A Semiotic Approach to the Intelligent Chinese CALL System Development

Chuprina Svetlana
Department of Computer Science
Perm State National Research University
Perm, Russia
chuprinas@inbox.ru

Osotova Tatyana
Department of Computer Science
Perm State National Research University
Perm, Russia
hvostya@gmail.com

Abstract— **In this paper, we present a novel approach to development of intelligent Chinese Computer-Aided Language Learning (CALL) systems based on ontological engineering methods and a semiotic model of Chinese hieroglyphs. The new features and methods such as "teaching by doing" method, mnemonic novel method and others make the learning process easier and are involved due to above mentioned approach. A semiotic model of Chinese hieroglyph in terms of the first-order logic is described. We have developed the system OntoKit 2.0 based on the results of our full research.**

Chinese CALL system; semiotic model of Chinese hieroglyph; ontological engineering method; mnemonic novel method

I. INTRODUCTION

An integration of traditional and intelligent information technologies (IT) is one of mainstreams of modern software development and of CALL systems implementation in particular. Recently, the most implementations of traditional CALL systems for non-native learners are focused on automation of learning material development, language assessment and learner's training as usual. In our opinion there are not enough in order to get the best learning outcomes. Additionally it is necessary to introduce creative elements to improve learner's cognition skills, to give an opportunity to reveal new regulations and patterns of application domain via self-training (for example, via cognitive games).

This paper presents one of possible approaches to develop intelligent CALL systems as an adaptive self training system and to solve mentioned above problems. We describe the implementation of an intelligent Chinese writing CALL system OntoKit 2.0 environment (see [1, 2] for details). Its ontological knowledge base includes description of semiotic model of Chinese hieroglyph as well as descriptions of the other types of knowledge (tree structure of hieroglyph, syntactical rules of Chinese language, etc). OntoKit 2.0 has became adaptive to different hieroglyphic Asian languages' features due to ontological reengineering methods and universality of supported paradigmatic relations.

In a way we are considering the problem of automation of Chinese characters training as an application domain of integration of such different technologies as pattern recognition (including Artificial Neural Networks), semantic networks and ontological engineering to improve the intelligent capabilities of Chinese CALL systems. The using of techniques mentioned above let us to add to automated training systems the new features to evaluate Chinese characters learning from a simple repetition to the creative process [3]. Ontology based approach makes it possible not only to introduce the new features and methods such as "teaching by doing" method, "mnemonic novel" method, taking into account the historical and culturological aspects of hieroglyphs' evolution into learning process but also provides an original ability of CALL system to advance itself by extension of knowledge base without the necessity to change the source code [1].

The phonological and the phonetic aspects of Chinese language learning are not the issue of our research.

II. METHODS OF ACCELERATION OF CHINESE HIEROGLYPHS MEMORIZATION

"Indo-European languages are based on a finite alphabet" and "letters do not carry meaning unless they are strung together into words" while Chinese language is "made up of symbols that themselves embody meaning", and "the number of possible symbols or elements in these languages is arbitrarily large and can be considered infinite" [4]. And for those who have just started acquaintance with Chinese written language, it seems that there is an incredible variety of characters, and it is almost impossible to remember that.

As mentioned above many CALL systems and Chinese CALL systems in particular are simply a set of lessons with some characters that you just need to learn, and there is no explanation of hieroglyph's structure, which would greatly facilitate the task of the study of Chinese written language for learners.

Chinese language is classified as ideographic, that is, to a system of signs which are used to record the lexical meaning of linguistic units [5]. We support the idea that an image of most of hieroglyphs maps the form of an object (physical or abstract entity) or a group of objects of the reality. In time the images of many Chinese characters became more schematic and abstract. These were the so-called "ideograms", which eventually got the form of modern Chinese hieroglyphs, a lot of which lost its original expressive image during the process of evolution. For example, the character shown in Fig. 1 is perceived not like a simple sequence of strokes, but as a schematic representation of the tree also.

