

Text Disambiguation by Educable AI System

Alexander VOSKRESENSKIJ
College MESI, Moscow, Russian Federation

Abstract. The structure of possible text understanding system is discussed. To store concepts and knowledge system uses multilayer ontology based on pragmatic memory model. The way of knowledge searching for system education is described. Main aim of project is development of system for translating text into sign language phrases.

Keywords. Natural language understanding, sign language, concept, memory, ontology, knowledge management, education

Introduction

This paper dwells on the problems related to the machine translation of Russian texts into the Russian sign language (RSL) used by deaf people. The project is now at its early stage implementation and it can be discussed only as a theoretical conception.

While translating from a verbal language into a sign language there comes a problem of disambiguation which in a certain sense different from the problem which we encounter while translating from one spoken language to another. What is unequivocally perceived in a verbal language may have several meanings in a sign language. This difference should be specifically expressed during translation. It is true not only RSL, but also of any other sign language, like Danish, for example [1].

Moreover, the same word in different contexts gets different senses. Therefore, fixed, universal systems of semantic coding are impossible. Certainly, there are words or word combinations whose values are fixed, such as some kinds of terminology (technical, medical, linguistic, etc.) or idioms. But it is frequent that word meanings, especially in informal conversation, are probabilistic. That is, the word does not have a fixed value. Instead, it is the index on some semantic field whose borders can be, in turn, indistinct.

For Nalimov [2], the "... words, on which our culture is based, do not and cannot have an atomistic meaning. It has become possible and even necessary to consider words as possessing fuzzy semantic fields over which the probabilistic distribution function is constructed and to consider people as probabilistic receivers" (p. 56).

Due to facts that any verbal language consists of millions words in contrast to approximately 6 (or maybe 8) thousands of sign language gestures, and sign language grammar differs from verbal language grammar, translating text to signs is possible only by a system which understands input text. The system must to select sign (or chain of signs) representing concept nearer by sense to translated words concept. Thus, text understanding in this case means narration using another set of words.

One of the preconditions of this work is the model of perception and processing of the information [3], partially explaining the probabilistic character of conceptual values. This model tries to find explanations of the distinctions between the cognitive abilities of the deaf and the hearing.

To store information about objects described in text a dynamic multilayer ontology is used which structure is based on said model. The general structure of estimated text understanding system is discussed.

The technique to select concepts with the same (or like) sense was offered which may be useful not only for described task but for searching for texts containing new knowledge for the subject domain also. An experiment has been planned and carried out on documents from the Internet to check the offered approach [4]. As far as we know, this is the first case of the application of design of experiment (DoE) [5] methods in linguistic research.

1. Memory: Pragmatic Approach

In many studies human memory is represented as having short-term memory and long-term memory parts and short-term memory is used for information input and output (see, for example, [6, 7]).

But this model is not effective with pragmatic point of view: we don't know a priori a value of new information so short-term memory can't be used as a filter of useful information entering long-term memory. Moreover, any experiment concerned with presentation of some information to human and subsequent answering interacts with output channel only because processing of information input by human mind is not available for direct experimental studying.

Based on observations presented in [3] possible simplified diagram of memory is shown in Figure 1.

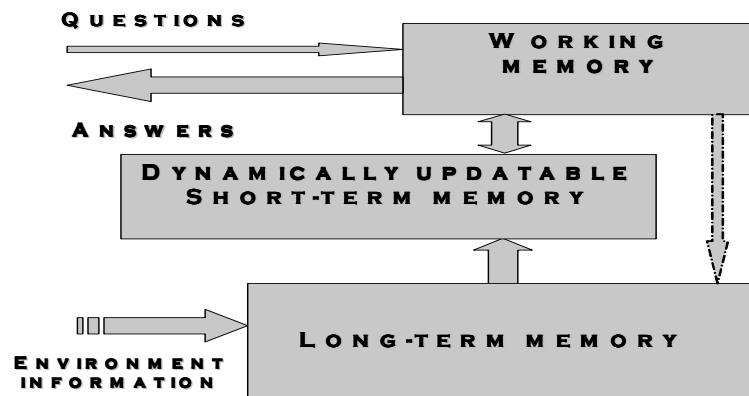


Figure 1. Diagram of possible memory structure

This model of memory may be described in the following way:

