Difficulties in Understanding Natural Language Texts in an Automated System for Solving Plane Geometric Tasks

Xenia A. Naidenova[a], Sergey S. Kurbatov[b], Vyacheslav P. Ganapolsky[a] and Taniana A. Martirova[a]

[a]*Military medical academy, Saint Petersburg 194044, Academic Lebedev street,6, 7, Russia*
[b]*Research Center of Electronic Computer Engineering, Moscow 117587, Warsaw highway, 125, 7, Russia*

Abstract
In this paper, the language and non-language difficulties in understanding texts of plane geometric tasks are presented: homonymy, ellipsis, ideographic synonymy, anaphora, quantifier expressions, syntax lexemes, and ambiguities of presentation. These difficulties have been revealed in the course of developing and using an automated system for solving plane geometry tasks expressed in natural Russian language. As one of progressive directions to overcome the difficulties listed above, a systematic approach based on forming cognitive schemes of plane geometric objects and configurations is proposed. The cognitive scheme of a task plays the role of deep structure synonymously reflecting the content of the task. The original description of a cognitive scheme in natural language as a combination of nominal, verbial, and propositional phrases can be transformed into the equivalent sentences with the use the transformation analysis taking into account all the linguistic phenomena in the real text of the task considered.

Keywords
Plane geometry, natural language processing, anaphora, ellipsis, cognitive scheme, transformation analysis, automated system.

1. Introduction

The analysis of language and non-language phenomena that hinder the successful understanding of texts in their processing by automated computer systems was undertaken as part of the implementation and development of an automated system for solving planimetric problems in natural Russian language. This system is described in a number of works [1–3]. The task texts were analyzed from the following sources [4–6].

It should be noted that the planimetric tasks are divided into separate classes, the texts of which have specific traditionally well-established linguistic features. For example, there are tasks for proof, for building, for calculating, for testing the theoretical knowledge of students. The textbooks highlight tasks aimed at studying only one of the planimetric figures, such as "inscribed and circumscribed circles", "length of circle and area of circle", or one class of ratios, such as "metric ratios in triangle". The structure of the text and its stylistic features depend on the class of the task.

In computer analysis, difficulties in understanding texts can be caused by such sentences, which, it would seem, correctly constructed.

This paper describes the most common difficulties in understanding texts in order to develop proposals for a common approach to overcoming them by applying cognitive patterns (schemes) of

DTTL-2021: International Workshop on Digital Technologies for Teaching and Learning, March 22-28, 2021, Kazan, Russia
EMAIL: ksennaidd@gmail.com (X.A. Naidenova); curbatow.serg@yandex.ru (S.S. Kurbatov); ganvp@mail.com (V.P. Ganapolsky); martta462@yandex.ru (T.A. Martirova)
ORCID: 0000-0003-2377-7093 (X.A. Naidenova); 0000-0002-0037-9335 (S.S. Kurbatov); 0000-0001-7685-5126 (V.P. Ganapolsky); 0000-0003-0000-6608 (T.A. Martirova)

geometric objects and relationships similar to the approach that had been proposed to resolve ellipses in the texts of planimetric tasks from the school geometry course [3, 7].

2. Language difficulties in understanding texts of planimetric tasks

Polysemy (a special case of homonymy) is a bunch of related concepts expressed by the common word. Consider, for example, the word "square". In one case, it means a geometric figure, in other cases – a calculated ratio:

Find the **sum of squared lengths** of the diagonals of ABCD if the length of segment OP and the radius of the circumscribed circle R are known.

On sides AC and BC of triangle ABC **squares ACA_1A_2 and BCB_1B_2** are constructed outwards. Prove that lines A_1B, A_2B_2 and AB_1 meet at one point.

"It seems that it was the problem on Hyppocratus' crescents that induced in many persons great expectations to the possibility of **squaring the circle**: the area of the figure formed by arcs of circles is equal to the area of a triangle [8].