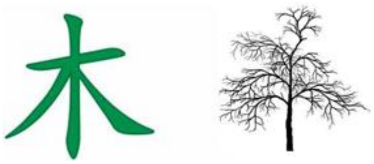


Figure 1. Chinese hieroglyph “Tree” (on the left) and picture of tree (on the right)

However, our imagination can extract the original visual image from the abstract scheme as before [3] if historical and culturological aspects are taken into account especially (see Fig. 2).

The graphics and semantics of Chinese hieroglyphics’ evaluation are considered as a common process. Not only complex Chinese character as a whole, but parts of it (graphemes) carry certain semantic load. Therefore, the meaning of a complex Chinese character depends on the meanings of its graphemes usually [3]. For example, Fig. 2 illustrates the image of Chinese character "Prosperity", which consists of the characters "Woman" and "Roof", so it is the symbol of prosperity for the Chinese people because it describes a situation when “a woman is at home” (under rooftop). And the character "Woman" itself depicts a sitting woman, which bended knees (this pose was accepted for sitting in the ancient China) and with folded hands in a sign of submission [5].

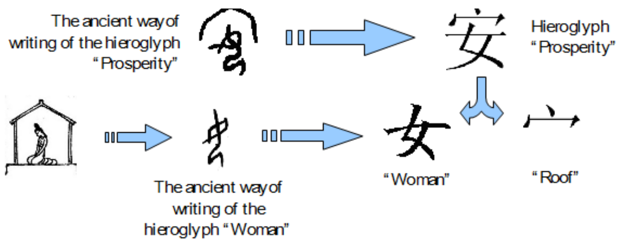

The ancient way of writing of the hieroglyph “Prosperity”
安
Hieroglyph “Prosperity”
女
宀
The ancient way of writing of the hieroglyph “Woman”
“Woman”
“Roof”

Figure 2. Historical and culturological aspects of hieroglyphs “Prosperity” and "Woman"

Taking account of historical and culturological aspects related to the evolution of the graphical representation of hieroglyphs during the process of Chinese characters learning, makes it possible not only to acquaint the learner with the history and culture of China and to feel the mentality of Chinese people better, but also to improve the process of Chinese characters learning itself [3].

However, it happens that it is not enough to learn only the appearance and modern spelling of some Chinese character to recognize its meaning from graphical representation [3]. Then for character learning it requires to create an additional stimulus in the form of visual metaphor to exert influence on learner’s memory [6] (see Fig. 3). Some kind of modification of the mnemonic novel method is used in our approach for this purpose: we take into account images of one and the same hieroglyph from different historical epochs and some additional facts (such kind of knowledge is stored at the ontological base) [3].

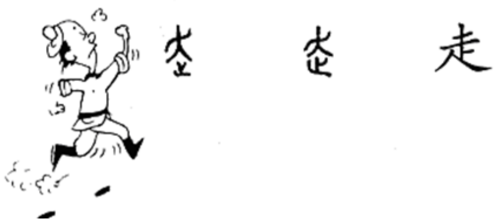


Figure 3. An example of the way of graphic visual metaphor creation for Chinese character "to go" [7]

III. Model of Chinese Hieroglyph

In contrast to the Indo-European languages single Chinese character is considered as a whole word or a part of the word. We have constructed a formal model for the description of the semiotic structure of the Chinese character to solve problems of explanation of Chinese characters’ structure and its features to learners.

The first-order logic is used to formalize the model. However, there are all sorts of ambiguities by reason of incomplete knowledge, because the Chinese language is extensible. In addition, it should be noted there is some inconsistency between different sources about the nature of many Chinese characters’ structure. For example, even such obligatory element of any hieroglyph as a radical has different definitions. In one case the radical is defined via graphemes, but in another case it can be defined via other graphical elements. Thus it is not enough to describe the formal model of Chinese hieroglyph in whole by the apparatus of first-order logic, and we use an apparatus of nonmonotonic logic, the default logic of Reiter (see [8] for details).

