Boris M. Putko

Introduction to

Structural Hieroglyphics

2018

This text is author's translation of the Russian edition of the "Introduction to Structural Hieroglyphics" ("Введение в структурную иероглифику", 2018, Lambert Academic Publishing, ISBN - 978-613-9-81566-1). Translation into non native language is not the best idea. But I don't see another way to introduce the readers to new method of input Chinese characters. I hope that after correcting the text with the help of Yandex, Google and my son Anton, the number of remaining errors does not spoil the text enough to make it unreadable. Effectiveness and simplicity of mastering composition input method make it irreplaceable for learning Chinese hieroglyphics generally. And I want to tell about this method to readers who interested in Chinese hieroglyphics and are ready to look for a new things in such a traditional area.

If you have something to say to the author about this text, about his (the author's) mistakes or misconception or just translation errors, please feel free to contact me at poutko@yahoo.com

Sincerely Yours,
B. Putko.

17.08.2018

"角之字形，乃刀下用也。今头上用刀，其凶甚矣！"

[The character] "horn" has the following structure: [above is written] "sword", and below "apply". The sword hangs over [his] head now, it's bad!
(Luo Guanzhong, Romance of the Three Kingdoms, ch.104)

I do not think that I will reveal to anyone a "terrible secret" if I say that hieroglyphic texts reading is a recognition of complex graphic images. Then, the study of hieroglyphics is the simultaneous solution of three different tasks in essence: the recognition of images, the fixation of these images in memory and the creation of graphic-semantic associations between these images and their meaning or graphic-sound associations between the image and its pronunciation or its "name" . Structural hieroglyphics, at first glance, is related to the solution of the first of these three tasks, but, as we shall see below, it is able to provide effective assistance in solving two other tasks, as well as in some utilitarian applications, which are extremely necessary for studying Chinese characters and mastering Chinese language in general, for language lives not only in sound but in texts too. And in addition to helping with recognizing and memorizing characters, structural hieroglyphics, as it turned out, can help us in finding unknown characters in a specially organized dictionary, and (Oh, God!) in the Holy of Holies of classical hieroglyphics - in writing of hieroglyphs. Of course, not in writing with a brush on paper or silk, but in the input of characters on digital devices from computer to phone.

Hieroglyphic dictionaries are already more than 2 thousand years old, and knowledgeable people have long understood the significance of these "devices" for the transmission of tradition. But in China (perhaps, and therefore) there has always been a close connection, if not to say a "cohesion" between the student and the teacher. Exactly within the framework of this close connection it was possible transmission of the tradition with the help of dictionaries with minimal distortion. The dictionary could be used only by a person who is well acquainted with the basics of hieroglyphics, having good skills received from the teacher. In those old times hieroglyphs in the dictionary (in the "lexicon") were located either "by themes" or "by radicals". As the Russian sinologist V. P. Vasiliev noted, " it is clear that neither one nor the other [method of arrangement] can be suitable for a beginner to learn; one must first learn the language in order to seek out [the necessary hieroglyphs] in these lexicons. The Chinese do so, because they always learn from teachers for several years, with the help of which they remember the reading and meaning of all the main characters. Without a teacher, with the help of one lexicon, it is not clear for the Chinese to learn their own language." [1, p. III, (translation into English is my - B. P.)]. With the advent of various recording systems for pronunciation of Chinese characters, the situation is unlikely to become easier. To find the character in the dictionary by pronunciation, you must already know its pronunciation.

Russian graphic system, the founder of which was V.P. Vasiliev, not to say that interrupted this tradition, but showed that a sophisticated analytical mind, "european genius", as said V.P. Vasiliev, is able to offer a fairly convenient methods of structuring graphic images. From my own experience I can say that the magnifying glass and half an hour were enough for me to find, without any help, the first Chinese character in my life in the Mudrov's dictionary [2], built on the principles of the Russian graphic system. This system had a noticeable effect on the encoding of the characters "by four corners" (Four-Corner method), adopted in China [3], but nevertheless, in the 20th century and at the beginning of the 21st, dictionaries, even electronic ones, remained essentially "flat", simple emulators of paper dictionaries. The simple idea of search a character not by one radical, but by any combination of its elements or by four corners or by the last element, it does not matter, still can not get enough distribution in the organization of electronic dictionaries. The idea of such a search for a character by a set of predefined components is trivial, and the implementation of such a mechanism has appeared for a long time. It is worth mentioning the word processor of the Australian NJStar Software Company [4], where, at least since the beginning of the 2000s, the "choice by radicals" (Radical Lookup) has already been implemented. In 2015, a new feature "Show entries containing character components: (汉字部件)" appeared on the network resource "MDBG English to Chinese dictionary" [5], which allows you to find a character by specifying one or more of its components from the list of 300 pieces. Pleco Software, the manufacturer of the well-known electronic dictionary Pleco, according to its founder Michael Love, in 2017 led his own development of such a tool [6], but I am not familiar with the results of this development.

Understanding of the structure of characters, regularities of their formation, and, most importantly, regularities of perception, recognition and memorization, in my opinion, is able unusually to facilitate the mastering of the hieroglyphics, is able to reduce the threshold of entering into the hieroglyphics for beginners and offer further opportunities of studying the hieroglyphics to the specialists.

The structural approach to hieroglyphics is by no means new, the obvious advantages of its use were clear to sinologists still the 19th century. The analysis of the graphic composition, the similarities and differences of characters without regard to their meanings and pronunciations provide additional advantages in memorizing characters. "Moreover, you have before your eyes a concentrated picture of the gradual growth of Chinese writing; if two similar hieroglyphs are placed next to one another, then you certainly remember them most likely; if you look for an unknown hieroglyph, then you will involuntarily remember those that are similar, and most importantly, with this system of growth, it seems only possible to learn to write in Chinese. It is not easy, for example, to write the hieroglyph 嬌 from the first time, but when you were first given 冂, then 高, then 喬, and finally 嬌, you will not only remember the hieroglyph, but before your eyes will be whole groups of other characters." [1, p. VIII (translation into English is my - B. P.)]. But too early attempts to use the analysis of the structure of characters for machine processing, in particular, to input characters ("cangjie" methods and "wubi") seem to only discredited the very idea of a structured approach to writing. Insufficiently clearly formulated task, arbitrary separation of characters on elements for their description, not associated with the natural way of perception , the desire of the authors "to embrace the immensity", the lack of instrumental capabilities of processing combination of elements, gave birth to strange monsters that can only live in the head of the developer and among the convinced adepts of these systems, and, as practice has shown, are difficult to perceive and easily rejected by users.

It seems to me the structural hieroglyphics are able to avoid the traps of unfounded claims. The simplicity and clarity of the implementation of the composite input, based on the rules of structural hieroglyphics is encouraging. We cannot say that the creation of Chinese characters was a kind of "project", that there was a certain "standard" of adding a characters into the General Treasury of characters, but it is impossible to say that the creators of hieroglyphs did it thoughtlessly. Without a doubt they had some thoughts on this issue, without a doubt they were guided by common sense, and tradition, and some aesthetic ideas. And the centuries-old "adjustment and running in", seemingly arbitrarily created characters, led to the fact that now we have what we have, maybe not fully understanding the causes and origins of the current state of affairs. And it seems to me permissible to refer here to the authority of Isaac Newton, who in the "The Mathematical Principles of Natural Philosophy", speaking of the reasons for the properties of the force of gravity, said that "And to us it is enough that gravity does really exist, and act according to the laws which we have explained," [7]. The mechanisms of perception of graphic images may differ from those described in this paper, may be quite different, but in this case, we are satisfied that these mechanisms does really exist, and act for many centuries, creating, maintaining, and providing future prospects for this amazing cultural phenomenon — hieroglyphic writing. And it is these mechanisms that allow the structural hieroglyphics to achieve what it so far only claims, but the claim is very reasonable. I hope that for the reader it will be quite clear from the following text.

1. Structural hieroglyphics. Prolegomena.

1.1. Elements of the characters. Graphs.

Structural hieroglyphics belong to that amazing category of subjects that stand on simple and obvious grounds, but the effectiveness or heuristic value of an entire subject becomes apparent only when these, in themselves, quite obvious grounds are gathered in one place, and in a certain order, strictly defined and aligned with each other to achieve a very specific goal, i.e. when the subject is sufficiently formalized. The claim to build a formalized system of hieroglyphics is one of those traps of unreasonable claims that I would like to avoid, but the attempt to construct something useful in the first approximation is, of course, connected with the need to more or less accurately imagine the task. In other words, the structural hieroglyphics in its current version can be thought of as formulation of the problem, namely as an attempt to "organize" a vast field of hieroglyphics in order to simplify the recognition, memorization, and use (input) Chinese characters.

At first glance, and this is correct, all the Chinese characters consist of separate elements. This maxim alone does not surprise anyone, but it does not clarify anything either. In classical hieroglyphics, for example, it is believed that the character consists of a set of strokes, which are distinguished from 32 to 38. In the conventional sense, the stroke is an element of a character that can be depicted without taking the brush away from the paper. But, this obvious statement, like other self-evident statements, can not give anything "neither to the mind, nor to the heart." Chinese characters itself originated as a depiction by draw a sequence of strokes on the carrier, but knowing this in itself does not give anything other than the ability to reproduce characters (as a kind of ancient function "copy-paste") – if you want to learn how to read and write hieroglyphic texts, you mast copy what you see in the right order. And the current computer systems of handwriting, in fact, have not gone away from drawing characters on the sheep's shoulder blades or on the shell of a turtle. Sophisticated machine algorithms used in handwriting are designed to make the machine understand what the man has drawn by his hand, to identify what was written, to keep it in an acceptable form for the machine. This is a huge leap in the development of machine algorithms, but this all has nothing to do with man writing hieroglyphs.

On a second glance, and it is also correct, the division of characters into individual strokes are not very informative. The sequence of strokes relative to a simple character is easy to remember and reproduce, but for a graphically complex characters to remember and reproduce the sequence of a large number of strokes simply, without additional tricks, is already impossible. Strictly speaking, people have never done this. Complex characters were originally created as compounds, as modifications of already existing characters by adding either additional strokes or entire "blocks" of such strokes, individual elements, which in themselves could have some meaning and could act as independent characters. It was also widely used in mnemonics, in writing characters and even in their interpretation (see for example, the epigraph to this text). In other words, the isolation and use of such blocks was a simple and *natural* way of developing hieroglyphics. The emphasis should be made on the word "natural", and, as we will see further, it is the natural way of perception of complex graphic images that create the basis of structural hieroglyphics.

It seems that now we can allow ourselves to claim that all the characters consist of a sequence of strokes, but we must add that most of the characters consist of some blocks of strokes, some elements that can be perceived as separate, easily recognizable entities, traditionally often having their own names: 手 — "an arm", 戈 — "a spear", 鸟 — "a bird". Combining these entities creates a new entity with a new name. "An arm" and "a spear" create a new essence 我 with the name "I". "I" and "a bird" give character 鹅 with the name "a goose". Note that even merging with each other, like "an arm" and "a spear" in the character "I" (我), and even having a common stroke, these entities are easily isolated by our consciousness, remain *invariant* for him.

The fixation of these easily recognizable entities was, in fact, the beginning of structural hieroglyphics. Their existence was known, was used in mnemonics and in the reproduction of complex characters, but was fixed for the first time in the compilation of dictionaries, when characters in the dictionary began to place in accordance with one of these elements, with the first, most often, when writing a character, and the main or most important when remembering or interpreting it. The Chinese called these elements "the sections headings" (部首) and they were right. Each character has its own sections heading. More precisely, the sections headings has a dictionary, and each character has a certain element, which is considered the main in the character and acts as a sections heading in the dictionary. Now these elements of characters are called, for the most part, *radicals*. It should be noted that traditional hieroglyphics is very free to use terminology. The same component of a character in different situations can be anything: a grapheme, a key, a radical, a determinative, even radical-determinative, a phonetic, – at the will of the author, who believes that everyone still understands what he means in this or that situation. And if somebody don't understand what the author has in mean, that is his own problem. Most often "a radical" in the traditional hieroglyphics is a certain part of the character, about which you are told — "this is the radical." We're going to do the same, if one say to us it is "the radical" then it is the radical. Sometimes the character as a whole itself can be a radical. Usually this is a fairly simple character, which is not that it can not be divided into elements (it is possible to divide any character, except for the character "一" or "乙"), but for some reason it is not necessary to separate it. It can be quite a complex character, for example, 鬼 or 黑. Then I will say that such elements are *atomic*, i.e. indivisible.

Thus, we can say that the structural hieroglyphics began with the isolation of a separate entity, which is still often called a radical. But the technological capabilities of that time did not allow to realize the obvious possibilities of combinatorics of radicals. Combination of the two elements can be easily implemented in a two-dimensional table, but what to do with increasing their number? The third element makes cross-tables inoperable. "Flat" dictionaries allowed you to simply fix only the radical index ("the section headings" in this case) and the "remainder", and somehow treat with this remainder after the section heading isolation for later retrieval within found section. Characters with this section heading could be arranged within the sections according to the thematic principle or pronunciation, but in the current dictionaries the variant was fixed by the number of strokes "in the remainder". This allowed somehow to solve the problem of classifying characters, helped to memorize them, and helped to reproduce the character, which was not clearly fixed in memory, when the first element was remembered and there was a hope that the "hand itself will remember" the rest in the process of writing. The number of these sections depended on the set of characters, their number and, to a large extent, on the arbitrariness of the dictionary compiler. From general considerations, it is understandable that in the absence of a standard in compiling a dictionary that itself became the standard for future generations of users, the number of sections directly depended on the number of characters in the dictionary - the more characters there were, the more sections. And the number of sections in the early dictionaries reached 540 as in the well-known dictionary "Shuowen Jiezi" (說文解字) of the Han era. In this state the hieroglyphics stayed for centuries. Until there was an urgent need to standardize the outline of the characters. At the same time it became clear that it is necessary to start with the standardization of the main identifiers of characters, with their section headings, and their number for the first time in the dictionary 字彙(Zìhuì) of 1615 year was limited to 214. This set of sections and, accordingly, set of their headings has been used for almost 100 years, when the then Emperor ordered for 5 years to publish the "only correct" dictionary, which was done in 1716. This quite reasonable Emperor, by personal name Xuanye (玄燁), better known for the era name of his reign as the Kangxi (康熙). And dictionary, published under his reign is still called the Kangxi Dictionary (康熙字典). And the set of section headings of this dictionary is still called Kangxi radicals. It is to this set of characters elements that we will refer later, but now from separate and well-defined considerations of structural hieroglyphics.

In addition to the role of identifiers in dictionaries, radicals have long been performing several tasks. Certainly, the radicals carry a powerful and proven for centuries the semantic load. As a rule, characters denoting certain concepts, objects, actions related to one class have the same radical in their composition. And one can find a lot of examples in the textbooks. The radicals as a rule perform the role of semantic determinative, textbooks love to to divide the characters on determinatives, and phonetics [8]. The determinative "serves to convey the approximate meaning" of sign, phonetic "conveys the exact or approximate sound" of the sign [ibid., p. 161]. What does "the approximate meaning" means? What does "the approximate sound" means? As long as you use a thousand characters specially selected for you by a teacher, everything seems simple and logical. But as soon as you go beyond the educational dictionary, it looks a little different. A simple example: the traditional hieroglyph 義 (yì), the determinative – 羊 (yáng), phonetic – 我 (wǒ)! And then came the reform to simplify the Chinese characters. And in a simplified outline this character looks like this - 义. Where is the determinative? Where is the phonetics? The fact is that the division of characters into determinative and phonetic has nothing to do with the structure of the characters. The sound of the character at the moment when it was created is a separate thing. How Chinese people draw character and how they voice to this character now, does not have to be identical to what took place at the time of this character's birth. Determinatives and phonetics are relevant only to mnemonics. And they depend only on the sharpness of the mind of the teacher. What mnemonic scheme he will come up with in order to make it easier for you to remember (attention!) three completely different things: the graphic image, its pronunciation and its meaning. Simple character 椅 (yǐ) – chair: determinative – 木 (wood), phonetic – 奇 (jī? or qí?). And try to understand now, what does wood have to do with the device on which you sit, which is regulated in three planes, and composed entirely of metal, plastic and fabric? Ah, you need to add in fact that such devices, no matter how it looked like, which used for the same purpose, previously was made from wood. And the result is that this associative information, being tied to a specific graphic image, makes it easier for you to remember it. Everything else is from the evil one.