The examples in Russian and their translation in English:

*Периметр прямоугольного треугольника равен 132, а **сумма квадратов сторон треугольника** – 6050.*

The perimeter of the right-angled triangle is 132, and the sum of squared sides of triangle is 6050.

*Дан параллелограмм, в котором острый угол 60°. Определить отношение длин сторон, если **отношение квадратов длин диагоналей** параллелограмма равно 197.*

The parallelogram is given, in which acute angle is 60°. Determine the ratio of the lengths of the sides, if the ratio of the squares of lengths of parallelogram's diagonals equals 197.

*На сторонах равнобедренного прямоугольного треугольника с катетом b построены **квадраты** во внешние стороны.*

On the sides of the equilateral right-angled triangle with a leg b, squares are built outwards.

***Квадрат** PQRS вписан в треугольник ABC так что вершины P и Q лежат на сторонах AB и AC и вершины R и S лежат на BC.*

Square PQRS is inscribed into triangle ABC so that vertices P and Q lie on sides AB and AC and vertices R and S lie on BC.

In the following tasks, the expression "enclosed in", "is between" have different meaning: either as a spatial relationship, or as a numerical expression:

*Докажите, что сумма расстояний от любой точки M, расположенной внутри неравностороннего треугольника ABC, до прямых AB, BC и CA **заключена между наименьшей и наибольшей высотами**.*

Prove that the sum of distances from any point M located inside the non-equilateral triangle ABC to straight lines AB, BC, and CA **is between** the lowest and highest altitudes.

*Докажите, что сумма расстояний от любой точки, расположенной внутри треугольника, до его вершин **заключена между его полупериметром и периметром.***

Prove that the sum of distances from any point inside the triangle to its vertices is between its half-perimeter and perimeter.

Some examples from original English texts:

Intercept (noun): the line segment **between two points of intersection** with other lines; the distance **between the centers of circles** is equal to d**;** prove that in triangle ABC bisector AE **lies between median AM and height AH**.

The following texts require the understanding of the semantics of words "divide", "dividing," "divided" expressing either an action in space or arithmetic operation:

*Основание треугольника **делится высотою на части** длиной 36см и 14 см. Перпендикулярно к основанию проведена прямая, делящая площадь данного треугольника пополам.*

The base of the triangle **is divided into parts of the lengths** 36 cm and 14 cm. Perpendicular to the base, a straight line, dividing the area of the triangle in half, is drawn.

***Разделим на 2 длины** высот треугольников.*

Divide by 2 the lengths of the heights of triangles.

Докажите, что высоты AA_1, BB_1 and CC_1 остроугольного треугольника ABC делят пополам углы этого треугольника.

Prove that heights AA_1, BB_1 and CC_1 of acute triangle ABC bisect the angles of triangle $A_1B_1C_1$.

We use (in Russian) the words “прямая”, “прямой” when we mean straight line or right angle. In English language there is not such difficulties.

Докажите, что точки A_1, B и B_1 лежат на одной прямой. Prove that points A_1, B, and B_1 lie on the same straight line.

Вписанный угол прямой. The inscribed angle is right.

Here are some fragments of task texts that include the polysemous word "image":

*N конгруэнтных окружностей, касающихся между собой попарно, касаются **внешним образом** окружности радиуса R.*

N congruent circles, tangent to each other in pairs, are tangent externally to a circle of radius R.

*Найти точку, которая является **образом точки A** при выполнении заданных отображений.*

Find a point that is **an image** of point A when performing specified mappings.

***Таким образом**, построенный треугольник является равнобедренным.*

Thus, the triangle built is equilateral.

In English, there are not analogical examples.

Ellipses. Ellipses or missing words in a sentence that are potentially restored by context, such as: “triangle is located in the left corner, square – in the right”. This sentence omits predicate and noun "corner" ("in the right corner").

The resolution of ellipses with a missed verb in the texts of planimetric tasks was considered in [3, 7]. These works propose an approach to restoring ellipses based on cognitive models of geometric objects and the relationship between them (as the deep structures of real sentences).