The formal model of Chinese hieroglyph in the OntoKit 2.0 is a triple

$$V=\{S,T,K\}, \qquad (1)$$

where S — a simbol of Chinese hieroglyph itself (symbol is considered as a part of Frege semiotic triangle (see [9,10] for details) showed on Fig. 4), T — translation of the Chinese hieroglyph into Russian language, K — knowledge about the hieroglyph, concerning both its structural component (image syntax) and semantics, including historical and culturological aspects [11].

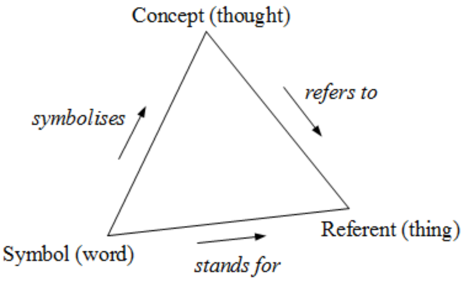


Figure 4. Semiotic Frege Triangle

Let us make some comments on terminology.

Any hieroglyphic character both complex and primitive consists of some number of standard graphic elements, named *strokes* [12]. The set of strokes is restricted (see, for example, Fig. 5). Graphical indication of the stroke is that it consists of one persistent line (see [13] for details).

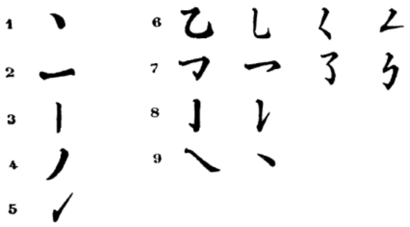


Figure 5. Base strokes of the Chinese hieroglyphics according to [14]

It seems to us that there are no requirements to organize the semantic search based on strokes only.

Additionally there are 24 basic graphical structures (some of them look as a stroke) so called "features" picked out especially for Chinese printed text (Fig. 6). Some dictionaries, for example [15], let conduct the search based on features mentioned above.

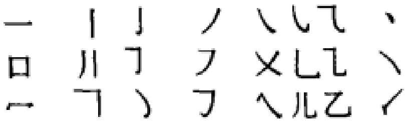


Figure 6. 24 typical features, which are marked out for Chinese printed text [15]

However, representation of hieroglyphic symbol only as a combination of strokes is not enough to understand its meaning. "Meaning" is considered here as a "referent" from the Frege triangle (see [9, 10] for details) showed on Fig. 4. That's why more complex graphical constructions such as graphemes and radicals are required.

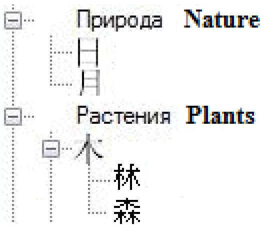
Combinations of strokes with fixed lexical meaning are called *graphemes* (see [12] for details). As the result of our review of the different interpretations of the structure of the Chinese characters information resources we have considered that not all graphemes have a self-contained lexical meaning. On our point of view the inner "context" within the character's image has the significant impact on the recognizing of the grapheme's meaning.

Hieroglyph's structure mathematically represents an ordered tree. Sharapov J.A. in [16] (the author of paper [16] is OntoKit development group member as well as the authors of this paper) points out that generally graphemes may be represented as the leaves of the latter tree. The graphemes' order in the hierarchical structure of hieroglyph is an important part to comprehend its meaning. That's why the ontological knowledge base includes semantic description of hieroglyph's structure. This description is handled by special component responsible to break down (to decompose) Chinese characters to its parts automatically. Description of the model of Chinese character represented as an ordered tree, the method of hieroglyph's decomposition to graphemes and the adequate internal view construction for Chinese characters is beyond the scope of this paper (see [16] for details).

Primitive Chinese character is hieroglyphic symbol consisting of a single grapheme.

Complex Chinese character is hieroglyphic symbol consisting of more than one grapheme.