The same can be said about the allocation of the group "graphemes" in the description of characters, the separation characters on "monograms", "heterogams", "ideographs", "phonoideograms". Such classifications do not have any structural sense, because in the structure of the character there are no criteria, features, parameters or properties that determine the belonging of the character to a particular group. Such arbitrary classifications allow only to structure the educational material, to facilitate its transfer, i.e. they carry a purely didactic load. While the authoritative teacher does not tell the student that the character 吃 is an ideogram, and the character 吧 is a phonoidentogram, there is no way to attribute the character to a particular group. Why 玉 is grapheme, and 王 is monogram, why 男 is ideogram, and 气 is phonoideogram, why 果 is monogram, and 栗 is heterogram? (All examples are taken from [8]).

Let's forget about such pseudo-classification of characters. Once again we reformulate the basic assertion of hieroglyphics: each character ("two-dimensional picture") can be represented as a set of some elementary units (at the same time, some characters can consist of only one such unit, i.e. of itself). These elementary units in different characters may have different relative sizes, occupy different positions relative to other similar units. I. e., relatively speaking, in a certain complex character unit A may be less than or more than unit B, it may be located above or below, to the left or to the right of unit B, while remaining itself, a certain elementary unit A, easily noticeable in the composition of the character and easily identifiable, i.e. some *invariant*. In this case, we can safely abstract from the mutual size of these units.

To get rid of the indication of relative position of these units in the composition of the character, we will use the concept of *the direction of decomposition*, we will "cut" the hieroglyph into components and describe its composition starting from the upper left corner and move consistently to the lower right — from top left to bottom right. Note, the tradition has done everything it could, so that we could come to this decision. First, traditional hieroglyphics requires that regardless of the mutual size of its elements, each character fits into a conditional square, the size of which should be preserved throughout the text, and secondly, the elements of the characters, basically, should be written from the left top to the right down. Tradition is a good thing, and you may break it only when you really need. Herewith, you need to understand why you do it. Thus, using the tips of tradition, we can describe a two-dimensional image of the character as a linear sequence of its elements, in which the element number reflects its position relative to other elements. Herewith, sometimes there are some special situations , but we will talk about them some later.

Pay attention, I just said a very important thing: structural hieroglyphics *describes* characters using a linear sequence of elementary graphical units. And then we should talk about these graphic units. Namely: to establish which graphic units can be used to describe the structure of characters, what properties they should have, how many they should be, what properties should have the whole set of these graphic units, what pool of characters can be described with this set of graphic units.

In the first approximation, these graphical units can be considered as components of the characters, and then, we can say that the character can be described simply by specifying the sequence of its components. And again, I will not surprise anyone if I say that the number of such graphic units necessary to describe characters depends on the number of characters that we want to describe with these components. It is believed that the number of characters in Chinese is unimaginable. Perhaps this is so, but it is absolutely impossible to work with such pseudo-quantities as "great", "huge", "unimaginable", etc. It is impossible to embrace the immensity and you need to start from something more mundane and accurate. For example, the UNICODE standard [9] currently describes 87870 characters in all blocks related to Chinese, Japanese, Korean, and Vietnamese. However, only the CJK Unified Ideographs block, which the standard defines as common [ibid., p. 683], contains 20971 characters. A set of IICore (International Ideograph Core), presented by the Ideographic Vice-Chairman Group in May 2004 (Last Updated: 2004.06.16) [10] contains 9810 characters, about which the source says it is a set of constantly used hieroglyphs throughout Southeast Asia, and that "they can handle all the needs of almost all of their customers " [9, p. 685]. 10 thousand is not 87 thousand. Another interesting source: a set of characters of modern Chinese language in simplified outline indicating the occurrence of the characters: The Lancaster Corpus of Mandarin Chinese - compiled by Tony McEnery and Richard Xiao [11]. The entire list consists of 6839 characters. This is already quite a visible amount. If you exclude from the list those characters that occur less frequently than 1 time per million, only 4336 hieroglyphs will remain. On the basis of such a set of characters you can already try to form a set of necessary graphic elements. But for this we need to imagine what we want from structural hieroglyphics in general and from this set of elements in particular.

The first and the main task of structural hieroglyphics is the identification of hieroglyphs. Herewith, we can, and should, strengthen the requirement for identification, to make it *unambiguous*. Whatever we do, our actions must have an unambiguous result, namely, one and only one character. The structural hieroglyphics describes graphically unique objects, therefore, the description of each object should preserve this uniqueness; a one-to-one correspondence between the hieroglyph and the sequence of its components should be established, as there is a one-to-one correspondence between the character and the sequence, the mutual arrangement and the mutual dimensions of its constituent strokes. If we break this correspondence when writing, for example, it is clear that "at the output" we get a completely different object. It is enough to place the "falling rightwards" with a small indent from the "falling leftwards" and instead of "a person" (人) we get the number "eight " (八). The model of one-to-one correspondence between the character and the method of its "handwriting on paper" is a good example for structural hieroglyphics, but this model is too complex to implement, requires a large number of parameters: the sequence of strokes, their mutual arrangement and mutual dimensions. Structural hieroglyphics claims that the indication of a simple sequence of components should be sufficient for the unique identification of the character. In other words, a properly composed sequence of components is able to uniquely identify each character within the scope of the set of such characters. It is clear that this requirement imposes certain restrictions on the list of required components. This list should be *necessary*, i.e. on the one hand, it should not contain components without which it is possible to do, on the other hand, the removal of any component from this list can make the description of some characters with the help of the remaining components insufficiently effective, and some of them impossible at all. At the same time, the whole set of components should be *sufficient* for unambiguous description of any character from the considered set, and any character should be divided into component completely, i.e. it should not contain anything other than the specified components from the above list.

From general considerations, it is clear that the simpler the components selected, the longer the describing character sequence will be, especially for graphically complex characters. Therefore, the use of a long and well-developed system of strokes for the identification of hieroglyphs is weakly effective. It should be noted that in view of the peculiarities of the "two-dimensional" writing technology, the division of the character into strokes before the appearance of printing was the only possible way of their "input". It is also should be noted that the memorization of the character as a sequence of strokes (namely sequence, not set) is one of the most difficult moments in mastering hieroglyphics and it is not possible without involving outside techniques, the main one of which is certainly the use of "muscle memory of the hand".

The most interesting from the point of view of "candidates for components" is the well-known above-mentioned set of Kangxi radicals, long used by tradition for other reasons and for other purposes. Even a quick look at the traditional set of radicals indicates a good graphical "coating" of characters, i.e. the vast majority of characters are naturally decomposed into radical-components. Thus, there is a direct reason to follow the tradition. Therefore, as a basis for the set of components, one of the many (and not completely matching) sets of hieroglyphic radicals was selected [12]. Herewith, some components have been removed from the set, and some of the missing components have been added. Some of the radicals are extremely rare, for example, radical 龠, which is also a separate character, in the Lancaster corpus of the Chinese characters is found with very low frequency: only 2 times on more than 58 million characters as a separate character, and never as part of other characters. In addition to traditional Kangxi radicals, there are "non-radical" components in the characters' composition, i.e. those which have no graphic similarity to the traditional Kangxi radicals, and which, nevertheless, quite accurately and unambiguously discriminated characters. For example, it's enough to look at the characters 乙 and 乞. The upper element of the second character is not represented in the list of Kangxi radicals, but it distinguishes this character from the previous one, which is exactly a radical in the traditional sense.

As a result of visual analysis of the above-mentioned Lancaster corpus of Chinese characters [11], a set of components was selected in the amount of 259 pieces. With this set of components, it was possible to describe 6839 characters (with the exception of a small number of single inclusions of "garbage" present in the original, which I could not recognize). In this case, each character is uniquely described by its sequence of components, unique within the boundaries of the considered corpus of characters. This initial set of components included all of Kangxi radicals, all of their variants, and a small set of non-radical components. But such a number of components are inoperable in specific applications. The selection matrix is too large to be memorized, or even to be placed on the screen of a digital device.

Let's turn to tradition again. The writing and reading brethren for centuries working with the radicals, constantly faced with the problems of reflection a limited number of radicals on a much larger field of characters, forced to constantly transfer their knowledge to the next generation, of course, already found a way out. In fact, they have carried out a kind of "initial formalization" of graphic images of radicals. It is quite clear (it is "obvious") that all three elements of the character 众, which differ from one another, are variants of one component – 人. The limited space for the character input led to the need to develop some calligraphic methods of depicting the components that preserve their commonality with the primary source. Most of these calligraphy methods are reflected in Appendix 1, which is still to be discussed. And if in the example with the character 众 this is obvious, then in other cases there are graphical variants of the same components witch are not obvious. The upper elements of the characters 少, 当 and 兴, for example, the tradition considers as the variants of the radical 小 from "truncated" to "turned upside down". The idea of "variants" is the primary unconscious and "not formulated formalization" of graphic images: we see one thing, but keep in mind the other one. And nothing prevents us from using this kind of formalization, but not for the purposes of remembering, interpreting or indexing the characters in the dictionary, but for the purpose of formal description of the sequence of the components of the hieroglyph. Undoubtedly, we will transgress the visual match of the components of the character and its description, but this centuries-proven technique will allow us to significantly reduce the number of components for the formal description of the component sequence of the characters. Another advantage of the idea of formal reflection, or *coding* some graphic units with other graphical units, will help us in eliminating some decomposition collisions. If we reject more or less obvious visual match of the parts or elements of the characters and the graphic units describing them, we can, for our purposes, introduce some "useful fictions" that do not exist in the outline of the characters, but for one reason or another they may be present in the description of the character as one of the equals with other components.

When I talked about the direction of decomposition as a panacea for the need to indicate the relative position of the components of the character, I was not accurate. This happens in most cases, but not always. There are situations when the rule of direction of the decomposition is insufficient. As an example, examine the hieroglyphs 员 and 呗, the sequence of components of which according to the rule of direction of decomposition is the same: 口 and 贝. In order to eliminate uncertainty in such cases, at the end of the complete sequence of components of the rarer character (the frequency of use of characters can be taken from the Lancaster corpus [11]) is added "empty" component, which does not exist in the graphic image of the character, *pseu-*

docomponent-modifier "*". This empty pseudocomponent-modifier is added to the description of the character to provide the uniqueness of the sequence of its components. The components sequence of the more frequently encountered character 员 (yuán) will be represented by the string "口贝", and the rarer character 呗 (bei) by the string "口贝*". Pay attention: "rarer", - it is important for the subsequent development of the system of input of characters in real texts, so often used characters are entered faster and easier to remember. This pseudocomponent-modifier "*" can be used to solve other collisions, for example to distinguish between characters 鸟 (niǎo) and 乌 (wū), a sequence of graphs which, respectively, will look like "鸟" and "鸟*" as the character 乌 (wū) is used less frequently. Pseudocomponent-modifier does not have its own graphic image in the composition of the character, but fixes with minimal cost the visual difference between the two characters, if only when it is really necessary, as shown in both examples.

The idea of using variants of the drawing of radicals was taken directly from the traditional hieroglyphics, and, after some adjustment of the distribution of these variants by components, allowed to noticeably reduce the number of components used by the structural hieroglyphics to describe the characters. Given the empty pseudocomponent-modifier "*", a sufficient number of components were reduced to 200. (In the future, when we become familiar with realization of the composition input system, based on the laws of structural hieroglyphics, it will become clear why it is 200.) Since these 200 components in this context are neither keys, nor radicals, nor determinatives, nor phonetics, i.e. nothing other than graphic images, for reasons of terminological certainty we will call them *graphs*. I will give a stricter definition of the graph later, but for now we will simply assume that the characters are described by a sequence of graphs. Graphs are not the radicals and not the characters, although they can graphically coincide with those and with others. Graphs can coincide with individual strokes, such as graphs 丿 or 丶, have a "strange" outline from the point of view of traditional hieroglyphics as graphs 𠂉 or 刂, or do not have a graphical representation at all, as an "empty" graph-modifier "*". Graphs are *formal, graphically* *represented codes* of certain components of the Chinese character. This kind of formalization of graphic images, on the one hand, allows us to "load" the computer, make it quickly and efficiently process the information we need to identify (and then input) characters, on the other hand, allows you to not lose the clarity of the process, its ease of development and accessibility for any person who studies Chinese characters regardless of the teacher (or with his minimal participation, what was advocated by V.P. Vasiliev more than 150 years ago [1]). There is a strong belief that hieroglyphics is such a complex and special area that it can be mastered only imbued with the roots, influences and generally the spirit of Chinese culture, only becoming in a sense Chinese. Structural hieroglyphics allows us to study Chinese characters, to learn to read and write in Chinese without any need to "become a Chinese" in any sense.

1.2. Problem definition.

Structural hieroglyphics is the study of hieroglyphs as complex graphic images, perception, recognition and memorization of which has its natural regularities. An attempt to identify these regularities and use them in the study of hieroglyphics can significantly facilitate the penetration into this mysterious area of Chinese culture, in Chinese writing. Understanding these natural regularities is able to simplify and facilitate both recognition and memorizing characters, as well as, most importantly, to provide a simple and effective method of input characters, regardless of their pronunciation, to provide a simple and clear way of identifying characters for searching them in the dictionary, when the sound of the character is not known or even the graphic image of the character may not be clearly presented, may be recognized approximately as "something very similar to something already known", but "a little bit not that", i.e. may be recognized in the likeness, or in a number of similar characters only. More formally, the tasks facing the structural hieroglyphic can be described as follows.

1. On the basis of visual analysis of the basic set of simplified Chinese characters which includes the characters most used in the modern Chinese found in modern texts at least once a million, to develop a set of formal graphical units to describe the full composition of the characters, which would allow to:

- unambiguously identify each character from the mentioned corpus as such;
- as much as possible freely choose the characters from the mentioned corpus according to any preassigned set of its components;
- to form groups of characters in the likeness of their components or entire blocks of components.

2. Having a description of the full composition of all characters of the specified set, find a mechanism for simple, visual and effective characters input on any digital devices designed for input, processing, storage and transmission of text information.

3. To develop software prototypes that use the established principles, mechanisms and algorithms and allow not only to master the methods of input, but also to use them in real life: tools for memorizing characters, for search for them in dictionaries and for inputting hieroglyphic texts.

Having a fairly accurate idea of what we want to get, and why we need it, we can move on to more precise definitions.

1.3. Definitions.

The structural hieroglyphics based on the fact that:

(1) each Chinese character can be described by a linear sequence of separate *indivisible* graphical units from a given set of those, and each such unit can correspond a separate component in the composition of the character or can reflect modifications or arrangement of components in the composition of the character. These graphical units will be called *graphs* further. The structural hieroglyphics states also that (2) the graphs e*xhaustively belong to the set of 200 units* (see Appendix 1), that (3) this set of graphs is *necessary and sufficient* for the unambiguous description of each character

from the considered set, that (4) the graphs describe the character *completely*, i.e. the graphs completely cover the entire graphical variety of components of character, the character are described by graphs and graphs only, and the description of the characters does not include anything not belonging to the mentioned set of graphs, that (5) the description of each character is *unambiguous*, and this means that the sequence of graphs describing each individual character is *unique*.