- The information from an environment enters directly in long-term memory where it is stored constantly.
- The information is presented by concepts and relationships between them forming knowledge, one of relationship parameters is the level of activation, which rate changes with time in due course from a maximum value down to a minimum value ($R_{\min} > 0$), simulating process of forgetting.
- In case the concept and relations are confirmed by input information, level of these relations activation is raised but no exceeds maximum value ($R \leq R_{\max}$). Levels of activation for tied concepts are rising too (probably, to a lesser degree).
- The knowledge which level of activation exceeds the certain value is come in the short-term memory (which is updated periodically). Because information entering from environment (in general case it may include visual, audio and other kinds of information, for example, books, films and so on) effects activation levels of concepts in long-term memory, content of short-time memory may be changed every cycle of short-memory updating. So content of short-time memory (or consciousness) is probabilistic to some degree.
- Working memory usually cooperates with short-term memory. The reference from working memory in long-term memory is possible during persevering remembrance of forgotten.
- This model allows reproducing processes of "attenuation" and "replacement" of knowledge, "changes of outlook" as a result of receipt of new knowledge.

The parameters of relationship between two concepts include causation, describing one concept as a cause and other as an effect. Because the model allows for concepts (in long-term memory) more than one relationship having different rates of activation level it's possible to change cause and effect (changing direction of relationship between concepts or selecting another pair of concepts) in short-term memory. It's a way for model's "changes of outlook" or domain tuning. So this model uses term logic rather than predicate logic. In this way it is closer to these AI systems as NARS or Novamente [8].

One of education results is new knowledge stored in mind. But some kind of knowledge is more valuable for subject than other; in many cases (but not always) new knowledge is more valuable than old one. If some kind of knowledge is confirmed more often than other, the former is more valuable than latter. Natural regulator of this phenomenon is forgetting. In our model changing of relationship activation level emulates the process of forgetting.

2. Multilayer Ontology

In our system there is a dynamic multilayer ontology proposed as said memory model.

In brief it may be described in the following way (see Figure 2):

- Each concept is represented by hyper graph which nodes are attributes of current realization of concept.
- Rib of hyper graph singles out concrete realization of concept from the general set of knowledge. The rib has numerical parameter of activation, and also

attributes of time, a place and the action which was a reason for the given rib formation.

- The parameter of activation of a rib has maximal value at formation, gradually decreases in due course.
- The top layer realizing a current “picture of the world” is formed by hypergraph, having ribs with activation level which exceeds the certain value.

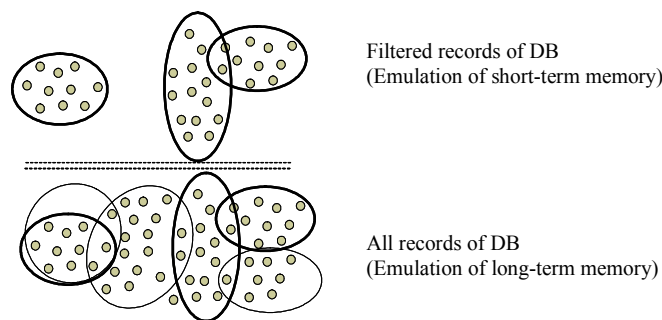


Figure 2. Formation of current “picture of world” by filtering an ontology content.

Realization of such structure is possible by means of a relational DB. The top layer (layers) is virtual and formed by the records of a DB filtered by certain rules.

3. Text Understanding System

On the basis of comparison of processes of verbal and sign languages communications it is shown [9], that the system of understanding of the text, besides the means providing linguistic analysis of the text, should include the block tracing changes of characteristics of objects described in the text (spatial position, the sizes, shape, age, etc.) and storing values of these characteristics in a binding to time of the text and to astronomical time. This way allows creating descriptions of situations at any moments of time.

But description of text objects will be not full without the explanation of their attitudes to each other, for example, friendly or hostile. This information can be absent in the given text, but can be derived from external sources of the information (for the human it can be the knowledge received during training, from the literature or other sources).

Originally language was developed from communication system composed of the vital signals for the human (danger, food, etc.) [10]. So it is possible to assume, that at reading any text the person builds system of the attitudes both to the text as a whole, and to objects described in it (these attitudes can change from negative up to positive, including neutral).