The *radical* is a semantic component of hieroglyph and hints at the meaning of the character. Within the environment of OntoKit 2.0 mentioned above a radical represents the belonging of any Chinese hieroglyph to the concrete semantic category of an appropriate domain specific ontology (see [3] for details). For example, consider a fragment of domain ontology shown in Fig. 7. We can see that the hieroglyph 森 has the radical 木 "tree" (it belongs to the category "Nature" of domain ontology). It signifies that translation of hieroglyph 森 is related to tree (the meaning of this Chinese character is "thick forest").



Природа **Nature**
- 日
- 月

Растения **Plants**
- 木
  - 林
  - 森

Figure 7. Fragment of domain ontology from OntoKit 2.0

According to [12] the radical is either some stroke (for example, radicals "one", "vertical", "tilting to the left") with no fixed meaning, or a grapheme, or symbol, consisting of 2-3 graphemes (for example, the radicals "to see", "chamois leather", "hemp", "turtle", "flute") as it is showed on Fig. 8. Earlier there were 214 radicals. A total list of radicals has been revised by Chinese linguists recently. The graphic elements, which were used in a limited number of hieroglyphs or rare, have been excluded. The modern list of radicals includes about

190 radicals (different versions of some radicals are taken into account also) [17]. There are a lot of dictionaries based on the radicals search.

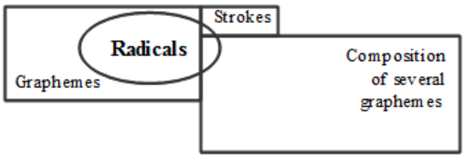


Figure 8. Illustration of corelation between set of Chinese characters' radicals and sets of other graphical elements (from the structure of the image point of view)

The model of Chinese hieroglyph in the language of first-order predicate is presented below (note that we give mnemonic names to predicates for convenience despite the fact that it is accepted to denote predicates by single symbols of the Roman alphabet):

"W" is a set of concepts.

"w" is some concept, $w \in W$.

"X" is a set of graphic elements derived from the strokes that can be called symbols in a semiotic perspective, that is, graphemes, including the strokes with independent meaning, and hieroglyphs

"x" is an element of the set X; $x \in X$.

S(x) asserts that x is a hieroglyph, $x \in X$.

Simp(x) asserts that x is a primitive hieroglyph, $x \in X$.

Comp(x) asserts that x is a complex hieroglyph, $x \in X$.

G(x) asserts that x is a grapheme, $x \in X$.

Key(x) asserts that x is a radical , $x \in X$.

BeKey(x,y) asserts that x is a radical to y ($x, y \in X$, where $Key(x) \wedge S(y)$ is correct).

Mean(x,w) asserts that graphical element x have meaning w, $x \in X, w \in W$.

P(x,y) asserts that x is a part of y; $x, y \in X$.

Eq(x,y) asserts that x is y; $x, y \in X$ (that is graphical elements x and y are equal by inscription).

EqW(x,y) asserts that x is y, $x, y \in W$.

$Op_{modif} = \{VC, HC, NTC\}$ is an operation of distortion, where *VC* is a vertical compression, *HC* is a horizontal compression, *NTC* is a nonformalizable transformation (it is usually associated with some fundamental simplification of the structure, and most of these transformations can be specified by the table).

$modif \in \{Op_{modif}, Op_{modif} \circ modif'\}$, где $modif' \in \{Op_{modif}, modif\}$.

1. Common statements:

$$(\forall x)(P(x,x)), \qquad (2)$$

$$(\forall x)(\forall y)(Eq(x,y) \rightarrow Eq(y,x)), \qquad (3)$$

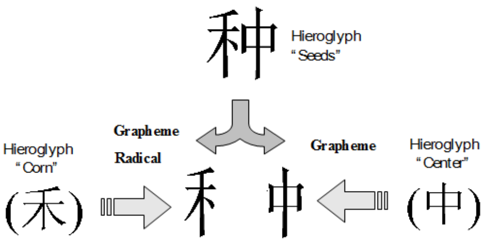
$$(\forall x)(\forall y)(Eq(x,y) \rightarrow (P(x,y) \wedge P(y,x))), \qquad (4)$$

$$(\forall x)(\forall y)(\forall z)((P(x,y) \wedge P(y,z)) \rightarrow P(x,z)). \qquad (5)$$