The rules for describing characters (what in traditional hieroglyphics is referred to as "decomposition rules") are extremely simple and consist in the following.

1. The sequence of the indication of graphs in the description of the character corresponds to the composition of its components in the direction from left-top to right-down, for covering components (i.e., those inside which other components are located) — from outside to inside. For example, the character 国 is described by a sequence of graphs 囗 and 玉.

2. In case of ambiguity of the division of the character into components, and this can happen for graphically complex characters, priority is given to the variant with fewer graphs. For example, the hieroglyph 喜 can be described by two sequences of graphs: "十豆口" and "士口丷一口", — the first variant is preferred by this rule. This rule can also be confirmed by the first property of graphs, their indivisibility. The graphs are *atomic*, i.e. indivisible by definition, so trying to separate a component that is completely described by graph 豆 into smaller graphical elements will result in an error.

3. All *collisions* of the description are resolved by adding to the description of one of the characters of the participants of the collision an "empty" graph-modifier, which does not have a corresponding component in the character composition, designated by the "*" sign. We will call collisions situations when according to the first two rules it is not possible to obtain a unique sequence of graphs, i.e. two or more characters have the same sequence of graphs (these situations are exhaustively reflected in Appendix 2).

These simple rules are enough to build a table describing a given number of characters (Appendix 3).

As I said, the basis for analysis of the composition of Chinese characters was a set of 4336 characters. The simple extension of this set to the full Lancaster corps in 6839 characters did not lead to a change in the principles of coding components by graphs or the set of graphs themselves. It turned out that the set of graphs and the principles of describing characters by graphs remained the same, despite the increase in the set of characters by more than half. And it is encouraging, because it shows the presence of some hidden regularities of perception of complex graphic images, which should have been guided, most likely, unconsciously, people who have used such writing systems for centuries. This may be evidence that further expansion of this approach to more powerful sets of characters will not lead to fundamental changes in the mechanisms and algorithms of structural hieroglyphics. But, however that may be, we can say with confidence that in the space of the most frequently used characters, these regularities work and bring real benefits.

The main result of the structural approach to hieroglyphics is the construction of the table of description of characters, which in turn is the base of applications using the formal description of the graphic structure of characters for applied tasks, primarily for the input of Chinese characters, for the construction of electronic dictionaries with the possibilities of many times exceeding those of traditional flat dictionaries, for the creation of applications that facilitate the process of memorization of Chinese characters.

1.4. Characters table.

Having a complete table of the description of the hieroglyphs of the considered set, one can start to engage in "entertaining arithmetic". And at that time, interesting things are being clarified (see Appendix 4).

The number of graphs in the description of each character varies from one to eleven. The maximum number of characters (1574 or of 36.23 %) consists of three components (more precisely, it is described by a sequence of three graphs). The proportion of hieroglyphs that has from one to three graphs in their description, in total amounts to 61.74%. As we remember, the sequence of graphs retains its uniqueness, therefore, for the unique identification of more than half of the hieroglyphs only three graphs are enough. And if we take into account the frequency of occurrence of characters in real texts, it turns out that in 82.25% of cases three graphs are enough. Strictly speaking, this was the starting point in an attempt to create a simple, effective and visual system of input characters, for the mastery of which nothing is needed, except a bit of attention, normal memory and some perseverance.

But what about the rest of the characters, which consist of a larger set of components? The analysis of the complete sequence of graphs for all the characters shows that in fact, the description of the characters we use carries some "information redundancy", and we may assume that at least for some characters, there are sets of graphs smaller than the complete sequence, but can identify this character uniquely. It turned out, for example, that the character 骤, which is described by a sequence of 9 graphs (马耳又丿丨丶丿丿丿) in the whole corpus of the considered characters, is uniquely identified by a sequence of two graphs : 马 and 耳. In other words, in this set of characters there is no other character, in which the graphs 马 and 耳 would meet together. In fact, it was found that for each of all 4336 characters exist such short sequences of graphs and the size of these short sequences do not exceed three graphs (see Appendix 3). And this is important: there are short sequences of graphs that do not exceed three graphs in size and at the same time preserve the uniqueness of the description of each character. Let us call such short sequences of graph *markers*. Theoretically, to input any predefined character, it is enough to input its marker, which, as we can see, consists of a maximum of three graphs. Moreover, 72% of the characters have markers consisting of one or two graphs. If we take into account the frequency of occurrence of characters in real texts, it turns out that in 82 cases out of 100 when you input the text, it is enough to specify not more then two graphs. In other cases, you will need three graphs, and never more. In fact, this is quite an unexpected result. In the following discussion, when analyzing the existing input methods, we will see that so far no one method has such efficiency.

2. Structural hieroglyphics. Application area.

In order to understand why this works, it is useful to consider the process of recognizing complex graphic images. One very important explanation regarding the content of this section should be made here. I am far from a strict scientific description of the processes of nervous activity in the recognition of graphic images. I can base on general biological ideas about perception and on my own experience only. And everything stated in this section is not the result of a scientific experiment, but only the result of careful observation of myself and my own feelings in the process of images recognition, it is my own characters recognition practice. For a man who is far from child age when the ability to direct and spontaneous images perception has long been lost, when any *image* has long been tied to the *name*, and consciousness operates with names with more satisfaction, return to the recognition and memorization of images is quite a difficult task. So difficult that it is not solved by itself, requires some effort and time. The most important thing is that this process is extended in time. So, you can watch him. And the inevitable multiple repetition allows you to observe the process in detail. I believe this is acceptable in the introduction of a new area, which, of course, expects to participate in its development of professional researchers in the field of human physiology, psychophysiology, pedagogy, linguistics and other areas for which knowledge of the material and methods of their science is much more useful than unprejudiced, but superficial look of the dilettante.

2.1. Recognition of graphic images.

The very first and the most difficult part for the analysis of the perception of complex graphic images is the recognition process itself, fixation of the graphic image as "already understood", the inclusion of it "in the processing" as a separate object, a certain invariant entity that can enter into associative relationships, processed as part of the graphic-sound associations that arise when learning characters and used when reading characters aloud, when after the appearance of graphic-sound associations arise sound-semantic associations, or as part of the graphic-semantic associations when reading to oneself. The matter is complicated by the fact that the processing of graphic images is such an ancient mechanism of information processing that it is evolutionarily "hidden" in the very depths of the subconscious. In other words, the perception of graphic objects takes place in the subconscious, and we cannot consciously influence the processes of image recognition. This, of course, is a drawback, but there is one very important advantage. It lies in the fact that the processes of recognizing graphic images, since they occur in the subconscious, if they go, then go catastrophically fast. It would be unwise not to take advantage of this.

Fortunately for us, we know in general terms how the brain processes complex graphic images. Fortunately for us, for many centuries of study and use of Chinese characters people, unconsciously using a cultural analogue of natural selection, when non-viable structures simply died away, implicitly brought the structure of characters in accordance with the natural regularities of perception of graphic images, in accordance with the ways that allow them the most easy and natural to remember, reproduce and convey the meaning of graphic images.

The first thing that should be noted here is that the brain stores the graphic image as a *whole*. And this is important. This is the basis for the use of cards when memorizing hieroglyphs. The student sees for a while the hieroglyph as "a picture", and after a while he is presented with its meaning or pronunciation. All that is needed, in order to remember the image, is watching on the picture during the necessary time. In this case, the activity of our brain is only try to remember this image. And the hint of its meaning or pronunciation creates only a graphic-semantic association, nothing more, and does not help memorizing the image itself. Fixation of the image in memory occurs spontaneously, by itself, regardless of our convulsive activity : "remember, remember, remember". But we know that we remember the image as a whole, and therefore we quite easily recognize the character, although at first it is not possible for us neither to reproduce it on paper, nor even imagine it in all the details in the mind. But when we see it, we recognize it immediately, instantly. That is why it is easier for us to remember the hieroglyph and recognize it in a series of similar ones. That is why phonetic input systems accelerate the input of characters in comparison with handwriting, but do not help memorize the character, because in response to entering the syllable they answer with a list of characters that *sound* like the desired one, and not *look* like it.

The second thing that should be noted here is that for the subconscious, *the similarity is the absence of differences*. And it allows us to understand how images recognition happens. Our brain finds the similarities, noting the differences. Where and how does he look for differences? Look at these pictures: 猴, 侯, 喉. You will immediately notice that these pictures are different and they differ in their first elements. And then, on second stage, You understand, that they are very similar by its last element. Therefore, we can see, two things: our brain analyzes the images by dividing it into some components and examines them sequentially, one-by-one. And the direction of this sequence of examination is determined by our cultural habits. (I suppose the Arabs, for example, can see the elements in reverse order, but it does not matter, we use direction from left to right or from top to bottom.) It's very simple: 什, 化, 代, 位, 们.

And what happens when there are many elements, and they do not fit into a linear sequence? If they are located simultaneously in two dimensions: left and right, top and bottom? 媛 or 暖, or 暧. Our brain is not inclined to complicate your life without necessity. On the first "picture" we see "女" and "something else" without having the names of neither "女" nor this "something else". On the second "picture" we see "日" and again the same "something else". It's enough now. But. On the third picture we see the same "日", and something, that is not "previous something", but very similar to "previous something", and it differs from previous by one little element inside in this "something": "爰" and "爱". OK. We remember, that "the similarity is the absence of differences". And brain seek for differences and agrees to a similarity if it did not find a difference. Where? Outside first, along the periphery of the image, at the beginning and at the end of the usual sequence, and then inside the picture only. (We have seen this in the examples.) Then our brain focus on this differences, memorize it and use it as a *distinctive element*, as a "label", for this images. But, pay attention, as a label that distinguishes only these two specific images. And then we understand why it is easier for us to distinguish and remember these two images in comparison with each other, because this is a natural way for our brain to recognize images. Thus, the first element (日 or 女) distinguishes 暖 from 媛 , and a small inner element (一 or 冖) distinguishes 暖 from 暧. But if there are some elements that distinguish one character from another, we may assume that for each character may exist a set of elements that distinguishes this character from all other characters from this set of characters. In this case, for our set of three characters, we can say that for the first character (媛) such element is 女, for the second (暖) elements 日 and 一 , for the third (暧) elements 日 and 冖. These sequences of characters components are their markers, which we have already found earlier analytically, without referring to the mechanisms of their appearance (see section 1.4). Remembering that the sequences of components of characters are described by a sequence of graphs (see section 1.3), we can state that *a marker is a unique sequence of graphs that distinguishes a given character from all other characters of a given set*. Now the most important thing for us is that the division of a complex graphic image into components, their sequential analysis and selection of markers for each graphic image is a natural way to recognize complex graphic images. It should be remembered that the graphic image is stored in memory as a whole object, not as a set of components. This kind of "graphical analytics" is needed only in the process of image recognition.

2.2. Fixation of graphic images in memory

The next important question is the fixation of graphic images in memory, a separate and uncontrolled by the consciousness process. It must be borne in mind that this process occurs spontaneously, without the participation of consciousness, and all our efforts to influence this process must be reduced to creating an enabling environment for such fixation. For example, a noticeable effect on the speed of memorization is provided by the comfortable dynamics of the appearance and disappearance of the image. This dynamic is individual, it must be empirically "adjusted" to perception, it must be synchronized with the process of perception. To remember the graphic image, you need to "see enough" it. At the same time, its appearance and disappearance somehow stimulates memorization. The question is only in the optimal alternation and comfortable time of viewing the image. Someone needs to look at image longer, someone less, someone needs more sessions of appearance-disappearance, someone less. But in dynamics, the appearing and disappearing image is remembered faster. You may to read the Chinese character, while you see it, you may to memorize its translation, when it disappears, you may imagine its individual components, but you need to understand that the construction of a graphic-semantic or graphic-sound association is a completely different process, and it occurs according to its own laws, regardless of the fixation image as such in memory. The image is fixed regardless of its name or sound first, and the associative connections of this image arise after that only.

A habitual way of identifying images by a name: 讠 - "to talk", 己 - "oneself", - generates and the habitual way to control memorization. Images? No, graphic-semantic associations. The image itself in this learning process is fixed in memory along with the development of graphic-semantic associations, in parallel and independently, fixed as a separate entity, an image separately from the name. Character 记 only at the first stage of learning is perceived by the analytical mind as a collection of names "to speak" and "oneself," but eventually, naturally and spontaneously, 记 becomes a totality of images 讠 and 己. The image "lives" in memory separately from its name, and the associative connection only unites them. The release of the old artificial and hammy association "remember — to speak oneself" leads to the emergence of a new graphic-graphic association "记" is "讠" and "己", and a new graphic-semantic association "记" – to memorize".

Training such a "nameless" fixing graphics is facilitated by several techniques. When we talked about the natural regularities of graphical image recognition, we paid attention to the fact that consciousness fixes the similarity only as the absence of differences, that's why it's simplest to remember characters in a series of similar ones, clearly fixing these differences of similar images: 瓣, 辩, 辫, 辨. This natural attention of our consciousness to the small differences of such objects is manifested in the fact that it is easier for us to guess the desired character in a series of similar ones than to reproduce it on paper, or even imagine it in the head, so to speak, "visualize" it in the mind. Then will come the ability to "see" the character in the head, just as to count in mind, the ability to decompose it into components from the standard set, which also need to "see enough", to be able to see them in the unknown characters. No wonder in the absence of training after some time in the analysis of the composition of the new character, the main error that leads to an unsuccessful search for such character in the dictionary is an attempt to decompose rare (and atomic by definition) components, which, due to the rarity of their use, tend to fly out of memory.

Therefore, it seems that when studying (especially initial) hieroglyphics it makes sense to make a bias towards graphic rather than thematic organization of educational material. By the way, the Chinese pedagogy clearly used this in the teaching practice. It is enough to look at the original Chinese materials for primary school back in the 80s of the last century [13]. Unfortunately, the new "schools" of Chinese language learning, which have flooded the Internet in recent years, rely on the "novelty" of the material, orient students to "modern Chinese" and, under the pretext of the fact that "now they do not say so", complicate the already difficult task of mastering hieroglyphics. No dialogue of today's teenagers will give so much for the mastering of hieroglyphics as a short fairy tale of the "小子下下山" [ibid.].

When teaching composition input to stimulate the process of "separating the image from the name" in the tools for memorizing graphs and characters (see section 3 below), the "picture mode" was specially added, when system as a question offers not a meaning, not a pronunciation, but an image of the graph or a character itself and waits input. No need think, no need remember, just repeat what you see. It's a some kind of graphical "voiceless dictation". And it turned out that to search for a graph on the matrix or to reproduce the proposed character, the time is spent almost 2 (!) times less than in traditional training modes, for example, on flash cards, when the meaning or pronunciation of the character is proposed for testing. Attempts to introduce elements of "rational comprehension" ("image-name-image") slow down the process of image recognition. Roughly speaking, the hand (with the mouse or with the touch screen) works faster than the head.

The image is fixed in the consciousness before its name or meaning or pronunciation. And this time lag is clearly visible in the process of fixing the image: it is immediately recognized as different from the others, but does not yet have its name. At first, it is discouraging and annoys: "Why I can't memorize it?" But that's normal. This means only that the image is already fixed in the subconscious and you needs only an associative connection of this image with its meaning or name. And as we already know, this is a separate task with its own characteristics and methods of solution. The absence of an image name in this conjunction is especially indicative when it is necessary to reproduce an unknown character some time after you met it first. Experience shows that it is extremely difficult, almost impossible to remember the character through the names of its components, as in the example with the character "remember" (记), but the graphic image itself will emerge in the mind, and it turns out that seen 2 - 3 hours ago somewhere on the signboard character can easily be found in the dictionary.