Consider one type of ellipses with a missed noun “length” and predicate “equal”, it is often found in planimetric tasks, for example, in such expressions: calculate the base of the triangle; the perimeter 2p of a quadrangle; a diamond with an acute angle β and side α; a circle with radius 4 cm; the sides of the triangle form an arithmetic progression; an equilateral triangle, the sum of base and height of which is equal to the double lateral side.

In these expressions we are talking about the length of perimeter equal to 2p, the length of the base of triangle, the length of side, equal to α, the length of radius equal to 4 cm, the lengths of the sides of triangle forming arithmetic progression.

It can be assumed that this kind of ellipses are associated with the properties of human cognitive sphere. Indeed, the examples show that the phrases "triangle’s sides," "sides of square", etc., that is, "sides of arbitrary figure in planimetry" are understood both as geometric objects, and as their lengths (values).

Let's call this phenomenon the transference of the object’s properties to the content or semantics of the object itself. That is, if the object has value, length, area, you can say "calculate the object", "object 2 cm" etc.

It is possible to note another phenomenon of the transference from a predicative form to an equivalent attributive form expressed by adjective or participle: has a length – having a length; circle touches a circle – tangent circle; sides form arithmetic progression – side forming arithmetic progression. There is a mental transference from action to its result: a triangle is given – a given triangle; a figure is built – a-built figure; is conducted – conducted, etc. In the semantics of object, the process is included as a result of which this object is formed. Or the semantics of object mentally includes the action the object produces. The phenomenon of transference can be used in the equivalent transformations of deep structure descriptions.

Ideographic synonymy: polysemy and partial intersection of lexical meanings of certain words. We consider this phenomenon in the following group of words: to divide, cut off, separate, break, dissect, bisect. These words are used when a figure or its element is divided into two parts or in some ratio, for example, “in half, into equal parts, in ratio 1:2, *m: n, m* to *n”,* or into 2 parts, the size of which is expressed in numbers.

The lexical meaning of words arises from the context of words. Thus, an analysis of the context of word “to divide” shows that this verb in the vast majority of cases in Russian is used when a figure or its element is divided into two parts. The words “разбивать” (to break), “рассекать” (to dissect) are mostly used when the result is several parts (more than 2) of figure or elements of figure.

In English, the verb "to bisect" is used with the same meaning that the verb "to divide" in Russian (the prefix "bi" comes from the Latin meaning "two' (for example, bicycle, having two wheels)). For example:

Определить отрезки, на которые стороны первого квадрата рассекаются вершинами второго квадрата.

Identify the segments into which the sides of the first square are cut by the vertices of the second square.

Вершины второго треугольника лежат на сторонах первого треугольника и делят каждую из них в отношении 1:2.

The vertices of the second triangle lie on the sides of the first triangle and divide each of them in ratio 1:2.

The **bisector divides an arc in halves.**

Prove that line AM **divides segment BC in halves**.

On heights of triangle ABC points A_1, B_1 and **C_1 that divide them in ratio 2 : 1** counting from the vertex are taken.

Prove that the broken line AOC **divides ABCD into two parts** whose areas are equal.

Anaphora: means the repetition of the same conceivable (thinkable) object in different parts of the sentence or in several coherent phrases.

In a particular case, anaphora is the repetition of the same word (the so-called exact repetition). In general, these are inaccurate repetitions. One of these repetitions is called **antecedent**, the other is a **reference**. As references in the inaccurate repetitions there are used nouns, pronouns, synonyms, etc. Most often pronouns are used. Examples of sentences with anaphoric reference:

Если какой-нибудь угол трапеции принадлежит двум ее сторонам, то они имеют общую точку.

If some angle of trapezium belongs to two its sides, then they have a common point.

Через точку пересечения диагоналей трапеции проведена прямая параллельно её основаниям.

In a trapezium, the straight line is drawn through the intersection point of the diagonals of trapezium parallel to its base.

The value of anaphoric pronouns can be fully described by the rules of their use through transformations [9], but the analysis of anaphoric pronouns has not yet been fully implemented.