2. Grapheme is a part of hieroglyph (Fig. 9)

$$(\forall x)(G(x) \rightarrow (\exists y)(S(y) \wedge P(x,y))). \qquad (6)$$

Figure 9. Illustration of the construction features of Chinese characters



3. A Chinese hieroglyph consists of one grapheme at least (Fig. 9):

$$(\forall x)(S(x) \rightarrow (\exists y)(G(y) \wedge P(y,x))). \qquad (7)$$

4. A Chinese hieroglyph is primitive if and only if it consists of a single grapheme (Fig. 10):

$$(\forall x)(Simp(x) \sim (S(x) \wedge (\forall y)((G(y) \wedge P(y,x)) \rightarrow Eq(x,y)))), \qquad (8)$$

where "$\sim$" is a symbol of biconditional operation.

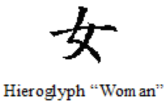


Figure 10. Example of primitive hieroglyph

5. A Chinese hieroglyph is complex if and only if it consists of more than one grapheme (Fig. 11):

$$(\forall x)(Comp(x) \sim (S(x) \wedge (\forall y)((G(y) \wedge P(y,x)) \rightarrow \overline{Eq(x,y)}))), \quad (9)$$

where "$\sim$" is a symbol of biconditional operation (this statement does not exclude the situation when the complex character consists of several identical graphemes).

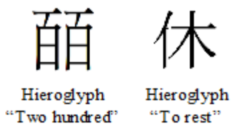


Figure 11. Example of complex hieroglyphs

6. A Chinese hieroglyph is either primitive or complex:

$$(\forall x)(S(x) \rightarrow (Simp(x) \oplus Comp(x))), \quad (10)$$

where "$\oplus$" is exclusive OR operation.

7. A radical is a part of hieroglyph (Fig. 9):.

$$(\forall x)(\forall y)(BeKey(x,y) \rightarrow (S(y) \wedge Key(x) \wedge P(x,y))), \quad (11)$$

8. There is a radical in any Chinese hieroglyph and it is just one (Fig. 9):

$$(\forall x)(S(x) \rightarrow (\exists y)(BeKey(y,x) \wedge (\forall z)(BeKey(z,x) \rightarrow Eq(z,y)))). \quad (12)$$

9. One and the same radical in the context of different characters can have different meanings [12]:

$$(\exists x_1)(\exists x_2)(\exists y_1)(\exists y_2)((BeKey(x_1,y_1) \wedge BeKey(x_2,y_2) \wedge Eq(x_1,x_2) \wedge \overline{Eq(y_1,y_2)}) \wedge (\exists w_1)(\exists w_2)(Mean(x_1,w_1) \wedge Mean(x_2,w_2) \wedge \overline{EqW(w_1,w_2)})). \quad (13)$$

An example of such radicals is shown in Fig. 12: Chinese hieroglyphs "City" and "Hill" in the function of radical have the same spelling. These radicals can be discerned in the character only by their position: "City" is used on the right, "Hill" - on the left.

10. There are radicals with the same meaning but different spelling:

$$(\exists x_1)(\exists x_2)(\exists y_1)(\exists y_2)((BeKey(x_1,y_1) \wedge BeKey(x_2,y_2) \wedge \overline{Eq(x_1,x_2)}) \wedge (\exists w)(Mean(x_1,w) \wedge Mean(x_2,w))). \quad (14)$$

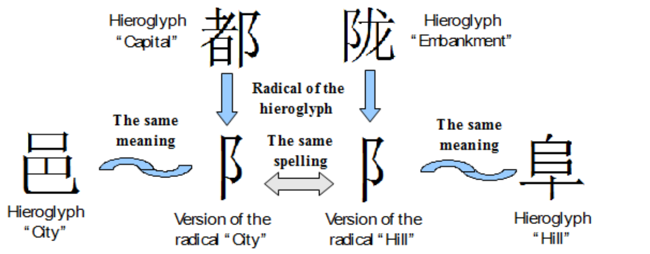


Figure 12. Radicals, which in the function of radical have the same spelling, but different meaning

Let us explain this statement with an example. In Fig. 13 there is shows the character "Fire", it is the radical of the character "Flame" (but in the character "Autumn" this character is only a grapheme). In addition, the character "Boil" has as the radical "Fire", although the spelling differs.