The composition input algorithm is organized so that the user sees the only character when the machine has already identified it. And at all intermediate stages of identification the user sees hieroglyphs chosen from the table by one or another principle of similarity, i.e. almost always surrounded by similar ones. Separate algorithms are specially designed to ensure that it was not accidental to form lists of characters, but purposefully select those have a certain degree of similarity: by initial components, by the end components , by any of the pre-defined components in any order, to identify frequently occurring blocks of components or to detect the occurrence of some characters in the other. Such work has always been done, but it was a "piece", painstaking work associated with huge labor costs, with large sets of cards and images. Now this task with the help of computer database management systems is performed by structural hieroglyphics "on the fly".

In preschool and early school age children calmly and effortlessly memorize images. In adulthood, a person, acquiring the ability to analytical knowledge, gradually loses the ability to direct visual perception. At first, our mind, spoiled by the habit to analytics, rebels, refuses to remember what does not understand. But it can be deceived. It is necessary to tell him that he understands and he stops interfere in this matter, and even begins to help in something. Being included in the composition of all sorts of "hammy associations", he thinks that he "analyzes" images. Let him do it, for at this time there are processes of memorizing images and organization of graph-semantic associations, which then will live in the subconscious. Let the analytical mind participate in the transportation of images to the subconscious, because there is a huge field of activity for the analytical mind, for the already conscious memorization of the meaning of the "ready" graphic image. Traditional mnemonics, associated with the allocation of the radicals in the character, with the division of the character into determinatives and phonetics, is forced to use the structure of the character, implicitly admit the presence of the essential characteristics of its structure. Inventing any kind of arbitrary links of structure with some of their ideas, which often have no relation to reality, or had them once, but have long lost them, they latently create useful associative connections. It should be understood that graphical and semantic associations are a fact of consciousness, not of reality, and for our consciousness their durability is important, and not the "correctness" of their occurrence. Often "wrong" associations help to remember the object easier, faster and stronger. There are many examples of such "wrong" associations in textbooks and language manuals, and even more may be created by yourself (see section 2.3 below). And it is more useful to do it yourself, because independent activity of consciousness is the main engine and inspirer of the process of memorization. With regard to structural hieroglyphics, it can be argued that the explicitly formulated principle of the composition of characters, the principle of combinatorics of components is the best help for the activity of this kind. It "unleashes hands" for this kind of work of consciousness, removes all the restrictions that are set, for example, by the need to divide the character into phonetics and determinative, it is this phonetic and this determinative, long and firmly set by tradition. Or to allocate this radical in the character. The freedom of associative creativity is much more useful than the habit of remembering "correctly".

2.3. Fixation of graphic-semantic associations.

On the next stage after fixing the image you must "denominate" it, set relation of the image and its "name" or meaning. Extraordinary polysemy of Chinese characters is able to discourage anyone. But it may help. A cluster of similar meanings is easier to remember than a single meaning. Look for abstract meanings for such clusters.

Character 嘴 – it is ideal for demonstrating such an abstract approach, because in fact it is any "protruding part": mouth, beak, nose, mouthpiece, nozzle, bottle neck — something long and protruding. 瓶 – is any vessel with a narrow throat: bottle, vase, pitcher, flagon.

The nonrandomness of the structure of Chinese characters and the nonrandomness of a set of their meanings, worked out over centuries of use, sometimes amazes with its logicality, sometimes discouraging by surprise – both of which help memorization of characters. Study the field of meanings, invent "meaning labels" for them.

I repeat, this has nothing to do with the structure of the characters. You are free to do whatever you want, to divide the characters into any elements: radicals, phonetics, determinatives, just into separate "pictures", or not to divide them at all. Use polysemy of characters, use results of your own artificial divisions of characters, make artificial, "hammy" associations: male beast and Duke – 公(both sound the same – gōng), a monkey 猴(hóu) and Marquis 侯(hóu)*. Nobody says that the Chinese called Duke "male beast", and the monkey is the "bestial" Marquis (犭 + 侯). But, by analogy with our "king of beasts", the lion 狮 is very similar to the "bestial" teacher (犭 + 师)**.

看 – the hand (手) above the eye (目), the man looks into the distance — a classic case. The basis of the example in the widespread mnemonics of the semantic load of radicals. Such associations you will find anywhere and as much as you want. In this case, the radical is "an eye" and "a hand" is not the radical. Why not vice versa? We will not discuss the "correctness" of this kind of association — the graphs are not the radicals, but what do we care about it now? Do they help you? And thank God, let them work. Hammy association "to look – a hand and an eye" makes life easier, and this is the most important, and the most interesting thing is that over the time and completely invisible to you with enough diligence and multiple repetitions, it will be replaced by a normal graphic-semantic association "看 — to look". Moreover, if you seriously consider mastering the matrix of graphs, then when you reproduce "看" it will not be "a hand and an eye" but "手" and "目", and "to look" will not "put your hand above your eyes," but simply "看" as a graphic image of the meaning of "to look".

This trick is possible not only with the radicals. Character 掰 - just the same puzzle: divide (分) between two hands (手), - "to break". there are two participants:

* – Count and Marquis is a very arbitrary translation of the name of the first two levels of the nobility of ancient China.

** – 犭 is a variant of the classic radical 94 (犬) commonly called "a dog" and meaning a four-legged animal in general.

one is not radical and one is radical. For the ability to read, and then for the ability to write, it is only important that the Association of the image 掰 with the meaning "to break" is firmly entrenched in the head.

Another example without the participation of radicals in general. 保 - a child's drawing of a man with a spear that stands and guards, protects something. There is no any connection with the meaning of radicals ("a man", "a mouth", "a tree"), there are no any "radicals", and there is just a picture like a child's drawing, which connects the image and meaning, which creates graphic-semantic association "保 — to protect".

2.4. Hieroglyphic dictionaries and search.

The ability to freely combine components eliminates the need to "correctly" select the key of the character to search for it in dictionary, there is no need to analyze and remember the exceptions and illogicality of choosing a radical for one or another character. Free combinatorics of components allows you to select from the table of description of characters all the characters containing a particular component, choose from the table all the characters containing any conceivable combination of components, regardless of whether these combinations are "correct" or not from the traditional point of view. The implementation of this kind of combinatorics is easy to shift on the shoulders of the computer and save the user from the routine of finding the result. This greatly facilitates the life of a particularly novice student, because it does not require any additional skills, except the habit of seeing these components as part of the character, which is easily and simply produced by practice and does not require a long process of support by the teacher, constantly "directing" and "correcting" the student. This does not require sophisticated algorithms, such as for the implementation of handwriting, and enough features of any very simple database management system. Technically, it is quite simple to add this kind of search to any existing electronic dictionary. At the same time, the ability to search for characters in the dictionary is simplified, and the search speed is increased by a multiple.

The naturalness of perception of characters as a set of components and the ease of mastering make it possible to begin studying hieroglyphics from the very beginning of learning. Moreover, the structural hieroglyphics allows not only to search for characters in the dictionary and thus to understand hieroglyphic texts, but also to write hieroglyphs without the generally accepted endlessly repeated writing them on paper. The structural hieroglyphics allows you do not waste time and paper in vain, and write quickly and clearly from the very first attempt. The reality of this life is that the manuscript, no matter how well it was done, requires input it on a computer (or any other digital device). Serious texts are written and transmitted by computer, not by notes on paper. So, the question boils down to the fact that on the basis of the principles of structural hieroglyphics it is necessary to develop an practical (simple and effective) input method.

3. Chinese characters input methods

There are three basic (in terms of prevalence) groups of Chinese characters input methods on digital devices.

3.1. Handwriting.

Handwriting method completely emulates the writing of a character on paper. Old and wise technology. The Chinese character is uniquely determined by the sequence, size and relative position of the standard strokes. The machine uses special and rather complex algorithms to "recognize" the entered character and offers the user an variant (or variants) to insert into the appropriate position in the input waiting program (the so-called focus). It is not necessary to talk about the speed of handwriting, because the number of strokes in the character can reach several tens. This is a delicate work: you need to follow the order of input, mutual arrangement and mutual size of the strokes. Of course, in advanced systems, the variants are offered even when the user has not yet finished drawing, advanced systems of handwriting recognition suggest the algorithm of adaptation to the peculiarities of the input of a particular user, but, in general, they can be used only from the inevitability, when there is nothing else. 人 – it is very simple. 犬 – it is simple. 狗 – it is already much more complicated. 猴 – ask any beginner to write in Chinese a simple word "hóu" (monkey). And what can you say about the character 藏?

It is clear that the speed of such handwriting of characters is not much higher than the speed of writing characters on paper. Old and wise technology, but slow and time consuming. The advantage of handwriting is only one – no new skill is required for those who have already completed a course in writing on paper.

3.2. Phonetic input.

In all types of phonetic input, the character is entered by its normative sound (for example, in Beijing dialect, Cantonese dialect, etc.), which is written either by Latin letters (pinyin) or Cyrillic letters (palladium), or letters of the Chinese phonetic alphabet (zhuyin/bopomofo). There are other ways of recording the pronunciation of hieroglyphs, but they are not widely distributed and essentially do not differ from the above. The most common phonetic input by pinyin. Among the main disadvantages of phonetic input are the following three.

1) There is no one-to-one correspondence between the character and its sound. A single character can be pronounced in different ways (have up to 8 variants of reading), a single syllable can correspond to several dozens of characters. There are ~25 thousand characters in the main CJK UNICODE blocks of the latest version v.10. All of them are "vocalized" with help of 1,314 Chinese syllables only, on average one syllable corresponds to 19.3 characters; only 57 syllables are recorded by one character, and the syllable "yì" (yi4) corresponds to 449 characters. Therefore, there is a problem of choosing the desired character from all the suggested characters in response to the input of the pronunciation. It is clear that the above choice of desired character, reduces the speed of input, required additional attention and additional manipulations from the user.

2) After the identification of the character is completed, even if this identification is made unambiguously and no choice is required, the phonetic input methods require mandatory manipulation (pressing a key on the keyboard or mouse button, or touching the touch screen) to fix the character in the focus. In other words, during the input process, after the identification of each character, the machine necessarily "asks" the user whether he agrees with the choice, even when this choice is unambiguous and does not imply any other variants.

3) Phonetic input is not possible for characters whose pronunciation is forgotten by the user or is not known to him at all, which often happens, for example, when searching for unknown characters in the dictionary.

3.3. Shape-based methods.

To eliminate the shortcomings of phonetic input, one can try using shape-based input methods. These methods, since they are based on the analysis of the graphic structure of the character, are not related to its pronunciation, therefore, allow to immediately eliminate the main lack of phonetic methods, i.e. allow you to enter characters with unknown pronunciation.

First of all, it is worth noting the methods of entering characters by key or radical, the key-based or radical-based methods which completely emulate the search for a character in a "paper" dictionary. Characters in many paper dictionaries are indexed by radicals, radicals in this index are arranged in order of increasing the number of strokes. Because radicals are used to index characters, they must be separate, easily identifiable, frequently occurring elements in characters. Most often, the radicals are located "at the beginning" of the character: on the left or on the top (as a consequence of the fact that the originally hieroglyphics texts were written from top to bottom). Some of the radicals are simple strokes, some are independent characters, some of the radicals have variants, sometimes significantly different in shape from the main radicals. In general, the variants differ from the radicals in that they in this outline can not act as a separate character, and are found only in the composition of other characters, keeping their radicals meanings, i.e. the characters, which include these variant in the index located in the same group of characters, as the characters with the main radical in its composition. It is clear that the set of radicals is tied to a specific dictionary, and in reality the number of radicals varies from dictionary to dictionary at the will of the compiler, and can reach together with their generally accepted variants of almost three hundred pieces. As was said in the first chapter (see p. 8), a set of radicals which is considered as a classic is Kangxi radicals that exists from the 18th century and composed of 214 radicals. It should be borne in mind that the radicals traditionally have a semantic component: each of them has a "name" and is identified with a certain group of phenomena, actions, objects. This semantic component of radicals is widely used in mnemonics to remember the meaning of specific characters containing a particular radical. Thus, the study of radicals is now a mandatory element of the study of hieroglyphics in general.

Radical input methods use either a virtual keyboard on the device screen, or even handwriting input. In the structure of the character, the only component is identified – its radical, the remaining part of the hieroglyph is characterized by one aggregate parameter - the number of strokes of the character except for the strokes of the radical itself. Thus, the non-radical components of the character do not matter: the radical is the basic identifier of the character, the others are not important. Coincide or not they (the remains) with other radicals combination or part of them — is irrelevant. The user needs to know the radical for each character and remember (or calculate in mind) the number of strokes in the remainder. The number of strokes of the character, excluding the number of strokes of the radical, is the second required parameter of the search (see, for example, a patent [14]). On the virtual keyboard described in this patent for radical input, they are located on increasing number of strokes. The user initially sees only groups of radicals from 1 to 10 strokes (the last group includes all other radicals that have more than 10 strokes), and must specify the desired group, after which a list of radicals appears with a specified number of strokes, from which you also need to make a choice. The user selects the number of additional strokes in the remaining part of the character (the number of strokes of the character, excluding strokes of the radical). After that, a list of characters containing this radical and the specified number of strokes in the remainder is displayed. The user must select the desired character from this list. The method is acceptable for searching for character in the dictionary, but as a text input method it is not suitable due to the large number of necessary manipulations, the large number of candidate characters in each "radical-strokes" group and thus the low speed of such input.

There are also methods that use the idea of radical combinations (see, for example, [15]). This patent describes a method and device for the identification of characters in an ideographic alphabet that allow the user to graphically describe a character using a set of its components. The method uses a combination of 82 radicals. The source does not indicate how many characters can be entered using these combinations. The user directly on the screen using the drag-n-drop operations of the components places them in the desired sequence. The matrix (or "canvas") where the hieroglyph is collected is divided into 9 sectors, and the machine "itself " takes into account the relative position of the components, because it "knows" in which sector the operator placed a particular component. Therefore, this method for identification of the character in addition to the components also takes into account their mutual arrangement. But to describe all the characters with such a small set of components is almost impossible, so the degree of uncertainty in the input remains high. Indeed, in the description of this patent, the character 若 is identified by three components. But in the same way by the same three components are identified another 4 characters. With this set of components, the hieroglyph 苦 is indistinguishable for the machine from the hieroglyph 若, for example. Therefore, the identification is ambiguous and to input the character requires additional user attention and additional manipulation to select the desired character from the 5 proposed. The advantages of this method include the use of radicals as a natural and familiar to Chinese hieroglyphics division of characters into components.

The following group of structural or shape-based methods of input for definiteness we will call *the methods of structural coding*. These methods are based on the idea of coding the graphic structure of the Chinese characters with letters of the Latin alphabet. First, some rules of decomposition of the character into "standard", pre-defined components are introduced. Each of these components is placed in accordance with a certain letter of the Latin alphabet and is located on the corresponding key of the standard Latin keyboard. The user presses the keys sequentially, i.e. enters a certain sequence of letters, reflecting the sequence of the components of the character. The machine identifies the characters from this received letter code and produces a list of candidate characters associated with the entered sequence of components. The user must somehow specify which of the candidate characters should be inserted into the text. From the methods of structural coding, the most famous are *wubi* and *cangjie*.