If to add to functions specified above the block which function is to mark objects of the text by various levels of attitudes "good" and "bad", this block will carry out (to some extent) functions of such mental phenomenon as individual "I" (in I.G. Fichte's interpretation).

The marking of objects may be realized on the basis of comparison and analogies to the objects marked at preliminary training of system.

The structure of such system is shown on Figure 3.

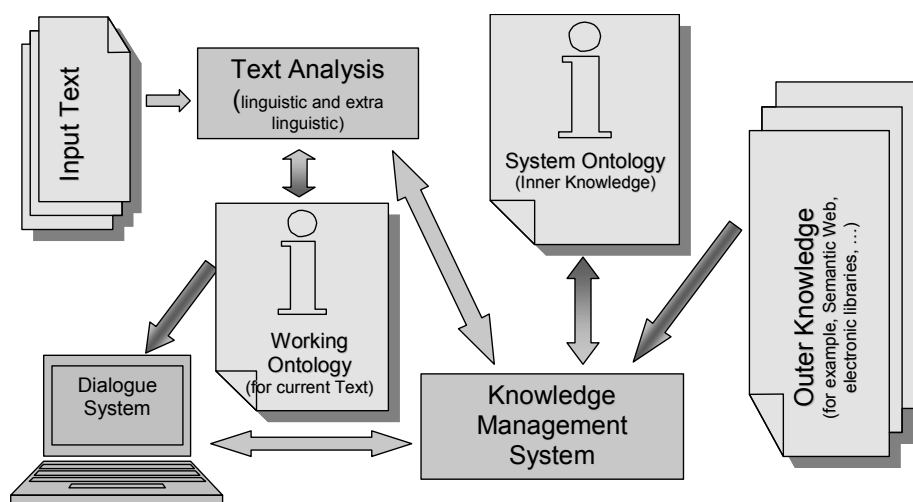


Figure 3. Structure of Text Understanding System

Input text at first is construed using linguistic methods (morphologic and syntactic). Defined objects (main actors, props and locations) with their attributes are loading into working ontology. If it possible objects information are enriched using knowledge from system ontology of computer and/or outer knowledge.

The system uses foregoing multilayer ontology. Such ontology can store the information on changes with time of the objects described in the text (and their moving), moreover, it can be used for short-term forecasting of topic evolution.

Aforesaid ontology includes objects of two types: concepts, containing descriptions of objects, and relationships which connect (group) concepts. Relationships are not attributes of concepts – they are independent objects having own set of attributes which includes references to concepts with which the given relationship cooperates.

Relationship attributes include time of an establishment of the given relationship (for example, astronomical time if relationship concerns to concepts of world around, or time of text if relationship concerns to concepts of a literary work), level of activation (or weight) of relationship and so forth.

Eventually the weight of relationship decreases, that simulates process of forgetting. The minimal weight of relationship is greater than zero, i.e. once the established relationship of the bottom level is never broken off.

At each updating of relationship (simultaneous supervision of the concepts grouped by given relationship) the weight of this relationship increases (but does not exceed the certain maximal value, identical, most likely, for all relationship). Simultaneously under the certain law the weight of each relationship connected with these concepts (in direct connection, and through other concepts) increases.

Interaction of the program agent who is carrying out semantic text processing, occurs only with top level of dynamic multilayer ontology.

The top level of dynamic multilayer ontology includes relationships with weight which exceeds the certain threshold level, and concepts which are tied by these relationships. Updating of top level of dynamic multilayer ontology is made periodically. This process models reception of new knowledge and forgetting old one (maybe like process is a reason for some brain rhythms?).

At inclusion in attributes of relationship the parameter defining a subject domain, fast tuning of ontology top level to the certain subject domain is possible.

Content of ontology top level has probabilistic character, i.e. system perception of the input text rather strongly depends on the information acting from external sources (not including the given input text). Respective alterations of word meanings (semantic fields) are close to the probabilistic model of language described in [2].

4. Education of System

One of definition in Google said that education is the gradual process of acquiring knowledge. If new concepts and/or relationships are contained in input text they will be loaded from ontology of text into bottom layer of computer ontology. Knowledge managing system will try to find new relationships between concepts using deduction rule Eq. (1):

$$\{A \rightarrow B, B \rightarrow C\} \mid A \rightarrow C. \quad (1)$$

Taking into consideration that creating new relationships or updating of existing ones changes activation levels of these and corresponding relationships this process may results to changes in top level of ontology representing learning outcome.