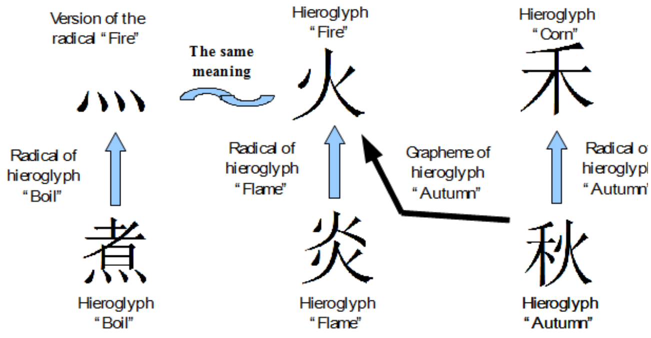

Version of the radical "Fire"
The same meaning
Hieroglyph
"Fire"
Hieroglyph
"Corn"
灬
火
禾
Radical of hieroglyph "Boil"
Radical of hieroglyph "Flame"
Grapheme of hieroglyph "Autumn"
Radical o hieroglyp "Autumn
煮
炎
秋
Hieroglyph
"Boil"
Hieroglyph
"Flame"
Hieroglyph
"Autumn"

Figure 13. Illustration of the features of radicals of Chinese characters

The statements represented below are noted with default logic of Reiter.

11. A radical is a grapheme (Fig. 9), if it does not contradict with other knowledge represented in the system (in Fig. 8 it is reflected that this statement is not always true):

$$(\forall x)(Key(x) \rightarrow (\exists y)(G(y) \wedge Eq(x,y))), \quad (15)$$

$$\frac{Key(x) : M(G(y) \wedge Eq(x,y))}{G(y) \wedge Eq(x,y)}. \quad (16)$$

12. Primitive hieroglyph associates with a grapheme with certain distortion (Fig. 9), if it does not contradict other knowledge represented in the system (some

primitive characters are presented by grapheme without distortion):

$$(\forall x)(\exists y)(Simp(x)\rightarrow (Eq(x, modif(y))\wedge G(y))), \qquad (17)$$

$$\frac{Simp(x): M(Eq(x, modif(y))\wedge G(y))}{Eq(x, modif(y))\wedge G(y)}. \qquad (18)$$

We use the model described above as a basis of the development of our CALL system OntoKit 2.0 and so called "teaching by doing" method in particular. Due to this method a learner has the facilities not only to decompose Chinese character to graphemes in order to know its meaning but to compose a new hieroglyph from different graphical constructions to "design" a new meaning based on the meanings of its source parts also [2]. It brings elements of creative work to the learning process and also provides an ability of CALL system to advance itself by extension of knowledge base without the necessity to change its source code [1].

A general scheme of decomposition process is illustrated in Fig. 14.