The wubi method (see [16]) uses in different implementations from 204 to 227 components, almost half of them graphically coincides with the classical Kangxi radicals or their variants, some are represented by individual elements of radicals, the other part is represented by individual strokes (about 90 non-radical components, i.e. those that are not graphically similar to radicals). All these components are "tied" to 25 keys of the standard Latin keyboard, so each key corresponds to from 3 to 14 components, and, accordingly, from 3 to 14 components have the same letter code. The logic of components location on the keyboard, their binding to specific keys, has a complex structure and is not connected with the convenience of input, but with an attempt to facilitate the memorization of this complex structure of components. The components used are not equivalent. There are 5 "basic strokes" according to which the other components are divided into 5 groups according to the first line of each component, 25 "basic characters", which, at the same time, can also act as components (and then they are entered according to different rules than the rules of entering the basic characters), and ordinary components. Decomposition rules are quite complex and depend on the type of components in the character, there are four such types of arrangement. Decomposition of some characters requires taking into account the relative position of the components in the form of the so-called "difference code". Difference code is a combination of "component number", which is determined by the belonging of a component to one of the five groups of basic strokes and types of relative position of components in the character: left-right, top-bottom and mixed type. Since the basic strokes 5, difference codes for a total of 15, therefore, 15 keys, in addition to entering the component codes, perform the function of entering the difference codes, and these codes are the same Latin letters that encode the components. In some situations, during decomposition, another parameter becomes important - the "last stroke" of the component. To minimize the use of difference codes, the developers have introduced different rules for 4 different types of characters to define this "last stroke" of component.

When you enter components, the order of writing strokes in the character is mostly repeated, but there are some exceptions. Characters are not equal in the way they are entered, i.e. there are different input algorithms for different types of hieroglyphs. For example, to enter basic or "main" characters, you need to press the corresponding key several times (from two to four times). For input a part of the characters you just need to press the keys in sequence, corresponding to the components of the desired character. For some characters, you must also enter the code for the relative location of these components, the above mentioned difference code. For some characters, you must enter two letters and then add up to four symbols by pressing key L.

All these algorithmic tricks are required in order to obtain a unique alphabetic code for each character, for example:

士 — fghg; 十 — fgh; 人 — w; 员 — km; 呗 — kmy; 手 — rtgh; 金 — qqqq ... and so on.

The machine "works" with these codes, choosing the necessary characters as you enter the letters (analysis of the possibilities of the method is based on the table of correspondence of letter codes to characters published in the network [17]). Using this code requires up to four manipulations (keystrokes), and only 636 characters out of all possible can be uniquely identified after pressing one or two keys. In the table of hieroglyphs, the developers found it possible to add multi-syllables words consisting of two characters and whole phrases (up to 9 characters). At the same time, the code does not provide a unique identification of a character or a word, and 16.75% of the entire set of the given codes correspond to two or more (in reality up to 45) characters or words, and 44% of the characters (or words) are not uniquely identified at all, i.e. have codes that overlap with other characters, and after entering these codes an additional user choice is required.

The *cangjie* method (see [18]) is used 24 primary and 87 secondary "signs", for a total of 111 signs. 44 characters of these are graphically different from the Kangxi radicals. The primary characters coincide with the radicals, and some of the secondary ones graphically coincide with the individual radicals, some with their variants, some with a strokes, and a small part — just graphical components of the character, not represented in the previous groups. Of the 24 primary characters only 13 are the same as the basic wubi characters. Among the secondary signs exist those coinciding with the components of wubi, but there are different one. The principle of this method is, as in the wubi method, in the "binding" characters to the codes of Latin keyboard keys. 24 keys are used to indicate the signs. One key in cangjie codes from 2 to 8 signs. Even the signs that coincide with the wubi components are, of course, tied to other keys, since the principles of the layout of the characters on the keyboard are different in these methods. Decomposition rules of cangjie method seem easier than the wubi but in this method, we have to take into account different types of characters (single-units, two-units and three-units characters) decomposition of which is different from each other. Due to the fact that the number of characters used by the cangjie method is almost 2 times less than in the wubi method, and the keys for encoding is used one less, the length of the resulting alphabetic code for graphically complex characters becomes larger. To speed up the input (reduce the number of manipulations), the developers of the cangjie method added rules for excluding certain signs to reduce the number of symbols in the received alphabetic code of certain characters. Despite this, the maximum number of symbols (or pressing a key on the keyboard) when entering individual characters with the cangjie method still reaches 5, and only 238 characters are uniquely identified after pressing one or two keys. At the same time, 22.2% of hieroglyphs are not uniquely identified. For such not identifiable characters, the user is offered a choice of 2 to 8 variants. This, of course, is better than in the wubi method, but still there is no one-to-one correspondence between the code and the character (see the correspondence table for letter codes and characters [19]).

Both methods have the same disadvantages: complexity of decomposition rules, heaviness and sophistication of organization of letter codes for characters, respectively, and difficulty of remembering the necessary sequence of keystrokes for inputting characters. The lack of one-to-one correspondence of codes and characters requires additional attention and actions of the user. The limited code binding field (25 and 24 keys) increases the number of necessary manipulations, and the input speed drops. On modern devices with a small touch screen, the usual advantages of using the keyboard for input speed are lost, because the input on these devices, as a rule, involves the use of only one hand, while the second one holds the device. At the same time, the disadvantages of the components distribution on a small number of keys remain. This distribution is uneven, it is not focused on the ease of input of these components or their sequence or frequency of use, which leads to unnecessary movement of the mouse pointer or finger on the touch screen, respectively, to reduce the speed of input.

Decomposition rules of the structured coding methods have two aspects. The first is the aspect of the developer who needs decomposition rules to obtain the minimum possible unique code for each character. As described above, this aspect does not work well because of the lack of one-to-one correspondence of codes and characters in both wubi and cangjie methods, despite their complexity, abundance of groups and types, and many exceptions. The second aspect is the aspect of the user who needs to know this complex system of rules, conventions and exceptions in order to exactly repeat the developer's actions and reach the same result. It is clear that eventually the user, regardless of the decomposition rules, should simply remember the sequence of keystrokes for inputting a particular character. During text input, the user has no more time to manage with the decomposition. But the user does not have another method to get information about the desired sequence of keystrokes than to try to perform decomposition one way or another, as he cannot get help from the machine. In the actual input process, the user does not even see the components he has already inputted, but sees only a set of letters representing the sequence of keys he pressed, i.e. the input is performed "blindly". In both structural coding methods, all the information on decomposition and matching the components to alphabetic codes is hidden from the machine; all of it is concentrated in the user's head. As already mentioned, the machine "knows" only the correspondence of letter codes to characters; complete information on actual graphical composition of character components after such encoding is lost forever and cannot be restored. Therefore, it is impossible to expect aid from the machine in situations of uncertainty: either the user inputted correct sequence of codes and got the desired character at output, or he made a mistake and got an absolutely unexpected result. Therefore, the abilities of these methods to find characters, by analogy with the known ones, based on the presence of similar component blocks, simply on the selection of all the characters having a particular component at the beginning or end of the character, are extremely limited. The artificial manner of isolating components, the lack of visibility both in decomposition and in binding components to a limited number of keys make it difficult for the user to master these methods.

3.4. Composition input.

The composition input is intended to provide a simple, visual and effective input of the most commonly used Chinese characters. The most commonly used characters will be considered as characters, which are used in the modern Chinese language with a frequency greater than once per million characters of the text. The frequency of characters were calculated on the basis of the Lancaster corpus of Chinese characters, compiled by Tony McEnery and Richard Xiao in 2006 the year [11]. By input efficiency here will be meant the value inverse of the number of user manipulations necessary for unambiguous identification of a hieroglyph by a machine. The more user manipulation is required, the less efficient the input method will be. It is clear that the input efficiency directly affects the theoretical limit of the input speed, so all other things being equal, the achievable input speed will be higher in the method, the number of manipulations of which is less.

Structural hieroglyphics provides us with the primary material for machine processing, a formal description of the structure of the characters, and a complete, without exceptions and omissions, formalized to machine codes, while leaving the user with a graphical representation of these codes in the form of graphs. The user operates on graphic images, the machine processes a sequence of codes describing the characters.

The idea of compositional input is that we have the ability to select the characters from the characters table (Appendix 3) based on the graphs table (Appendix 1). For example, we can ask the machine to show us a character "consisting of" or "containing in its description" a single graph 人, we will get a unique character 人. If we ask the machine to display a character containing in its description 2 graphs 人 following one another, we get the only unique character 从. Adding to the query one graph 人, we get character 众. Since we have a description of the complete sequence of the components of each character, we can implement any conceivable query to the character table. In fact, we have unlimited opportunities to select characters, and the only thing is how we organize the selection algorithms, how to optimize them to find any character with minimal cost. Then the task of the user is to consistently enter the desired graphs. In accordance with the established mode of selection from the table of characters, the machine will give him one or another result.

3.4.1. Graphs matrix.

For graphs input by the composition input method in its current implementation (see [20]) uses a specially organized virtual keyboard, although theoretically nothing prevents implementing the input with a separate hardware keyboard. The convenience and speed of input depends on the configuration of the keyboard. It is not difficult to implement a keyboard for both ten-finger blind writing and input on portable devices with one hand. Existing prototypes provide both of these features. The matrix of graphs is realized in them in two form factors: 20 x 10 and 10 x 10 (Fig. 1, 2 and 3).

On the matrix of 20 x 10 graphs are arranged in accordance with the decrease in the frequency of their use from the center to the periphery. Since frequently used graphs are concentrated in the center of the matrix, minimal movement of the positioning device (mouse pointer, stylus or finger on the touch screen) is required for input. Simple calculations show that the graphs located in the 10 central columns of the matrix 20 x 10 provide input 90% of characters, and graphs located in the 10 peripheral columns (5 on each side), are needed when you enter only 10% of characters. For reasons of minimization of the occupied space these graphs can be hidden for the time and called only as necessary by a separate key, just as the usual *Shift* key on the keyboard causes capital letters. The arrangement of graphs on the main matrix 10 x 10 repeats the Central part of the full-size matrix 20 x 10, and the second part of this reduced matrix repeats the arrangement of moved to the center 10 side columns of the full-size matrix. Then, on the second part of the matrix, the principle of graph distribution by frequency of use is preserved, the relative arrangement of graphs is preserved and, most importantly, when moving from a full-size device to a portable one, and vice versa, the skill acquired by the user on another type of matrix is preserved. It is actually quite easy to remember such a terrifying matrix, especially since there is a simulator (see [20]), with the help of which an untrained user can memorize the arrangement of 200 graphs on the matrix in 8 - 10 lessons of 15 - 20 minutes only.

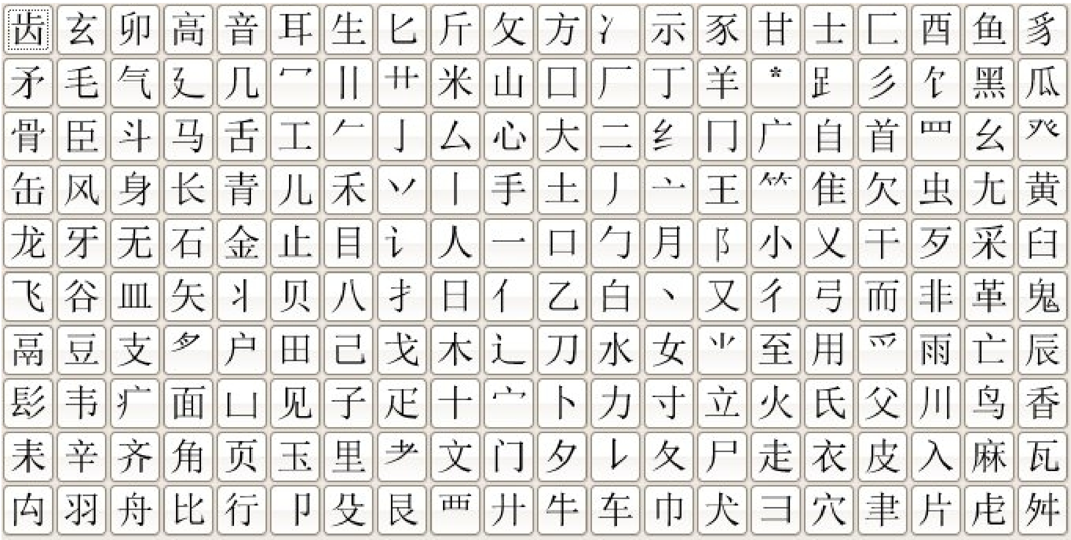

Fig.1. The full matrix of graphs of the virtual keyboard.

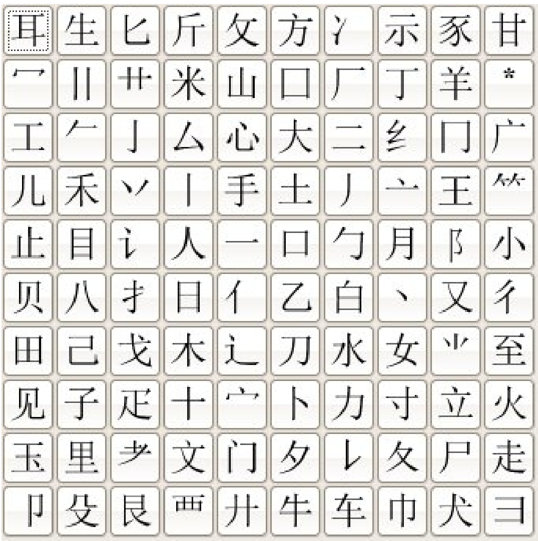

Fig. 2 The compact matrix of graphs of the virtual keyboard. Part 1.

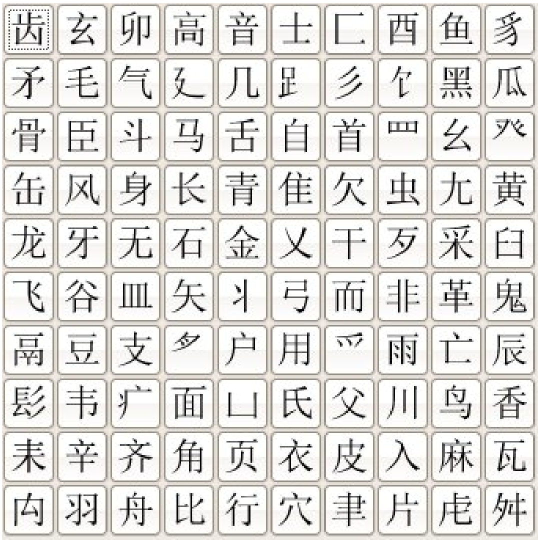

Fig. 3 The compact matrix of graphs of the virtual keyboard. Part 2.

The matrix of graphs is the basis of the composite input system and it is present in all prototypes of the necessary tools. These tools are currently implemented in the Java and H2 Database Engine, so they can be run on any device where the Java machine is installed. The tools were tested on Windows XP, Windows 7, Windows 8, Windows 10, Linux (CentOS-6.5 and CentOS-7.0) and Mac OS X. There are no licensing restrictions on the download and use of software products. The tools offered on the developer's website include a program for memorizing the graph matrix, a program for training "recognition" and facilitating the memorization of characters as a set of graphs, and a program for entering characters [21]. If you experience an unexpected difficulty in the composite input process, you can use phonetic input to resolve your difficulty. Phonetic input module is implemented for both pinyin and palladium input. A brief but quite clear description of the instruments can be found there [22]. It is clear that the method of composition input involves its use in any operating system as an input method, and when it is sooner or later implemented, then the implementation of all the above tools will change. Now it is important to pay attention only to the fundamental possibilities of realization of all the advantages of composite input, which provides us with structural hieroglyphics and to algorithms that allow us to realize these advantages.

3.4.2. Algorithms and selection modes.

The one-to-one correspondence of the character and its description by a sequence of graphs ensures unambiguous identification of any character from a given set. By sequentially inputting the graphs of a particular character, the user as a result of a query to the characters table will ultimately receive one unique character corresponding to inputted sequence of graphs. Herewith, as we remember, we can get a unique character with much less effort. For this purpose was developed a special algorithm for the identification of characters on the entered graphs. As the graphs are entered, this algorithm extracts the characters from the characters table, looking at the entered graphs in the following order:

1) one inputted graph is considered to be the first graph of the complete sequence of graphs;

2) two inputted graphs are considered sequentially as:

a) first and last graphs of the complete sequence of graphs or

b) first and second graphs of the complete sequence of graphs;

3) three inputted graphs are considered as:

a) first, second and last graphs of the complete sequence of graphs or

b) first, second and third graphs of the complete sequence of graphs;

4) four inputted graphs are considered as first, second, last and next to last graphs of the complete sequence of graphs.