Of course, in the beginning stage when computer ontology is empty, process of self education is impossible. So the preliminary loading of computer ontology is needed. Conditions when self education starts are not known as yet.

For system education may be used outer knowledge sources also.

Ability of AI system for self education is very important, because, for example, we can't teach system all rules of text semantic interpretation: we don't know the whole list of these rules as yet; moreover, it may be changing due to probabilistic nature of language. Ability of AI system constructs these rules using term logic open a way to solve the problem.

5. Search of New Knowledge

To accomplish procedures of self educating the system must have means for access to new knowledge. One of the ways is to search it in Internet. An experiment has been planned and carried out on documents from the Internet to check the offered approach. As far as we know, this is the first case of the application of design of experiment (DoE) methods in linguistic research. Detail description of this experiment was published in Russian [4] and later in English [11]. Here shorten description of this experiment is presented.

Composition of queries, loading these queries into search engine, and results treatment was accomplished 'by hands'. Of course, these procedures may be automated, but in this case it was inefficient.

5.1. Semantic Search Technique

Dekang Lin has stated that: *“Two different words are likely to have similar meanings if they occur in identical local contexts.”* [12]. This statement was used in cited work for a choice of the correct word meaning in the dictionary, using explanatory texts for entries.

However, while translating the text into sign language, use of a local context appears insufficient. For example, the word “fence” is transferred into different signs of RSL depending on whether (inside or outside of the territory surrounded by a fence) a subject is located. This information can be found only in the general context of the text.

Let's expand on D. Lin's statement and formulate it as follows: *Two words or phrases can have a similar value if they co-occur in identical (or similar) contexts.* We have removed the restriction of a local context.

In this case a word or phrase whose value is defined by comparison of the context surrounding it and the context of a cue word or phrase is only a “plug”, reserving a place in a context. This “plug” can be excluded if the measures providing corresponding wildcard in a context are stipulated.

The way to search for similar documents semantically on the basis of the comparison of their lexical vectors is known. But in the task of searching, including all or the most significant components of a lexical vector of the document, the documents concerning different yet close subject domains will be eliminated.

The method presented for semantic search uses the replacement of the most significant component of a query vector by the wildcard. In this case, as a result of search, there are documents that will be retrieved belonging to domains that are different, but close, to the subject domain. These describe distinct concepts corresponding to a component excluded from the lexical vector and some of these concepts are previously unknown to the user.

In Google, the role of wildcard is handled by the symbol “*” where the number of asterisks should not be less than the number of excluded words. Russian search engine Yandex¹ has the special operator for these purposes².

¹ www.yandex.ru

² http://www.yandex.ru/ya_detail.html

5.2. Design of Experiment

A technique for checking Internet documents has been selected in connection with their availability, their huge quantity and the variety of their content.

Two measures of the query results were considered: the whole number of documents found as a result of the query (Y_1) and the number of relevant documents contained in the first 50 selected documents (Y_2). The choice of the second measure was defined by the difficulty for the experimenters to estimate the relevance of all received documents. Thus it is supposed that the search engine returns the majority of relevant documents in the top of the list of references.

The search engine Yandex was used.

The experiment looked at research on the influence of three factors:

A – The fragment of text passed (is substituted by wildcard). The excluded element of the text shall be designated as **X**.

B – Word order in the query.

C – Morphological forms of words.

We vary for all three factors. Factorial, the experimental plan corresponds [5] to 2^3 including eight combinations of query variants (Table 1), where maximum factor value is designated as “+”, minimum – as “-”.

Table 1. The order of queries to the search engine

Trial	A	B	C
(1)	-	-	-
a	+	-	-
b	-	+	-
ab	+	+	-
c	-	-	+
ac	+	-	+
bc	-	+	+
abc	+	+	+

The queries were transformed to the syntax required by Yandex. When the maximum of the **A** factor is shown it means that it was set by the operator “/(+2 +5)”, meaning that the query terms are separated from one to four other words in the resulting text.

This provides a template in which words and expressions can fall synonymous with the excluded element **X**. We call this template the “semantic trap”.