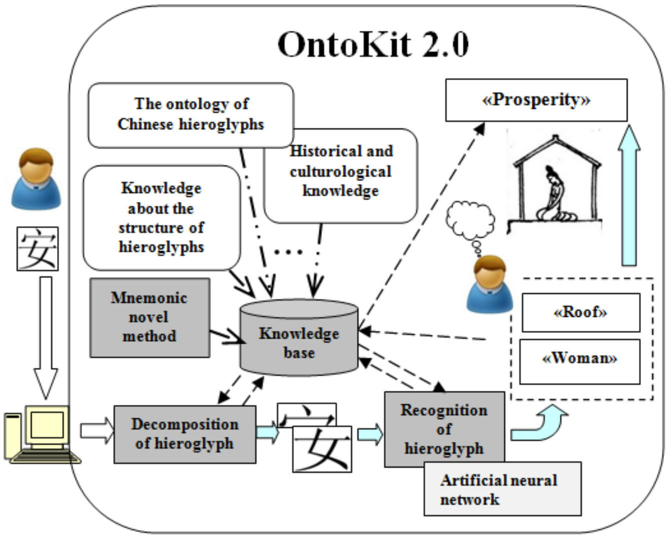

OntoKit 2.0
The ontology of Chinese hieroglyphs
Historical and culturological knowledge
«Prosperity»
Knowledge about the structure of hieroglyphs
Mnemonic novel method
Knowledge base
«Roof»
«Woman»
Decomposition of hieroglyph
Recognition of hieroglyph
Artificial neural network
安

Figure 14. A general scheme of automation of Chinese characters learning process [18]

There is a scheme of hieroglyph's image analysis in Fig. 15. An image of hieroglyph (in bmp or jpg format, for example) is an input of this process. The system breaks down image to images of hieroglyph's components automatically. As the result of this process the internal view of hieroglyph's components description in ordered tree form is generated due to recognition based on the convolutional neural network using [19]. The system uses the results of recognition to help someone learn the Chinese hieroglyphs, memorize its spelling, or perform other learning-related activities via cognitive games.

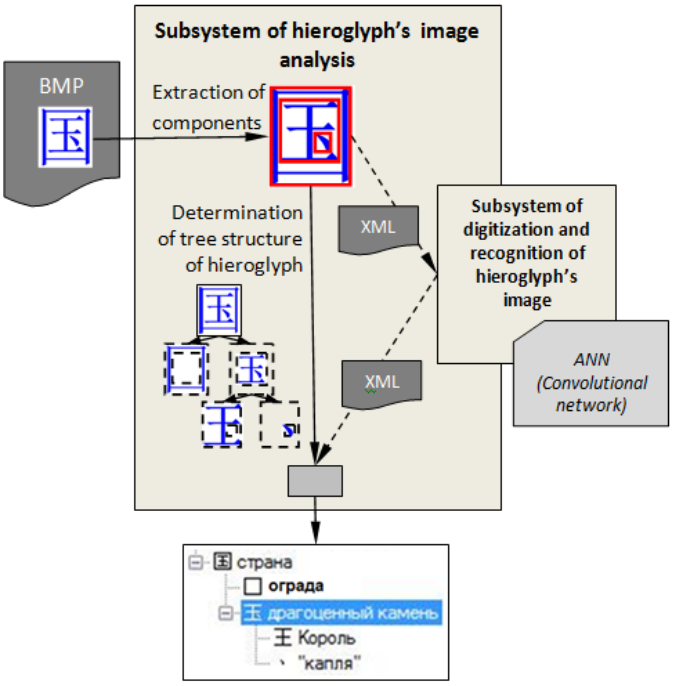

Subsystem of hieroglyph's image analysis
BMP
Extraction of components
Determination of tree structure of hieroglyph
XML
XML
Subsystem of digitization and recognition of hieroglyph's image
ANN (Convolutional network)
страна
ограда
драгоценный камень
Король
"капля"

Figure 15. OntoKit 2.0: a scheme of hieroglyph's image analysis

Conclusion

This article focuses on the problems of automation of Chinese characters training and development of an extensible Chinese CALL system with new intelligent facilities based on the original ontological approach and semiotic model of hieroglyph.

The semiotic model of Chinese character in terms of the first-order logic and non-monotonic default logic of Reiter is represented.

We have implemented the research prototype of OntoKit 2.0 based on the proposed approach by using C# and C++ languages. So the viability of described approach have been proved. That's why we consider our results as full research.