Analysis of variants in this sequence allows unambiguous identification of all characters from the plurality considered in this way. The algorithm is constructed such that it returns a list of characters, in which the first position of the list is occupied by a character strictly corresponding to the inputted graphs, followed by characters similar to the first one in the graph composition, but requiring the input of additional graphs for identification uniqueness. Thus, the graph returned in the list of characters in the first position will be considered the uniquely identified characters in a given set of graphs. As a result of analysis of operation of this algorithm, it was found that to uniquely identify almost all characters consisting of more than three graphs, with the exception of 42, it is sufficient to take into account only the first, second and last graphs of their complete sequence. These sequences for each such character will be called fsl-sequences (from "first, second, last"), accordingly, the algorithm for isolating these sequences will be conditionally called fsl-algorithm. Among other things, "fsl" ("first, second, last") is a good mnemonics for the user, especially at the initial stage of training, which allows inputting unfamiliar character without knowing its marker yet (the markers are described in detail below in the context of selection modes).

The importance of fsl-sequences in character identification, as we now know, is hardly accidental. The development of Chinese character has been going on for centuries, and the experience of many generations in the memorization and use of characters has led to the fact that the general structure of characters as complex graphic images gradually, implicitly and unconsciously, was brought into line with the natural regularities of perception and recognition of complex graphic images: a holistic coverage of the image, its division into components, the transition from the outside into the inside. From the point of view of the naturalness of perception, the selection of the first and last component of the graphic image is absolutely inevitable. This is also confirmed by the fact that there are characters (40 characters of the 42 characters mentioned above) that cannot be identified by three graphs and are identified only by four: first, second, next to last and last, as if embracing the perimeter of the character and passing inside it. If distinctions between structures of two characters (in fact, two graphic images) are concentrated in their very center, then it is not possible to "get" to them through the description with the aid of the first, second and last graphs. The two remaining of the 42 hieroglyphs: 暖 and 暧, each consisting of 5 graphs (日⺥一手又

and 日爫冖手又), differ only in the central (third) graph and they are identified only by the first, second and third graphs.

It is also quite natural in the recognition of complex graphical images is the selection and fixation of certain "marks" (already known to us as markers), characteristic of these images, and allowing them to identify, without paying attention to the remaining details that are "insignificant" from this point of view. It has also been observed that selection of characters by fsl-algorithm often "works" earlier, after entering only two graphs - first and second or first and last, i.e. there is information redundancy, which can be eliminated. On the basis of the "early response" of the fsl-algorithm for each character, after a mandatory check for uniqueness, a minimal sequence of graphs was found, its marker, that preserves the one-to-one correspondence with the corresponding character. Since, in most cases, the fsl-algorithm responses on the maximum on three graphs, the size of markers also does not exceed three graphs. For the above mentioned characters, which are identified by the fsl algorithm only by four graphs, it was also possible to form markers from three graphs. Thus, for all characters of the set under consideration, markers with a size of no more than three graphs were generated, which allowed organizing input of characters of said plurality using no more than three manipulations. It should be noted that markers of all characters in the first position always include the *first* graph of the complete sequence of graphs. The second marker graph is either the *second or last* graph of the complete sequence. The third marker graph is *almost always the last* graph of the complete sequence, and only nine characters have the third marker graph being the third graph of the complete sequence. That is why memorizing markers for each characters is not difficult for the user. It is clear that this invariance of the position of the graphs within marker reflects the natural regularity of perception of graphic images and facilitates the memorization of markers for each character.

The composition input allows to identify unambiguously after two manipulations 72.5% of the used characters (3146 characters), which is incomparably more than wubi (651 characters) and cangjie (238 characters). If we take into account the frequency of use of these characters in real texts, it turns out that in 80% of cases when inputting the average text, only two manipulations are needed to identify the characters. The maximum required number of manipulations to identify a single character at the current input method is equal to 3, whereas in wubi it is 4 and in cangjie it is 5. It is also worth noting that the described methods of structural coding after the identification of the character require user intervention for the act of input, i.e. to transfer the character into the focus of waiting input application. The composition input in marker mode allows the machine to track the *completed input signal* and automatically move the inputted character to the focus without user intervention (see below for details). Thus the user "saves" one manipulation, he does not need to indicate for machine the character, which should be transferred to the focus.

In the case of composition input, the selection of the character corresponding to the user-entered graphs is carried out according to the characters table in accordance with the user's preset selection mode. Since the character description table stores information about the complete sequence of graphs for each character, the method can use different selection modes depending on the task or the user's level of training. In general, it is possible to implement a selection of characters for any conceivable combination of graphs: the first and/or the last, one after another, but even every other, — for any combination. But experience shows that the preferred options include the following selection modes:

Free mode is a mode of selecting all characters, in which the graphs inputted by the user are met in any order. The mode can be used to input a character unknown to the user, for example, from a paper document, to search it in a dictionary or to get reference information on the composition of graphs or the marker of the needed character (see below). Having determined several graphs of the desired character, the user enters them, and visually finds the desired character in the returned list of hieroglyphs.

Sequential mode is a mode of selecting characters, in which the user-inputted graphs appear in the sequence of graphs of characters one after another and not necessarily from the beginning of character; the mode is useful for searching "by analogy" characters having common graph blocks that can also act as separate characters, which, among other things, allow tracking the inclusion of simple characters in more complex ones. Isolation of similar components of characters is a useful procedure; it facilitates memorizing characters and is traditionally widely used in teaching hieroglyphics. Now it can be done "on the go", without special pre-training of educational materials.

Fsl-mode is a mode of selecting characters, where the system interprets 3 graphs inputted by the user as the first, second and last graphs of the desired characters. It is a very useful way of quick input, if the user at the initial stage of learning still does not remember the markers. This mode allows identifying the majority of characters (69% or 2987) after inputting no more than two graphs and almost all (except for 40 characters) after three graphs inputted. To enter the remaining 40 characters, it suffices to add the next to last graph of the complete sequence to the three inputted graphs (see above description of fsl-algorithm). As to the 2 characters (暖 and 暧) mentioned above, they are inputted in this mode by their first, second and third graphs.

Marker-based mode is a mode of selecting characters, where the system interprets the graphs inputted by the user as marker graphs and returns characters containing the inputted graphs in the order specified in the markers of the characters description table. This is the fastest input mode. As mentioned above, 80% of characters of the average statistical text in modern Chinese are identified in this mode just after two manipulations.

Marker-based input offers the user some more advantages. At marker-based input, the mentioned above signal of completed input can be tracked. It is necessary to recall that in all characters input methods the input procedure as such is divided into 2 stages.

1. Process of identification of the desired character by some user inputted parameters and presentation of this identified character to the user for input or, in case

of uncertainty, formation of a list of candidate characters and presentation this list to the user for selecting the needed character.

2. Actual input of the desired candidate into the focus of a program waiting for input from the keyboard (performed by the user "manually", by separate manipulation).

At the marker-based input, in response to inputting a first graph by the user the system returns a list of characters, in which the first characters is that, whose marker either consists of the single inputted graph, or begins with the inputted graph. If the returned list contains only one character, then no graphs are to be inputted further. Otherwise, after inputting a second graph, the only character in the returned list will be the same. After inputting a third graph, if it was possible, it is already pointless to continue inputting, because there are no markers larger than three graphs in this version of composite input. The size of the returned list consisting of one character is precisely the signal of completed input, which can be automatically detected and used in the input algorithm. After such detection, the second stage of the traditional procedure involving the user becomes meaningless, it can be omitted, because the obtained character can be sent directly to the focus of waiting input application automatically. In the marker-based process, the input of 93.2% of characters ends upon detecting the signal of completed input. In other words, in 93.2% of cases the composite input method increases the speed of input owing to elimination of one manipulation. In the remaining 6.8% of cases, when the potential input of graphs is still possible in principle, but the needed character is already presented in the returned list in the first position, the user may send it to the program input focus of the program waiting for input from the keyboard by a separate manipulation.

Second advantage of the marker-based input is that after inputting the first graph, it is very simple to compute, using the characters description table, the set of marker graphs possible (or "allowed") for subsequent input and, thus, exclude those combinations of graphs that do not occur in the markers and, as a result, return an empty list of characters. To do this, it is sufficient to select in the characters table those marker graphs that occur in the next marker position after that inputted for characters from the list of characters returned by the system at the previous stage. Then, the remaining graphs not included in the resulting list can be locked on the matrix of graphs or switched to idle mode. This key lock mode not only allows getting rid of some amount of mechanical errors, but also makes it easier to master the method, giving the user a visual hint of possible options directly at the time of input.

Finally, the third advantage is that the use of the present input method in early stages of training contributes to better memorization of characters. While the structural coding methods with their complex decomposition rules can be applied only after the character is already "known", i.e. can be easily differentiated from others and its structure is already fixed in memory, the composition input method enables writing characters at the first stages of training, which just helps to understand and remember the structure of the characters as such. The mastering of the inputting of new characters is extremely simple, the input system suggests the possibility by a simple click, "on the fly" without leaving the frame of the application to know the composition of any of the character included in the system. It is enough by a convenient selection mode visually find the character and "ask" the system about the sequence of its graphs and the marker.

4. Structural hieroglyphics. Problems and prospects.

Structural hieroglyphics do one simple and obvious thing. It looks at the characters from the side of its structure, regardless of its sound, meaning, regardless of the primary composition of its strokes, looks at it as a composite image, assembled as a puzzle of several simpler units. Herewith, without being burdened with the technological limitations of the traditional method of entering, searching and remembering characters, the structural hieroglyphics suddenly finds that the character can not only be "divided" into components, but also "collect" it from the components can be as fast. It is not clear yet whether it is "gift of God" or "intrigues of Satan", but the composition input is much more effective than the phonetic one, i.e. faster, shorter in the number of manipulations required to input a single character. And it is already clear that the use of composite input slightly shifts the emphasis of teaching methods. And the most important thing is that it shifts them towards simplifying the teaching of hieroglyphics, accelerating the acquisition of reading and writing skills, i.e. reducing the threshold of entering the area of the Chinese language, which traditionally seems excessively high.

The "detachment" of Chinese character from the sound, the ability and the need to form directly graphic-semantic associations, bypassing the usual for cultures with alphabetical writing bunch of graphic-sound and sound-semantic associations, allows one to quickly develop the habit of "dumb" or speechless speed reading. Of course, Chinese children, when they learn to read, first pronounce the characters. But they already know that bāo is both "wrapping" and "ripping", and the teacher will tell them that in the text the first will look like 包 and the second as 剥. And 和 in the text may sound like hé and hè, and huó, and huò, and hú. And if you need to list a few things in the text, you need to write 和, not 河, although both may sounds the same (hé). The child simply has no other choice than to pay attention to the image in the text and to abstract from the sound, which can, of course, be spoken, but the sound itself will not give anything, and it is easy to give up.

In general, the habit of carriers of alphabetic writing read silently historically quite late acquisition. The first texts were written for their pronunciation, for their pronouncing aloud. Today it seems to us that consonant (or abjad) writing in general can neither be read nor understood without a pronunciation it aloud. The separation of meaning from sound has been (and is) very slow. At the very end of the 4th century AD, Saint Augustine described the ability of Ambrose the bishop of Milan to read silently as rare and surprising [23]. Unfortunately, I do not know such evidence for cultures with hieroglyphic writing. The appearance in the early middle ages in the texts the spaces between words, appearance paragraphs, tables of contents, abstracts, in fact, produced a "quiet revolution" by teaching people in the mass to "silent reading" [24, p. 384]. The meaning began to "break away" from the sound, the transfer of meaning ceased to require sound reproduction. It should be noted that when writing this separation of meaning from the sound is not obvious, if at all. During the writing, even if we do not pronounce the text in a voice, we "tap" it with ours fingers on the keys, "pronounce" it by symbol, one-by-one. The same "tapping" of Chinese texts leads to the need to double the attention, to describe without errors the flow of sounds, to convert them without errors into a flow of graphic images separated from these sounds. This leads to the appearance of strange and very rare for traditional hieroglyphics errors, when in text appears meaningless in terms of context character, which sounds the same as the correct one, but it looks different. (See, for example, [13, p.68], a characteristic typo when in the word tiānyá (天涯) instead of character 涯 appears has long been outdated character 漄.) And once you make a mistake when specifying the tone of the character which you try to write as appears one, well, very similar to the desired, but with a different meaning, and "paddle" (桨) turns into a "starch" (浆) [ibid., p. 84].

There are no natural reasons and rules of handling with images or with sounds that make this image sound that way, and this sound to fix on the letter in this way.

The phonetics of Chinese characters is initially arbitrary, as is the shape of the letters that transmit sounds in alphabetic writing systems. And if it is very difficult for representatives of cultures with alphabetic writing to get rid of the sound of the text, especially difficult for speakers of inflectional languages, the representatives of cultures with ideographic writing easier to move to silent speed reading, it is easier to "turn off" graphic-sound associations, when they are not necessary, and use directly graphic-semantic ones. Perhaps this is one of the reasons that Chinese children are faster than their European counterparts, for example, master reading. Chinese character initially is a container for transfer sense. In languages with alphabetical writing there is only a fixation of the flow of sounds, the carrier of meaning is the oral word, the sound conveys the meaning.

Without trying to figure out "which came first: the chicken or the egg", – we can still presume that the simplicity of operating with "image of sense" or the simplicity of conveying sense through images accompanies the image perception of the world. And the use of phonetic input, a return to the graphic-sound associations can if not destroy this "company" then break it thoroughly. With the advent of structural, non-phonetic input, effective enough to use it without loss, but in fact with a noticeable increase in the speed of input, it is possible to formulate the problem of explanation the features of the interaction of perception methods and methods of implementation of the transfer of sense. I'm afraid to talk about the "value" for the culture of these different essentially ways of conveying meanings, their relative simplicity or complexity, I do not know, is it "good" or is it "bad", but by the example of structural hieroglyphics, we can see how natural for the human consciousness was the emergence of such a "strange" (from our point of view, from the point of view of "alphabetical culture") method of transferring sense. It seems to us that the "fixed" sound is able to provide the interaction of individuals as effectively as the sound signal provides the interaction of the pack members during the hunt, for example, quickly and reliably. But after the hunt, outside the pack, in another context, when the sound is no longer heard, when you need to communicate with a "different level of being" (with "heaven" or "God"), or with another "time" (far ancestors or remote descendants), it turns out, fixing the image is as effective as fixing the sound. I'm afraid, as if not even more effective, judging by the time interval in the history of using this method of fixing and transferring senses. You can also see some kind of accompaniment (I will not pronounce the word "correlation", because the correlation is the properties of statistics, the correlation is considered and expressed in numerical language) and the correspondence of hieroglyphics and tradition as the properties of culture. At first glance, I do not know whether it is true or not, it seems that the use of hieroglyphs and increased attention to the preservation of tradition are created for each other. And again, the use of another (not better or worse, just another) principle of transmission of sense violates this correspondence, "tears" it. And the current Chinese teenager, who is using phonetic input on his smartphone all the time, in response to the indication of the erroneous character in his text, calmly declares: "What's the difference, it sounds right." It sounds right, but it is written wrong. Because with such extent of homophony as in the Chinese language, to separate two same-sounding words (or syllables), otherwise than to depict them otherwise is simply impossible. And no context will help here.