For an estimate of model adequacy, each of the experiments included two independent returns. The first return was obtained by the query: «*географические атласы стран Европы*» (“*geographical atlases of the countries of Europe*”) where **X** = «*атласы*» (“*atlases*”). The second was: «*поиск новых знаний в интернете*» (“*search for new knowledge on the Internet*”) where **X** = «*знаний*» (“*knowledge*”).

Thematically, these two phrases are not connected with each other.

The randomization of results was reached through the casual order of the query task. Results are shown (Table 2).

It is apparent (Table 2), that the results are rather changeable, especially for Y_1 . There is an observed stratification of results between returns. The reason for this can be

the influence of unconsidered factors, such as the number of query terms. To eliminate this factor, a normalization function was applied to the return for each query (Eq. 2):

$$\hat{Y}_{ji} = (Y_{ji} - Y_{imin}) / (Y_{imax} - Y_{imin}) \quad (2)$$

Where: j = trial number; i = number of return; Y_{jimin} = minimum value of the return function for the i^{th} result; Y_{imax} = maximum value of the return function for the i^{th} result.

Table 2. The experiment results

Trial	Return	Result Y_1	\hat{Y}_1	Result Y_2	\hat{Y}_2
(1)	1	48607206	0,975772	23	0,437500
	2	279573078	0,999740	15	0,789474
a	1	47980297	0,963187	21	0,375000
	2	279567901	0,999722	15	0,789474
b	1	18	0,000000	9	0,000000
	2	0	0,000000	0	0,000000
ab	1	1098	0,000022	33	0,750000
	2	1188	0,000004	19	1,000000
c	1	49814093	1,000000	13	0,125000
	2	279645655	1,000000	17	0,894737
ac	1	48880091	0,981250	12	0,093750
	2	14676	0,000052	15	0,789474
bc	1	147	0,000003	41	1,000000
	2	17	0,000000	5	0,263158
abc	1	17907	0,000359	30	0,656250
	2	254	0,000001	6	0,315789

Mathematical models of results were constructed and after excluding of insignificant factors a mathematical indicator of query relevance was obtained (Eq. 3 and 4):

$$\hat{Y}_{1r} = 0,433 - 0,865B \quad (3)$$

$$\hat{Y}_{2r} = 0,517 + 0,157A + 0,207AB - 0,264AC + 0,122BC - 0,246ABC \quad (4)$$

Here, symbols B and C designate the contribution of syntax (word order) and morphology respectively. It is characteristic that there was only an interaction of these factors is significant.

This shows that in semantic search it is still necessary to consider both the morphology of query words and syntax as well. It is obvious, that for this purpose, it is necessary to analyze the texts of documents but not indexes of contemporary search engines, because search engine indexes to not keep punctuation which is essential for text processing.

5.3. Discussion of Experimental Results

Assumption laid in the basis of the experiment was justified. For example, for second query «поиск новых знаний в интернете» ("search for new knowledge on the

Internet") in place of excluded word «знаний» ("knowledge") these words and word-combinations were obtained: «*талантливых авторов*» ("gifted authors"), «*каналов коммуникаций*» ("channels of communication"), «*информации*» ("information"), «*тематических ресурсов*» ("thematic resources"), «*православных страниц*» ("Orthodox pages") and so on.

It is obvious that content of the query is too small for keep retrieved results in one domain. On the other hand retrieved results offers permissible forms of initial phrase overpatching for other domains. It is open a way for automated smooth widening of knowledge sphere of trainable system not limited by initial dictionary of system.

Other experiment results pertinent to linguistic and data extracting was discussed in [4 and 11].

6. Process of Text Disambiguation

Due to Frege's composition principle sentence meaning is a function of meanings of the sentence parts and way of these parts combining. So to text understanding there is important to find accurate meanings for words and word-combinations of the text.

In our task there are several main types of disambiguation needed to be resolved:

- a) Words polysemy and ambiguity. In most cases it can be resolved by means of morphological and syntactic analysis. If proper meaning will not be founded on this stage, all possible meanings will be loaded in bottom layer of text ontology for posterior disambiguation by semantic value and created relationships with other concepts of text.
- b) Identification of concepts and proper names to aggregate their attributes in bottom layer of ontology. This task is general for automated annotation and referring systems. Example of like task is shown in [13].
- c) Identification of pronoun reference. Be guided by [6, 14] the search of noun or proper name to which a pronoun makes reference is limited, as a rule, by one paragraph. After resolve the reference, attributes defined by pronoun (and its reference) are aggregated in bottom layer of text ontology.
- d) Identification of references in compound sentence. In this case an area of reference resolving is limited by this compound sentence. After resolve the reference, attributes of concepts are aggregated in bottom layer of text ontology.
- e) Resolving of anaphora, ellipsis, incomplete sentence will be accomplished using information extracted from previous part of text and stored in text ontology.