References

[1] Chuptina S.I., Sharapov J.A., Osotova T.V. "The ontology approach to creation of an automated adaptive system for teaching chinese characters" // Proceedings of XXXVI International Conference «Information technology in scince, sociology, economics и business» IT + SE'09 The Autemn session. M.: 2009, pp. 53-55. (rus)

[2] Osotova T.V. "The approach to Chinese character recognition at the CALL system OntoKit 2.0" // Proceedings of scintific academic conference "Modern problems of mathematic and its applied areas" Perm: Perm State University, 2010, pp. 114-118. (rus)

[3] Chuptina S.I., Sharapov J.A., Osotova T.V. "Automatisation of Chinese language training: historical and culturological approach" // Proceedings of scintific conference «Historical and Cultural Heritage and Information and Communication Technology: retention analysis». Perm: Perm State University, 2009, pp. 202-213. (rus)

[4] Jurgens H., Peitgen H.-O., Saupe D. “The Language of Fractals” // Scintific American.1990. №10. P. 36.

[5] Volkova O.N. “Culture-oriented linguistics of the first foreign language (chinese language). Module 1. The Chinese grammotology” Version 1.0 [Online resource]: online educational manual. URL: http://www.files.lib.sfu-kras.ru/ebibl/umkd/346/u_course_1.pdf (access date: 30.01.2012). (rus)

[6] Vurdov A.M. “Japanese language for pleasure. Kanji essays” Syktyvkar: Uki, 2006, p.528. (rus)

[7] Lusya V., Starostina S.P. “The Chinese-Russian educational dictionary”.– M.: AST: Vostok – Zapad, 2006, p.382 (rus)

[8] Ueno H., Koyama T., Okamoto T., Matsubi B., Isidzuka M.. “Knowledge representation and deployment” Trans. from jap. – M.: Mir, 1989, pp.199 – 207. (rus)

[9] Nesterov A.V. “On semantic, pragmatic, and dialectic triangles” // Automatic Documentation and Mathematical Linguistics, Vol. 43, №3. Allerton Press, 2009, pp. 9-14.

[10] Frege: “On Sense and Denotation” // UW Faculty Web Server. URL: http://faculty.washington.edu/smcohen/453/FregeDisplay.pdf (access date: 30.03.2012).

[11] Osotova T.V. “The role of semiotic description at automated Chinese language training” // Proceedings of IV International Online Scintific Students’ Conference «Students’ scintific forum» URL: http://www.rae.ru/forum2012/pdf/2721.pdf (access date: 26.03.2012) (rus)

[12] Kondrashevskii A.F. “Practical cource of the Chinese language. Hieroglyphic’s guide’” Part I. M. : PH «Muravey», 2000, p.152 (rus)

[13] Hieroglyph’s structure // Educational center «Sakura» URL: http://www.sakura-inyaz.ru/struktura_ieroglifov.html (access date: 15.12.2011). (rus)

[14] Wieger L. “Chinese Characters. Their origin, etymology, history, classification and signification. A thorough study from Chinese documents” – New York: Paragon book reprint corp. — P. 12.

[15] Panasyuk V.A., Suhanov V.F. “The Great Chinese-Russian dictionary” Vol. 1. M.: «Nauka», 1983, pp.10 – 12. (rus)

[16] Sharapov J.A. “The algorithm of inventive tasks solving at the task of Chinese characters’ structure description” // Proceedings of XXXVII International Conference «Information technology in scince, sociology, economics и business» IT + SE’10 The Spring session. M.: 2010, pp. 86-88. (rus)

[17] Internet-school of the Chinese language. Issue 4. The Chinese writing. Knowledges aboute the structure and spelling of hieroglyphs // Internet-school of the Chinese language URL: http://chinese-school.narod.ru/004.html (access date: 23.12.2011). (rus)

[18] Osotova T.V. “The semiotic approach to aftomation of Chinese character training” / The bulletin of Perm State University. Mathematics. Mechanics. Informatics. Issue 3(7). Perm, 2011, pp. 59-62. (rus)

[19] Quen-Zong Wu, Yann Le Cun, Larry D. Jackel, Bor-Shenn Jeng On-line recognition of limited-vocabulary Chinese character using multiple convolutional neural networks // Yann LeCun home page. URL: http://yann.lecun.com/exdb/publis/index.html (access date: 23.05.2009).