u_{\mathrm{A}}(X)$ are lower than the corresponding type B values $u_{\mathrm{B}}(X)$ for all ACSs by approximately 1.5. At the same time, the percentage of uncertainty does not exceed 4% of the value of the mathematical expectation $\overline{q}$ for all pairs $\mu_Q(\langle i_l, i_j \rangle)$, other than the pair "know-I" (up to 22.03%), which has an understandable explanation, given the selected excerpt 1 in the text by Lewis Carroll, this pair being found relatively more often than in the whole book 2 (Alice in Wonderland) in general. These results allow us to hope that the proposed approach will improve the quality of problem-solving in automatic semantic analysis of texts, in particular, the identification of authors. However, it is likely that a similar comparison of pairs which are at the bottom of sorted lists which are rarely found, may demonstrate high uncertainty.

Further research is also required to define the laws of the distribution of experimental values $\mu_Q(\langle i_l, i_j \rangle)$ and to obtain subjective characteristics for an ACS knowledge base to enable dynamic uncertainty measurement.

8 Conclusion

The research resulted in solving the task of obtaining values of the uncertainty of sense for NLCs as components of ACSs, which is directly related to the problem of understanding the sense of textual information. Further, a method for measuring sense in an NLC was further developed based on fuzzy relationships, which, unlike the existing methods, is based on two formal terms of artificial cognitive systems and linguistic image that enables output statistical data to be obtained, in order to evaluate the results of uncertainty measurement of types A and B. For the first time we obtained and interpreted formal values of the uncertainty of measurement results of the sense of NLCs that enable us to take into account information on links between lemmas of a text to solve the tasks of identifying informative features of a text.

The practical significance of the results is to obtain software technology to produce tools based on the DKPro Core linguistics package, which allows us to implement our proposed method for semantic analysis of English-language texts. The results of a series of experiments revealed that the distribution law links between lemmas of a text is Pareto-like, but has significant differences from a formal and classical Pareto distribution, including significantly higher values of mathematical expectation λ (up to 46.3%) and mean square deviation σ (by several orders).

In terms of this proposed approach to determining the sense of NLCs, the augmentation of the size of a text by number of words and, accordingly, the size of its vocabulary by the number of lemmas does not affect the parameters of the distribution law and uncertainty of the sense of each individual ACS. Analysis of the results obtained suggests that the parameter of the average number of links of a language image be considered as the most informative characteristic of the text, since the Pearson correlation coefficient between it and the parameters related to the uncertainty of the sense is greater than 0.945.

Comparison of uncertainty values for 4 pairs of language images, used by 3 different authors, showed a high degree of similarity in the sense of such pairs according to the approach put forward. This type A uncertainty

values $u_A(X)$ are proportionally lower than the corresponding type B values $u_B(X)$ for all ACSs by about 1.5 times, which allows us to only obtain a single value $u_A(X)$ for uncertainty.

The results of research that were obtained were, among others, formal parameters for the uncertainty of sense and the average number of links of language pairs, which provide potential improvement in resolving the tasks in semantic analysis of NLCs, including clustering, classification and definition of authorship of texts.

Nomenclature

$u_A(X)$	evaluation of type A uncertainty
$u_B(X)$	evaluation of type B uncertainty
$\lambda, \overline{q}$	mathematical expectation
σ	Standard quadratic deviation (SCR)
NLC	natural language construction
ACS	artificial cognitive system

References

1. O.V. Bisikalo, S. Cieszczyk, G. Yussupova, Solving problems on base of concepts formalization of language image and figurative meaning of the natural-language constructs, in *Proc. SPIE 9816, Optical Fibers and Their Applications 2015, December 18, 2015* (2015), 98161U, doi:10.1117/12.2229046
2. O.V. Bisikalo, I.A. Kravchuk, Methods of obtaining knowledge from natural language texts, in *Perspektywiczne opracowania są nauką i technikami – 2012: Materiały VIII Miedzynarodowej naukowi-praktycznej konferencji, 07–15.11.2012*, Vol. 19, Przemysl (2012)
3. O.M. Vasilevskyi, Calibration method to assess the accuracy of measurement devices using the theory of uncertainty, Int. J. Metrol. Qual. Eng. **5** (4), 403 (2014)
4. Evaluation of measurement data – Guide to the expression of uncertainty in measurement: JCGM 100:2008, Sevres: JCGM, 2008, 120 p.
5. ISO/IEC Guide 98-1:2009, *Uncertainty of measurement – Part 1: Introduction to the expression of uncertainty in measurement* (ISO, Geneva, Switzerland, 2009), 32 p.
6. O.M. Vasilevskyi, A frequency method for dynamic uncertainty evaluation of measurement during modes of dynamic operation, Int. J. Metrol. Qual. Eng. **6** (2), 202 (2015)
7. A. Gut, *Probability: a graduate course (Springer Texts in Statistics)* (Springer-Verlag, 2005), 603 p.
8. R.N. Kvetny, O.V. Bisikalo, O.I. Osmolovsky, I.A. Kravchuk, Morphological analysis of input information in intelligent robotic systems, in *Aviation in the XXI-st Century: Proceedings of the fifth world congress, September 25–27, 2012* (2012), Vol. 1, pp. 1.9.54–1.9.56
9. O. Bisikalo, A. Yarovenko, I. Kravchuk, I. Nazarov, Search method based on figurative indexation of Folksonomic features of graphic files, TEM J. **2** (4), 297–304 (2013)
10. H.-J. Zimmermann, *Fuzzy set theory – and its applications* (Kluwer, 2001), 4th ed., 519 p.
11. ISO/IEC 17025:2005, *General requirements for the competence of testing and calibration laboratories* (ISO, Geneva, Switzerland, 2005), 28 p.
12. I. Gurevych, M. Muhlhauser, Ch. Muller, J. Steimle, M. Weimer, T. Zesch, *Darmstadt knowledge processing repository based on UIMA [Electronic resource]*, February 9, 2007 (2007), Available from: https://www.ukp.tudarmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2007/gldv-uima-ukp.pdf
13. O. Bisikalo, I. Kravchuk, Automation of course content construction, *SWorld: Scientific research and their practical application. Modern state and ways of development 2013, 1–12 October 2013* (2013), Available from: http://www.sworld.com.ua/index.php/ru/technical-sciences-313/informatics-computer-science-and-automation-313/19442-313-0895
14. Stanford dependencies, Universal dependencies, The Stanford NLP Group, Available from: http://nlp.stanford.edu/software/stanford-dependencies.shtml
15. Free ebooks, Project Gutenberg, Project Gutenberg Literary Archive Foundation, Available from: https://www.gutenberg.org/
16. O. Bisikalo, I. Kravchuk, *Formalization of semantic network of image constructions in electronic content* (Cornell University Library (Computer Science, Computation and Language), 2011), arXiv:1201.1192v1, January 2011, p. 4, Available from: http://arxiv.org/abs/1201.1192v1

Cite this article as: Oleg V. Bisikalo, Oleksandr M. Vasilevskyi, Evaluation of uncertainty in the measurement of sense of natural language constructions, Int. J. Metrol. Qual. Eng. **8**, 6 (2017)