Our context based approach proposes to use surrounding context of a text fragment for definition of this fragment meaning and for to resolve disambiguation (homonymy, homograph, anaphora, and so on).

The frequency analysis techniques are used for selection of words or word-combinations having most probability of accordance with given semantic field (with high occurrence frequency) and discriminate analysis techniques are used for selection of words or word-combinations having most significance in given domain for semantic field delimiting (its occurrence frequency is low – an analogue of "slang" method [15] used, for example, for author's identification).

7. Possible Application

One of the distant learning problems is laboriousness of learning courses creating. To meet specific, immediate, and unique learner's needs for personal goals and tasks IBM Corporation developed the solution named Dynamic Learning [16].

One of this solution features is ability for dynamically create a custom "course" out of modular learning objects. To prepare this learning objects specific procedures are used, but operation of metadata editing is executed manually (Figure 4³).

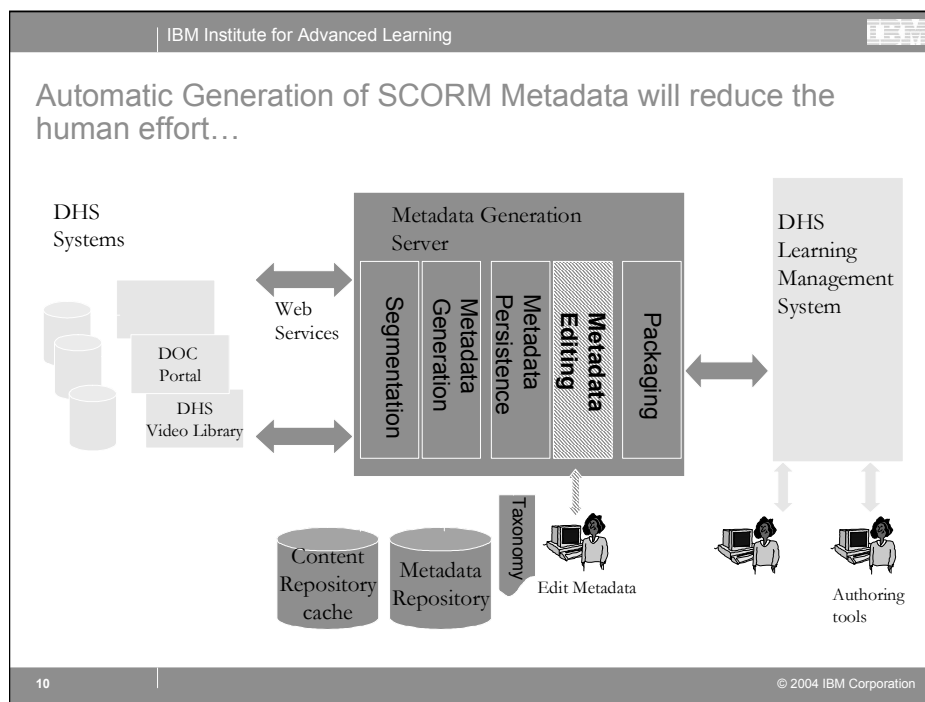


Figure 4. Automatic Generation of SCORM Metadata will reduce the human effort...

Creation of text understanding system can to solve a problem of metadata editing. Furthermore it can to make easier the creating of integrated learning courses by giving access to learning modules related to different taxonomy classes (different educating area) but having common features.

Acknowledgements

This work is partially funded by Human Capital Foundation (<http://hcfoundation.ru>).

³ This slide from [16] is cited with permission of Dr. Y. Ravin.