The composition input forms a different approach, a different view of the character, forms different attitude to it than the phonetic input do, it does not turn the character into a sequence of sounds, it retains the integrity of the graphic image in memory and ensures the purity of the graphic-semantic association. And this purity allows us to hope that along with "silent reading" it is possible "silent writing". Not only because this writing will be faster, it's a technique, it's a little interesting. But this writing can have a completely different result, not related to the flow of sounds, not expressed by the sounds of a living language. The mere mention of calligraphy as a separate area of culture may take us too far away. But quite understandable analogies can be found outside of calligraphy. The properties of any numbers have no relationship to his name: 12 — "twelve", "shíèr"; 3 — "three", "sān". But we will never get from the set of these sounds the number 9, which is easy to get by a simple arithmetic action: 12 - 3 = 9. The symbolism of formal systems is essentially a hieroglyphics rather than an alphabet, because in formal systems the meaning is so thoroughly detached from the sound that it is not to comes to our mind to record arithmetic operations with the words: "one add two".

The structural approach to hieroglyphics allowed us to find some critical points in the composition of hieroglyphs, which determine the uniqueness of each character and, thus, provide the speed of the recognition process. These are markers, these are something that are immediately evident. The impossibility for me of setting a correct series of experiments to clarify the relative importance of these critical points led to the inevitable arbitrariness in the choice of the marker. It is clear, first graph of each character dominates of course, but what about the others? 2051 characters having more than two graphs in a complete sequence has markers consisting of 2 graphs. Herewith, most of these characters allow the use of both the second and the last graph as part of the marker. This is especially true for relatively simple characters of three components, and hence, described by three graphs. Immediately, without thinking: the character 以 is uniquely identified by the sequence of graphs 𠄌, 丶 and also by the sequence 𠄌, 人. Which sequence is best used as a marker? In this case, it seems that the sequence 𠄌, 人 better reflects the specifics of the perception of this image, and therefore it will be easier to remember. But it is not always the case. Two more examples. For character 霞, both 雨, 口 and 雨, 又 work as a marker, and both 厂, 欠 and 厂, 丷 work for the character 厥. The last example is especially illustrative. For a person unfamiliar with the hieroglyphics, the 厂, 欠 sequence seems visually comfortable, but for a person familiar with the rules of writing characters, it is easier to use the 厂, 丷 order, since he is already used to writing it in this order. And there are a lot of such examples. Proponents of handwriting will immediately notice that the graph 辶 can not be the first in the description of the component sequence, because in most cases it is written last in order. But I see it first, because it is located on the left.

Next example is set of characters with 曷 element at the end of its structure. There are 12 such character in the our set: 喝, 渴, 竭, 揭, and so on. Graph 勹 located inside these characters, but visually it is the last element of these characters. And I see it as the last. But by formal consideration of structural hieroglyphics the last graph is ㇙, and we is forced use it as the last graph for markers. Regularities of perception contradicts to formal rules of structural hieroglyphics. The worse for these rules. I suppose, we must use 𠃌 as the last graph of markers in these cases. It will be easier to remember such markers.

In any case, all this examples requires competent discussion.

For any field of science one of the most important points is the question of its heuristic value. With the pragmatic value of structural hieroglyphics, everything is clear, it helps compositional input to achieve results that are not available to other input systems. In the composition input method, there are algorithms that allow to identify (and, accordingly, to enter on the computer) any given character maximum by three of its components. Therefore, three manipulations are enough to input any character. For the simplicity of the calculation we will not take into account the manipulation of the translation of the selected character in the input field — sometimes with phonetic input they may need more than 2, and with composite input only one or, depending on the program settings, may not be necessary at all. Also, we will not take into account any kind of tricks of predictive input, which are the subject of a separate discussion. But, as we have already seen, with the help of composition input of 72 percent of characters only 2 manipulations are required. And if we take into account the real frequency of occurrence of characters in texts, this number increases to 80 percent. The phonetic input method for entering the shortest syllable requires at least 2 manipulations, and for entering any character it requires on average 4.2 manipulation. Thus, composite input accelerates the input of characters at least two times.

But it may be much more interesting to extend the principles of structural hieroglyphics to other systems of hieroglyphic writing. There is almost no doubt that it is possible to implement the principles of structural hieroglyphics on a set of Chinese characters of traditional style. Moreover, since traditional hieroglyphs did not pass through the "needle's eye" of simplification reforms, the principles of their arrangement are closer to the original, mnemonics is not distorted by the replacement of some elements by others in the process of reform, the composition of their components, at first glance, seems more logical and easier. And new circumstances may arise when trying to apply the principles of structural hieroglyphics to a lot of Japanese characters, as they also changed in detail, but with their own, Japanese, specificity.

Solving purely practical tasks of developing input systems, organizing dictionaries, which is valuable and interesting in itself, the structural hieroglyphics allows us to see the unresolved issues in other areas of science. Psychology and pedagogy, for example, basing on the peculiarities of the perception of complex graphic images by the child, can understand what is the specificity of teaching hieroglyphic writing as opposed to teaching alphabetical writing. Equally, and vice versa: what is the specificity of the perception of simple, but "voiced" images-letters by child, who before learning to read sees the world of "pictures" with a name, and the division of this name into individual sounds for him a separate and difficult task. Why should the letter in the ABC first have a name: "a", "bee", "cee"; "dee", — and only then it acquires a sound? Why a child first counts specific objects and the very procedure of counting for him is only a sequence of sounds: "one", "two", "three" etc. And only a few later this sounds acquires for him the image of digits which are a graphical image of some abstract sense or are hieroglyphs in fact .

What is the impact of the specifics of teaching hieroglyphics, its difference from learning alphabetic writing has on the individual development of the child? What is the impact of hieroglyphic writing with its specificity at the level of individual perception on the vision of the world as a whole, on culture as a whole. How far are the differences between alphabetic and hieroglyphic cultures?

Structural hieroglyphics not only helps to formulate these questions, but also allows us to hope that someday its will be answered. I think that structural hieroglyphics will be a good help for people who decided to go to search of these answers or go to exciting "journey to the East", and especially for those who decided to do it without professional guide or, in the words of V.P. Vasiliev, quoted in the introduction, "without a teacher, with the help of one lexicon."

Appendixes.

Appendix 1.

Graphs table.

For those who are familiar with any variant of hieroglyphic radicals this description may seem redundant, and you can quickly glance over it paying attention only to the non-radical components, and those who are not familiar with the radicals it will be useful to view it carefully. The graphs are arranged in the table in descending order of frequency of occurrence. I draw the attention of the "traditionalists" that the first column contains the sequence number in the table (it can be considered as a *frequency rating*), and not the number of the Kangxi radical in traditional understanding. The third column is really just the name of the graph and nothing else, so it is quite arbitrary and does not have to match the name or meaning of a graphically similar radical. Each graph encodes a hieroglyphic component similar to it, as well as the components marked in the "variants" column. An empty description means that there are no appreciably different variants of components for this graph, of course, except for the size changes and minor changes in proportions. In other words, if the graph coincides with the Kangxi radical (or one of its variants), then all other variants of the given radical are encoded by the given graph, and therefore, as a rule, they are collected in one row in the table. With the exception of two graphs: 亻 (No. 4) the variant of the radical 人 (no. 5), defined as a separate graph, and 扌 (No. 20) the variant of the radical 手 (No. 9) defined as a separate graph too. The separation of these two Kangxi radicals and the use of their variants as separate graphs associated with a high degree of uncertainty otherwise. For example, if the graph 扌 is considered as a simple variant of the graph 手 (as tradition does it in respect to the radical 手), then for the machine the characters 右 and 扣 will be indistinguishable, since the sequence of graphs of both characters will be the same: 手 and 口 (the first graph of the character 右, as you can see from row 9 of the table, is a variant of the graph 手). The same is true for characters 左 and 扛, and in other similar cases.

No	picture	name	variants	description
1.	一	one	七 刁 之	may be slightly sloping and moderately sloping, and in general the curve, as in the last example
2.	口	mouth		be careful, not to confuse "mouth" with the "surround" (№ 50). They differ only in size (the only case in hieroglyphics), but in the "mouth" nothing happens, and in the "surround" you can find a lot of things...
3.	乙	second (cyclical sign of ten)	艺 也 九 买 诀 乃 及 扬	we will adopt following the tradition that as his variants should be considered: 乛, 𠃌, 乚, ⺄ etc.. A few variants, be careful. From simple to sophisticated, as in the last three examples. In the second example there are 2 variants of "second" together .
4.	亻	man (on the left)		
5.	人	man	众	The diversity (three variants) of "men" is collected into the "crowd". Note that the standard "falling rightwards" turn effortlessly to the "dot", when necessary, that actually allows us to incorporate them in the graph № 10 (this "calligraphic trick" we will meet again and again).
6.	日	sun	昌 昭	in the examples, two variants of the sun, just to eyes accustomed: one very "stretched", another extremely "elongated"

7.	勹	wrap		
8.	白	white		
9.	手	hand	我 掰 帮 友	covers all conceivable variants of "classic" hands except "扌" (№ 20). In combination with the "spear" (№ 27) it is not clear who owns the horizontal line. In the second example, meets his variant, in the third pulls up "the finger". Well, in the last example is extremely simplified "hand", which you will always be confused with the combination of "falling leftwards - one" (№ 12 - № 1). Be careful.
10.	丶	dot	冬 尺	same as "falling rightwards", i.e. the line going from left to right, top to bottom, regardless of its length. Three variants of "dots" on the example of two characters in the next column.
11.	土	earth	垅	differs from similar "scholar" (№ 102) ratio horizontal lines ("earth" has shorter upper stroke then lower). In straitened circumstances slightly "tilts" the bottom line, as, for example, before "dragon" (a very common calligraphic trick).
12.	丿	falling leftwards	少 夭 囱	can "lean" a little bit, being almost vertical, can recline thoroughly as in the first example, and can almost lie down supine, as in the second example. In the third variant is reduced almost to a point, but always retains its direction from top-right to left-down.

13.	刀	knife	剪 尔 师 临	In the first example, the two "knives": almost classic, only slightly compressed by the weight from above, and above it "knife on the side". The second character is really "knife on top". Next two variants of "knife on the side", but they are all "knives" for us.
14.	木	tree	枉	"tapering" of the "tree" by type "man" (№ 5), reducing "falling rightwards" to "dot" (№ 10).
15.	辶	walk	嗵	It's almost always located on the left side of the character and underlines it from the bottom, but sometimes "hiding" inside.
16.	丨	vertical stroke		vary only in length (sorry, height :).
17.	月	moon	那 肯	in addition to the "standard" types there are 2 variants, but always recognizable.
18.	水	water	永 涵 承	in version of the classic "water" is slightly changed middle hook. In the following example 2 "water": "water on the left and frequent variant "water" when it is "closely". And a unique case, when the water is "covering" sign and incorporates 3 horizontal line (an example of "neglect" tradition, which believes that the radical is "hand" here).
19.	讠	say (abbr.)	言	"full version " of radical (occurs much less frequently).
20.	扌	hand (var.)		easily recognizable sign, variations only in height.
21.	二	two		encodes all the paired horizontal lines.

22.	心	heart	意 蕊 性 恭	slightly flattened when it's located at the bottom. Habitual and frequent version of "heart on the left" (example 3) and quite rare in the last example.
23.	丷	antennae	兰 业	is not necessarily at the top, may be lower, but then he "covers" some graph. Never found at the very bottom.
24.	厶	private		
25.	女	woman		
26.	亠	top		
27.	戈	spear	弋	"reduced" version of the spear, in the tradition known as the radical "to hunt".
28.	大	big	因 爽 拳	several variants of the "big" quite recognizable and caused only by changes in the height or width of the sign.
29.	宀	roof		
30.	卜	foretell	卡	two variants of "foretell" in one character.
31.	十	ten		
32.	力	strength		
33.	亅	hook		

34.	寸	inch		
35.	纟	thread (abbr.)	纆 紊	sometimes you can meet variant which probably has slipped through "reform of simplify".
36.	疋	leg	是	in some cases there is no hook on its upper horizontal
37.	八	eight	公 共 兆 亦	sometimes skips inside itself another "objects" or able to "cover" the other graphs as in the last two examples.
38.	又	again	对 祭	two variants, the second is found only in derived from the given character.
39.	阝	beta		combines traditional "town-mound", because the style they are not distinguishable, and he is located on the left or on the right, is determined by the sequence of graphs in the description of the character, therefore, the separation between "town" and "mound" seems redundant. No variants.
40.	目	eye	且 睢	common variant "eyes" with elongated lower horizontal.
41.	王	king	玨	predictable variant - it's the expected change of the lower horizontal ("two-in-one" in the example)
42.	禾	grain	秆	variant formed by the same type as that of the "tree" (№ 14) and of the "man" (№ 5) - reduction of falling rightwards stroke.

43.	冂	down box	响 冒 澳 鲷	one of the "covering" graphs, inside of which is never empty, except for one character in combination with the "surround" (№ 50). Four variants are shown.
44.	⺌	topknot	肖 学	tradition regards it as a variant of "small" (№ 61), but it seems easier to consider it as a separate graph with two variants.
45.	己	self	己 巳 改	of the three variants, the second one is "dangerous" because of its resemblance to the variant "seal" (№ 88).
46.	𠂉	visor		in the traditional set there is no such thing, but sometimes it greatly simplifies the description.
47.	子	son	子	a slightly modified horizontal alternate forms of this graph
48.	立	stand	旁	when it located on the "side" (№ 73) and on the "turban" (№ 84), apparently, for more stability, based not on the "one" (№ 1), but on the "cover" (№ 94)
49.	山	mountain		
50.	囗	surround	回 凸	compare with "mouth" (№ 2). If the inside of the object there is something, so it's not "mouth", but "surround" (funny first example is the "mouth" in the "surround") Covering graph. It can be empty inside, but rarely (example 2 is interesting combination of two covering graphs "down box" and "surround").
51.	儿	child	兆 枕 胤	pay attention to the third example, in which this graph passes the other graphs inside itself.

52.	夕	evening		sometimes confused with the "moon" (№ 17) but more like a "meat on left" (№ 128)
53.	门	gate		
54.	米	rice	籽	a variant of rice that you will recognize without my prompt.
55.	㇄	hook-to right	东 乐 甚	in the traditional set of radicals is missing. In addition to direct correspondence to the image, it encodes also "a lower left corner" right up to right angle .
56.	厂	steep	反	sometimes with a sloped top.
57.	攵	strike (on right side)		be careful, it tends to be confused with "walk slowly" (№ 63).
58.	文	language	刘	
59.	丁	tau	叮 可	changes only the length of the top horizontal.
60.	艹	grass (on top)		
61.	小	small	尖	occurs only slightly shortened version.
62.	耂	old (abbr.)	老	tradition views it as variant of "old" (see example). But we will do the opposite, since this variant is encountered incomparably more often than traditional.

63.	夂	walk slowly	处 卤 夜 修	variants are very similar and recognizable. One of them even allows to put a "dot" inside of it.
64.	丨丨	two vertical		in tradition it absent, but very useful because it make our life easier :).
65.	匕	dagger		
66.	羊	sheep	羔 差	shortened from the bottom, variant "sheep" loses "leg" (or "tail"?) and sometimes just slightly bent.
67.	里	village		
68.	尸	corpse		not to be confused with "door" (№ 110)
69.	止	stop	此	also quite predictable variant.
70.	田	field		
71.	廾	two hands		
72.	贝	shell		
73.	方	side	万 嗷	located separately (example 1) or merging with other graphs (example 2) sometimes loses the top "dot".
74.	彳	step		

75.	工	work	功	predictable variant.
76.	牛	cow	告牲	two variants: a "shortened" and "compressed"
77.	⺮	bamboo (on top)		traditional "full" bamboo, it seems, does not occur at all...
78.	火	fire	燃	two variants of the same character: the first one is simply compressed laterally and easily recognizable, and the second... well, let us get used to that "the fire from below" also is "the fire"
79.	至	to		
80.	广	exten-sive		
81.	见	see		
82.	斤	ax	丘	based on "one", change its falling leftwards stroke to vertical stroke
83.	示	spirit	祀	"spirit on the left" is less similar to its prototype than to variant of clothes (№ 117, clothes on left). Be careful.
84.	巾	turban		
85.	冫	ice		be careful, it's very similar to variant of "water" (№ 18, water on the left).