One of the fundamental studies of anaphoric statements in speech acts was carried out in [10]. This paper notes that many existing theories and approaches to describing anaphoric relationships create an eclectic and disjointed picture. This paper provides for the first time a review and analysis of existing anaphora models. Both structural-linguistic anaphora theories, such as the binding theory, binding primitive theory, centration theory, accessibility theory, and discursive theories such as the waiting hypothesis, cognitive quantitative model, theory of the cohesion of discourse are considered. The strengths and weaknesses of each theory have been identified, and experimental research has been carried out on factors whose role in the process of establishing and decoding anaphoric dependencies has first been identified.

Based on the results, a new model of anaphora mechanisms has been proposed, describing both the speaker's position and the position of the recipient, as well as the point of common ground of these positions.

Quantifier expressions. Quantifiers are divided in the following groups [9]: a) **quantifier adjectives** – *всякий* (every, each, everybody, everyone, any, anybody), *все* (all, everybody, everyone), *любой* (any, any one), *каждый* (each, every), *какой-нибудь* (some, some kind of), *какой-либо* (some, any), *некоторый* (a certain, some), *один* (alone, one), *никакой* (no, not any), *ни один* (no one, nobody), *какой бы то ни было* (whatever); b) **quantifier nouns** – *всё* (all the, everything), *что-нибудь* (something, anything), *что-либо* (something); c) **quantifier adverbs of place** – *везде* (anywhere), *где-нибудь* (somewhere), *где либо* (somewhere else); d) **quantifier adverbs of time** – *всегда* (always), *иногда* (sometimes, occasionally, now and then, once in a way, now and again, once in a while), *когда-нибудь* (some day), *когда-либо* (sometime); e) **quantifier verbs** – *существует, имеется, есть, нет* (exists, is, does not exist, is not). For example:

*Если **любой** угол четырехугольника прямой, то это прямоугольник.*

If **any** angle of a quadrangle is right, then it is a rectangle.

*Из **всех** прямоугольных треугольников данной высоты h найдите треугольник наименьшего периметра.*

Of **all** the rectangular triangles of given height h, find a triangle of the smallest perimeter.

*Докажите, что для **всякого** прямоугольного треугольника имеют место неравенства…*

Prove that there are the inequalities for **every** rectangular triangle...

*В каком случае **каждое** из этих неравенств обращается в равенство?*

In what case does **each** of these inequalities turn to the equality?

*Докажите, что **все** прямоугольные треугольники, длины сторон которых образуют арифметическую прогрессию, подобны.*

Prove that **all** right-angled triangles, the sides' lengths of which form arithmetic progression, are similar.

The analysis of quantifiers is given in [9]. The analysis includes a list of deep lexemes sufficient to present the meaning of sentences with quantifiers and to describe the transformations allowing us to pass from deep structures with these lexemes to real sentences with their syntax representation. This work offers a special language to describe quantifiers in deep structures.

Including analytical (arithmetic, algebraic) expressions in the text: it is quite common for a task to make up an algebraic expression, that is, a formula for calculating a certain magnitude or interpreting a formula that is found in the text. Solving a task may require composing or transforming formulas.

Syntactic lexemes: these are words that have a unique set of syntax relations and demand unique requirements for the correctness of syntax structure of sentences.

As it is pointed out in [11, p.261], "in many cases adequate identification of a syntactic construction relies upon semantic agreement of words". Thus, semantic features can make a substantial contribution in syntactic parsing".

For example, the following words fall into the category of syntactic lexemes: *пропорционально* (proportionally), *сравните* (compare), *does not exceed* (не превосходит), *выразить через* (to express through), *противоположный* (opposite), *справедливо* (justly), *обратно* (back), *обращается в* (to turn into), *конгруэнтный* (congruent), *вывести формулу* (to derive a formula), *истинность* (the truthfulness), *аналогично* (analogously), *равносильно* (tantamount to), *соответственно* (respectively), *иметь* (to have), *таков* (such), *связаны соотношением* (connected by a ratio) and others.