References

- [1] Jette Kristoffersen, Thomas Troelsgaard, Janne Boye Niemelä, Bo Haardell. How to describe mouth patterns in the sign language dictionary. Theoretical Issues in Sign Language Research 9. Florianópolis, December 06 to 09 2006, Universidade Federal de Santa Catarina. Florianópolis, SC Brazil. (<http://www.tislr9.ufsc.br/index.htm>).
- [2] Vasily Nalimov. Realms of the Unconscious; The Enchanted Frontier. ISI Press, 1982, 320 p.
- [3] Alexander Voskresenskij. Forgetting as a Factor for Knowledge Forming. Proceedings of First Russian Internet-conference on Cognitive Science, 2004 (In Russian: Материалы Первой Российской Интернет-конференции по когнитивной науке — М., УМК «Психология», 2004, С. 150 – 155.)
- [4] A. Voskresenskij and G. Khakhalin. Composition of queries to search engine for knowledge retrieval from Internet. Proceedings of International Conference on Computing Linguistics and Intellectual Technologies “Dialogue’2005” (In Russian: Воскресенский А.Л., Хахалин Г.К. Формирование запросов к поисковой машине для извлечения знаний из Интернета. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог'2005" — М.: Наука, 2005. С. 86 – 91.; <http://www.dialog-21.ru>)
- [5] D.C. Montgomery. Design and Analysis of Experiments (4th ed.), New York: Wiley, 1997.
- [6] A. Fenk, G. Fenk-Oczlon. Information processing limitations and linguistic structure. Proceedings of the Second Biennial Conference on Cognitive Science. Saint-Petersburg, Russia, 2006.
- [7] Carlo Geraci, Marta Gozzi, Costanza Papagno, Carlo Cecchetto. Short term memory and sign languages. Reduced resources and full languages. Theoretical Issues in Sign Language Research 9. Florianópolis, December 06 to 09 2006, Universidade Federal de Santa Catarina, Florianópolis, SC Brazil. (<http://www.tislr9.ufsc.br/index.htm>).
- [8] Artificial General Intelligence /B. Goertzel, C. Pennachin (eds). — Springer, 2007.
- [9] A. Voskresenskij and G. Khakhalin.. About model of NL-text understanding. Proceedings of the Second Biennial Conference on Cognitive Science. Saint-Petersburg, Russia, 2006. (In Russian: Воскресенский А.Л., Хахалин Г.К. О модели понимания ЕЯ-текста. // Вторая международная конференция по когнитивной науке: Тезисы докладов: В 2 т. Санкт-Петербург, 9 – 13 июня 2006 г. — СПб.: Филологический факультет СПбГУ, 2006. — Т. 1, С. 238 – 239).
- [10] V.G. Red'ko. Problem of Cognitive Evolution Modeling. Proceedings of First Russian Internet-conference on Cognitive Science, 2004. (In Russian: Редько В.Г. Задача моделирования когнитивной эволюции. // Материалы Первой Российской Интернет-конференции по когнитивной науке / Под ред. А.Н. Гусева, В.Д. Соловьева — М., УМК «Психология», 2004. С. 14 – 28).
- [11] A. Voskresenskij and G. Khakhalin. Semantic Search Engine in a Multimedia Russian Sign Language Dictionary. Proceedings of XII International Conference “Speech and Computer” SPECOM’2007. October 15 – 18, 2007. Moscow, Russia. Volume 2. pp. 739 – 744.
- [12] D. Lin. Using syntactic dependency as local context to resolve word sense ambiguity. Proceedings of the 35th annual meeting on Association for Computational Linguistics. Madrid, Spain, 1997.
- [13] Z. Kazi and Y. Ravin. Who’s Who? Identifying Concepts and Entities across Multiple Documents. Proceedings of the 33rd Hawaii International Conference on System Sciences – 2000. (0-7695-0493-0/00).
- [14] A.A. Kibrik. Reference and Work Memory: On Interaction of Linguistic with Psychology and Cognitive Science. (In Russian: Кибрик А.А. Референция и рабочая память: о взаимодействии лингвистики с психологией и когнитивной наукой. // Материалы Первой Российской Интернет-конференции по когнитивной науке / Под ред. А.Н. Гусева, В.Д. Соловьева — М., УМК «Психология», 2004. С. 29 – 43.)
- [15] S. Khaitun. Sciencemetrics. (In Russian: Хайтун С.Д. Наукометрия. Состояние и перспективы. — М.: Наука, 1983).
- [16] Y. Ravin. Innovation in Learning from IBM: Examples from the Institute of Advanced Learning. // Learning-on-Demand European Meeting, London, England, 6 December 2004.