86.	覀	west (on top)	洒	the full (traditional) variant of the "west", as in the example, is less common than abbreviated variant, and it is very similar to "wine" (№ 159). To avoid confusion in matrix an abbreviated variant of this graph is included to it.
87.	车	car		
88.	卩	seal	节 范 剜	of the three variants of this graphs, the second is "dangerous" because of its similarity with the variant of graph "self" (№ 45).
89.	艮	though	即 退	two compacted variants.
90.	犬	dog	独	"dog on the left" occurs much more often then its traditional variant.
91.	生	give birth		
92.	豕	pig	象	in combination with "knife on top" (№ 13) - "mouth" (№ 2) or after the "pig's nose" (№ 103) loses its top horizontal.
93.	殳	weapon		
94.	冖	cover		"roof" (№ 29) without "dot" and without variants.
95.	*	asterisk - blank graph		the graph modifier, the blank graph, the asterisk ("*" as it appears in the description of the graphs sequence of some charact-ers and in the graph matrix) is a graph which does not have its own "image" but demonstrates the some modification of the graphs sequence.

96.	页	page		
97.	玉	jade		
98.	走	leave		
99.	耳	ear		
100.	甘	sweet		
101.	乂	cross		
102.	士	scholar		Chinese "scholar" is different from "earth" (№ 11), except the first elongated horizontal, so that is always right, regardless of the size or tightness. No variants.
103.	彐	pig's nose (abbr.)	彔碌缘	sometimes stretches out the "nose" trying to look like "brush" (№ 141), but never linked with the "vertical stroke" (№ 16) , "hook" (№ 33) and "falling leftwards" (№ 12). Sometimes, as in the second example, pulls the lower horizon-tal. Full version meets today rarely (example 3).
104.	金	gold	银滏	most often used in the abbreviated form "gold on the left" (example 1). In the second example, the cu-rious case of merger "father" (№ 131) and "gold"
105.	弓	bow		

106.	隹	bird without tail		
107.	丬	chip	叫 寐 藏 鼎	a simplified version of "chips". Sometimes it may be situated on the right and then it looks diffe-rently (example 1). Full traditional variant is very rare, with slightly varies (examples 2, 3 and 4).
108.	几	small table	微	reduced small table "flexes" one its leg
109.	青	blue		
110.	户	door	所	the variant is old "door" (the rest of traditional style).
111.	用	use	甬	sometimes "straightens" the falling leftwards stroke.
112.	氏	clan	民 旅	the top falling leftwards can be "masked" by another graph, and the horizontals turn into falling leftwards
113.	舌	tongue		
114.	自	oneself		
115.	凵	recep-tacle	凹	usually covering graph himself comes inside "surround" (№ 50).
116.	矢	arrow		

117.	衣	clothes	裛 袁 衤	unexpectedly turned out to be covering graph. Lets place the other graphs under the "hat", and sometimes just loses it as in second example. But more often you may meet the "shortened" version of the "clothes" – the so-called "clothes on the left". Beware, this variant is very similar to the "spirit on the left" (№ 83), added only a small falling leftwards on the right.
118.	⻊	foot (on the left)	促	the full version of "foot" is quite rare.
119.	音	sound		
120.	行	go	街	covering graph. It may have a lot of things inside it. Be careful when decomposition.
121.	穴	cave	空 罕	expected modification of the "cave" in the first example and the unexpected loss of the top dot in the second...
122.	干	stem		
123.	石	rock		
124.	而	and		
125.	长	long		
126.	欠	owe		

127.	马	horse		
128.	⺼	meat (on the left)	肉	very similar to the "moon" (№ 17) and they are often confused, but it is a separate thing, please pay attention to it, I specifically chose a font that draws "meat on the left" by other way than the "moon". The full version of "meat" (see example) occurs much less frequently.
129.	爫	claw (on top)		
130.	首	chief		
131.	父	father		
132.	面	face		
133.	廴	journey		
134.	彡	hair		
135.	角	horn		
136.	川	stream	侃 巡 粼 带	in the first example "stream" flexes his last vertical line. There are another version of "stream" (example 2) and a shortened "version of version" (example 3). And the last reduction to the state of the three verticals.

137.	皮	skin		
138.	高	high		
139.	比	compare		
140.	匚	enclose		
141.	聿	brush	尹 争 妻	compare with "pig's nose" (№ 103)
142.	豸	reptile		
143.	皿	utensil		
144.	无	without	僭	the second horizontal line takes a hook on the left when the graph is repeated (compressing the right leg) or is located after the "though" (№ 89)
145.	歹	evil		
146.	非	non-		
147.	身	body		
148.	虫	worm	瞶	located on the "shell" it loses the right "dot".
149.	支	branch		

150.	雨	rain	雪	the "flat" version of "rain" is more often than traditional (51:2)
151.	斗	dipper		
152.	罒	net (abbr.)	网	rare "full" version of the "net". In combination with other graphs are not noticed.
153.	疒	sick		
154.	气	air		
155.	饣	food (on the left)	食	in seclusion acquires a traditional "classic" look.
156.	齐	neat		
157.	入	enter		
158.	卯	fourth (cyclical sign of 12)	留 卿	lying on top is compressed, as in the first example. In one case acts as a covering graph, passing inward "though" (№ 89).
159.	酉	wine		
160.	舟	boat	彤	it was discovered only one variant when the "boat" lost the bottom "dot"
161.	片	slice		
162.	牙	tooth		

163.	谷	valley		
164.	采	pick		
165.	革	leather		
166.	风	wind		
167.	豆	bean		
168.	匚	box		
169.	尢	lame	尴	in this embodiment the "lame" has feet clearly different length. Of the 18 characters with graph the lame this variant, when it underlines the next graph, is found three times.
170.	臣	servant	巨	in some cases, loses internal vertical lines.
171.	幺	insignifi cant	乡	in a variant "dot" changes to fall-ing leftwards. This variant, by the way, occurs in texts much more then standard
172.	韦	soft leather		
173.	鸟	bird	乌	variant without an internal "dot". In graphs sequences referred to as 鸟*.
174.	毛	wool		
175.	黑	black		

176.	辛	pungent		
177.	麻	hemp		
178.	玄	abstruse		
179.	缶	amphora		
180.	鱼	fish		
181.	羽	feather		
182.	虍	tiger		
183.	臼	mortar	舆	it often acts as a covering graph, significantly changing in only one variant.
184.	龙	dragon		
185.	飞	fly		
186.	鬼	ghost		
187.	黄	yellow		
188.	鬲	jug		
189.	辰	morning		

190.	骨	bone		
191.	癶	footsteps		
192.	髟	long hair	套	tradition claims it's "long hair on the head" but in the example it lost "hair".
193.	香	perfume		
194.	矛	lance		
195.	瓜	melon		
196.	耒	plow		
197.	瓦	tile		
198.	齿	teeth		
199.	禸	track	禺	changes only in combination with the "field" (№ 70)
200.	舛	error		

Appendix 2.

Collisions and their solutions.

Collisions of the *first kind* – uncertainties in the sequence of graphs. These collisions are eliminated by adding a graph-modifier (*) to the rare character. Note that it is added *at the end* of the sequence — this makes it easier to enter "exceptional" characters. For this kind of modification are selected, of course, more rare characters.

No	rare character	graph sequence	frequent character	graph sequence
1.	未	木一*	本	木一
2.	末	木一**	本	木一
3.	杏	木口*	束	木口
4.	呜	口鸟*	鸣	口鸟
5.	沮	水目*	泪	水目
6.	呗	口贝*	员	口贝
7.	晾	日亠口小*	景	日亠口小
8.	抻	扌丨日*	抽	扌丨日
9.	旧	丨日*	由	丨日
10.	申	丨日**	由	丨日
11.	叭	口八*	只	口八

Collisions of the *second kind* – slightly different characters. These collisions are also eliminated by adding a modifier sign to the description of the rarest character. This addition of a modifier is often relevant only for characters consisting of a single component. In characters consisting of several components, this addition of a modifier is often irrelevant and is not really made, as for example, for the character 县, described by graphs 目 and 厶, this sequence of graphs simply does not have anything to be confused with. The characters in line 3 occur with approximately equal frequency (~0.15%), so the choice was made in fact arbitrarily, based on the "logic" of the image (see line 4 with an even rarer character, where 2 modifiers are used).

No	rare character	graph sequence	frequent character	graph sequence
1.	曰	日*	日	日
2.	且	目*	目	目
3.	已	己*	己	己
4.	巳	己**	己	己
5.	万	方*	方	方
6.	孑	子*	子	子
7.	巨	臣*	臣	臣
8.	乌	鸟*	鸟	鸟
9.	弋	戈*	戈	戈
10.	代	亻戈*	伐	亻戈
11.	汩	水日*	沓	水日
12.	裸	衣田木*	裹	衣田木
13.	晖	日冖车*	晕	日冖车
14.	邑	口己丨*	吧	口己丨
15.	赫	土丿亅八*	赤	土丿亅八
16.	审	宀丨日*	宙	宀丨日

Characters table.

This Appendix provides only a portion of the table of characters, reflecting its essential attributes. Those table fields that are certainly important in the selection algorithms, predictive input, blocking matrix elements, but are not required to explain the principles of composite input is omitted. It is clear that the publication of the entire table for 4336 characters is meaningless, here are only 23 lines from it as an example. Who is interested in the composition of all characters, can contact the developer's website [20].

No	character	graphs sequence	marker
1	一	一	一
2	二	二	二
3	三	一二	一二
4	丑	二‖一	二一
5	不	一亻 丶	一亻 丶
6	业	‖丷一	‖丷
7	专	丨二乙丶	丨二
8	我	手戈	手戈
9	鹅	手戈鸟	手鸟
10	国	囗玉	囗玉
11	鸟	鸟	鸟
12	乌	鸟*	鸟*
13	世	乚凵一	乚凵
14	丘	斤一	斤一
15	丙	一人冂	一人
16	东	乚一小	乚小
17	丞	乙水一	乙水
18	丢	丿土厶	丿土
19	员	口贝	口贝
20	呗	口贝*	口贝*
21	喜	十豆口	十豆口
22	暖	日爫一手又	日爫一
23	暧	日爫冖手又	日爫冖

Table 1. Distribution of characters by groups along the length of the full sequence of graphs (according to T. McEnery and R. Xiao [11]). The difference in the total amounts comes from the fact that the table of characters were added 8 characters, absence in the Lancaster corpus).

number of graphs in full sequence	characters	% in corpus	accumulated %	characters in group	% characters in group	accumulated % characters in group
1	145	3,34	3,34	7253559	12,43	12,43
2	963	22,17	25,51	22486596	38,54	50,97
3	1574	36,23	61,74	18256981	31,29	82,25
4	996	22,93	84,67	7683814	13,17	95,42
5	488	11,23	95,90	2076157	3,56	98,98
6	125	2,88	98,78	452481	0,78	99,75
7	40	0,92	99,70	124637	0,21	99,97
8	7	0,16	99,86	16343	0,03	100,00
9	3	0,07	99,93	2159	0,00	100,00
10	1	0,02	99,95	146	0,00	100,00
11	2	0,05	100,00	171	0,00	100,00
total	4344	100		58353044	100	

Table 2. The distribution of characters in groups along the length of the markers (according to R. T. McEnery and Xiao [11]).

number of graphs in marker	number of characters in group	% in corpus	accumulated %	total number of characters in group of corpus	% characters in group of corpus	accumulated % characters in group of corpus
1	152	3,50	3,50	7317715	12,54	12,54
2	2998	69,01	72,51	41399460	70,95	83,49
3	1194	27,49	100,00	9635869	16,51	100,00
total	4344	100		58353044	100	

References.

1. Васильев В.П., Графическая система китайских иероглифов. Опыт первого китайско-русского словаря. (Репринтное издание, оригинал 1867-го года.) — Санкт-Петербург, 2010.

2. Баранова З.И., Гладков Е.В., Жаворонков В.А., Мудров Б.Г.; Под ред. Мудрова Б.Г., Китайско-русский словарь. — М., Рус. яз., 1988.

3. Сторожук А.Г., Предисловие к переизданию, в кн: Васильев В.П. Графическая система китайских иероглифов. Опыт первого китайско-русского словаря. (Репринтное издание, оригинал 1867-го года.) — Санкт-Петербург, 2010.

4. NJStar Software Company. URL: http://www.njstar.com/cms

5. MDBG English to Chinese dictionary. URL: http://www.mdbg.net/chindict/chindict.php?page=chardict.

6. Michael Love, private message, 2015. "Thanks for your email, but we're actually already working on something like this - just refining our component data (in the CHARS tab) first." - 16.09.2015.

7. Newton Isaac, "The Mathematical Principles of Natural Philosophy", translated into english by Andrew Motte, University of Adelaide, URL: https://ebooks.adelaide.edu.au/n/newton/isaac/mathematical-principles-of-natural-philosophy/book3.4.html.

8. Задоенко Т.П., Хуан Шуин, Начальный курс китайского языка, Ч.1 — М., Восточная Книга, 2007.

9. The Unicode Standard Version 10.0 - Core Specification, 18.2 Ideographic Description Characters. URL: http://www.unicode.org/versions/Unicode10.0.0/ch18.pdf

10. II Core Version 2.2. URL: http://appsrv.cse.cuhk.edu.hk/~irg/irg/IICore/IICore.htm

11. McEnery Tony, Xiao Richard, The Lancaster Corpus of Mandarin Chinese (LCMC), — Lancaster University, 2004. URL: http://www.lancaster.ac.uk/fass/projects/corpus/LCMC/default.htm (adres of main material URL: http://www.lancs.ac.uk/fass/projects/corpus/data/Chinese_character_frquency_list.zip).

12. The Unicode Standard, Version 10.0, Kangxi Radicals, Range: 2F00–2FDF. URL: http://www.unicode.org/charts/PDF/U2F00.pdf

13. Китайский язык. 50 текстов для чтения. Начальный и средний уровень. — М., Восточная книга, 2006.

14. Zhang Jishan, On screen chinese keyboard, Unated States Patent US 6,809,725 B1, 26.10.2004.

15. Lakritz David, Method and apparatus for identifying characters in ideographic alphabet, Unated States Patent US 5,586,198 A , 17.12.1996.

16. Wubi method: Wikis. URL: http://www.thefullwiki.org/Wubi_method.

17. Wubi86 table. URL: http://raw.githubusercontent.com/definite/ibus-table-chinese/master/tables/wubi-jidian/wubi-jidian86.txt.

18. Overview Of The Cang-jie Method. URL: http://www.cocoanutstech.com/cang_jie/www.cjmember.com/the_cj_method.htm

19. Index: Chinese Cangjie. URL: https://en.wiktionary.org/wiki/Index:Chinese_Cangjie.

20. 11-е декабря (эл. альманах), № 4. URL: http://poutko.ru/China/EnglishIndex.html

21. DownLoads. URL: http://poutko.ru/China/DownLoadsEn.html

22. Short description of tools. URL: http://poutko.ru/China/HowItWorks.html

23. Августин А., Исповедь. — М.: "Ренессанс", 1991. - Кн.2.III.3.

24. Шкуратов В.А., Историческая психология. Учебное пособие. — М.: "Смысл", 1997.

Contents.