Transformational analysis of the lexemes "*имеет/имеют*" (has/have), "*таков*" (such), and "*соответственно*" (respectively) has been given by E.V. Paducheva in [9].

3. Non-language difficulties in understanding texts of planimetric tasks

The non-language difficulties are caused, first, a vague text language that is not logically and linguistically clear. Secondly, these difficulties are induced by the necessity to attract general geometric knowledge related to objects and relationships in the texts of tasks. Thirdly, some difficulties are explained by the need to choose from several building options, or to formulate additional considerations (conditions) for drawing. The latter circumstance requires the involvement, in the process of drawing, various assumptions and logical conclusions.

An example of a lack of text clarity might be the task: "two circles of radii r and R (r < R) are located such that one of their internal tangents is perpendicular to one of their external tangents. Find the area of the triangle formed by these tangents and one of the internal tangents".

Some examples of involving the common geometric knowledge are:

"Three circles, the radii of which are 1, 2 and 3, touch in pairs externally. Calculate the radii of two circles, each of which touches to three given circles". (Here it is required understanding how to build two additional circles).

"Through point R, lying on the continuation of diagonal AC of quadrangle ABCD and the middles of sides BC and CD, are drawn two straight lines crossing sides of AB and AD, respectively, in points E and F. Prove that the straight lines EF and BD are parallel". (In this task, it is necessary to take into account that through two points you can draw only one straight line, and two straight lines should intersect at point R).

"Find the corners of an equilateral triangle if its altitude is half the bisector of angle at the base". (Here you have to decide what altitude is meant).

"A square is inscribed in the other square. Calculate a smaller angle between the sides of the squares if their areas are related as 2:3". (It is important to consider the position of the vertex of the inscribed square).

A line that has exactly one common point with a circle is called a line tangent to the circle. Through any point A outside the circle, exactly two tangents to the circle can be drawn. Let B and C be the tangent points and O the center of the circle. Then: a) AB = AC; b) ∠BAO = ∠CAO; c) OB ⊥ AB. Sometimes the word "tangent" is applied not to the whole line AB (AC) but to the segment AB (AC): the tangents to one circle drawn from one point are equal.

There are the tasks for which drawing their conditions is possible only after their solution. For example: "Is there a rectilinear polygon in which the length of one of its diagonals equals the sum of two other diagonals?"

Call these difficulties cognitive expectations. Cognitive expectations are apparently quite common when generating natural language texts. That is why we come to the idea of involving in the analysis of texts cognitive graphics and relations pre-formed in the system of solving planimetric problems. We can use them during the visualization of a task condition in dialogue with a user. The user can be a high school student, a teacher, and a schoolboy [12].

4. Conclusion

It may be expedient to overcome the challenges listed above in the framework of a unified approach by examining how the sentences describing geometric configurations are constructed.

In the 60s and 70s of the previous century, a systematic study of the meaning of language units larger than the word began – sentences and texts [9]. At the same time, the meaning of sentence began to be understood as a language expression, which would serve as an invariant of sentence in its synonymous transformations. Thus, the languages of the deep semantic level of expressions of in NL and the rules of transforming deep sentences into all possible expressions equivalent in meaning began to be proposed.

Information-logical language (ILL), described in [13] became one of the deep languages. The language of multi-sorted predicate calculus with some extensions is taken as the basis of this language so that ILL is not limited to geometric vocabulary. In practice, however, the language has proved to be quite complex. First, it can be considered as a descriptive language, and secondly, from the most general theoretical positions, it considers all possible syntax control models for each lexeme.

However, when we deal with the text of geometric tasks, the syntax models that are consistent with the geometric content of the texts have the priority. It is more appropriate to consider the visual image (drawing) of a geometric situation displayed in a sentence as a sentence's deep model invariant to it in meaning. This invariant can be considered as a cognitive model of the geometric situation, or conceivable image of sentence. The language of cognitive model description can be a generative language similar to a programming language. At the lowest level, it is advisable to place operations or commands: "build a figure", "draw a line", "drop perpendicular from point to line", etc. The next level can be combining noun, verb, propositional phrases, and other clauses consistent with the expressions of NL, meeting in the real texts of the tasks and describing the constructions created by the lower-level commands. The third level can be associated with the inclusion of logical quantifiers, if necessary, and possible transformations of cognitive scheme descriptions.

The process of understanding the text of task at the entrance of a automated solver turns into two step-by-step sub-processes going towards each other: constructing a drawing of the text of task (cognitive scheme) and generating a text describing the cognitive scheme as accurately and fully as possible.

In parallel, there may be an equivalent transformation of the description of the cognitive scheme being formed to the real text of the original geometric task, including the transformation of complete sentences into elliptical ones. The convergence of the converted and the original texts will prove that the original sentence is correct.

The language of cognitive scheme description has yet to be created. But such a work has already begun. A library of cognitive schemes is being developed with their visualization and description in NL and syntax diagram. Some approach to using cognitive schemes in the analysis of texts in planimetry are described in [3, 13].

5. Acknowledgements

The research was partially supported by the Russian Foundation for Basic Research (RFBR), project no. 18- 07-00098a.

6. References

[1] S.S. Kurbatov, A.B. Voroybev, Ontological solver of geometry tasks described in natural language, in: Proceedings of the 15th National Conference on Artificial Intelligence with International Participation, CAI-2016, 1 (2016), 56–63.

[2] S.S. Kurbatov, I.B. Fominykh, A.B. Vorobyev, Software support for an algorithmic and cognitive agent based on the Polya methodology, Software products and systems 1 (2018) 12–19.

[3] X.A. Naidenova, S.S. Kurbatov, V.P. Ganapol'skii, An analysis of plane task text ellipticity and the possibility of ellipses reconstructing based on cognitive modelling geometric objects and actions, in: A.M. Elisarov, N.V. Lukachevich (Eds.), Proceedings of Computational Models in Language and Speech Workshop, CMLS 2018, colocated with the 15th TEL International Conference on Computational and Cognitive Linguistics, TEL-2018, vol. 2, the Academy of Scieces of RT, Kazan, RT, 2018, pp. 70–85.

[4] N.P. Antonov, M.Ya. Vygotsky, V.V. Nikitin, A.I. Sankin, Collection of the tasks on elementary mathematics, Nauka, Moscow, Russia, 1979.

[5] E.G. Gotman, Z.A. Skopets, Solution of geometric tasks by the analytical method, Prosveschenie, Moscow,Russia,1979.

[6] A.N. Kolmogorov (Ed.), Geometry: Textbook for 8th grade high school, Prosveschenie, Moscow, Russia, 1980.

[7] X.A. Naidenova, S.S. Kurbatov, V.P. Ganapolsky, Cognitive models in planimetric task text processing, International Journal of Cognitive Research in Science, Engineering, and Education, 1 (2020) 25-35.

[8] V. Prasolov, Tasks on geometry, 5th ed., Moscow Handbooks, Moscow, Russia, 2006.

[9] E.V. Paducheva, About semantics of syntax, URSS, Moscow, Russia, 2019.

[10] V.K. Prokopenya, Anaphora mechanisms in speech generating and understanding: experimental research with the use of Russian language material, Ph.D. Thesis, St. Petersburg State University, Russia, 2016.

[11] Y. Apresyan, I. Boguslavsky, I. Iodmin, A. Lasursky, L. Mityushin, V. Sannikov, L. Tzinman, Linguistical maintenance of the system ETAP-2, Nauka, Moscow, Russia, 1989.

[12] S.S. Kurbatov, I.B. Fominykh, A.B. Vorobyev, Interactive visualization of cognitive structures in an integrated system, in: Proceedings of the 17th National Conference on Artificial Intelligence with International Participation, CAI-2019, 1 (2019) 222–230.

[13] A.V. Kuznetsov, E.V. Paducheva, N.M. Ermolaeva, About information language for geometry and an algorithm of the translation from natural language to information language and back, in: Linguistic research on machine translation, VINITI, Moscow, 1961, pp. 40–74.