Article

The Think-Aloud Method for Evaluating the Usability of a Regional Atlas

Tomas Vanicek [ID] and Stanislav Popelka *[ID]

Department of Geoinformatics, Palacký University Olomouc, 17. listopadu 12, 77900 Olomouc, Czech Republic
* Correspondence: stanislav.popelka@upol.cz

Abstract: The think-aloud method is a widely used method for evaluating the usability of websites and software. However, it can also be used with cartographic products, an area which has been neglected up to now. It is a method in which test participants verbalise all their thought processes aloud. The participants are given a test scenario containing tasks to be completed. The method aims to reveal the participants' subjective attitudes toward a product in order to evaluate its usability. The present paper describes the use of the think-aloud method to evaluate the usability of a cartographic work—the regional atlas *of the Moravian-Silesian Region*. The study includes (I) a complete review of the method, based on the studies conducted; (II) testing tools for working with recorded data; (III) designing an experiment for evaluating the usability of the atlas; and (IV) the resulting qualitative and quantitative evaluation of the atlas based on the obtained results. During the study, three approaches were proposed to process and analyse the audio recordings. The first option was to separate the audio recordings into individual annotations and analyse them. The second option was to convert the recordings to text and perform a linguistic analysis. The third supplementary option was to use all the material produced and to analyse it subjectively and retrospectively, from the researcher's perspective. All three options were used in the final assessment of the atlas. Based on the participants' statements, any shortcomings in the studied atlas were identified for each topic (e.g., non-dominant maps or exceedingly complex infographics), and recommendations for their elimination were proposed.

Keywords: think-aloud; cartography; evaluation; usability; experiment

1. Introduction

Usability is an essential property of a product and is determined to be one of the most important properties of a suitably designed system. Participants must feel comfortable working with a product while they efficiently find the information needed. A product with good usability should reduce the cognitive load of the user to a minimum and be clear and understandable. According to ISO standard 9241 [1], usability has three aspects (effectiveness, efficiency, and satisfaction) by which specific users achieve specific goals in a specific environment. In usability testing, effectiveness addresses the accuracy and completeness of answers (e.g., the success rate of questions answered), efficiency addresses how answers are achieved (e.g., the time taken to complete a task), and satisfaction focuses on participants' attitude and comfort. The most common method for usability assessment is through user testing. The goal of this form of testing is to uncover the most serious issues that may cause problems for users [2]. The level of usability can be determined using various methods (e.g., questionnaire, interview, focus group, and eye-tracking). For this study, the think-aloud method was chosen. The purpose was to see if the method could be applied to a more complex cartographic product, specifically the evaluation of an atlas. According to Nielsen [3], think-aloud is the most valuable method for evaluating product usability. Through this method, any system, whether a physical product, a web environment, or software, can be evaluated qualitatively and quantitatively to improve its

usability. According to Chen [4], think-aloud is an effective method for collecting many product statements related to user satisfaction while performing tasks. The method is still used by many researchers [5–11].
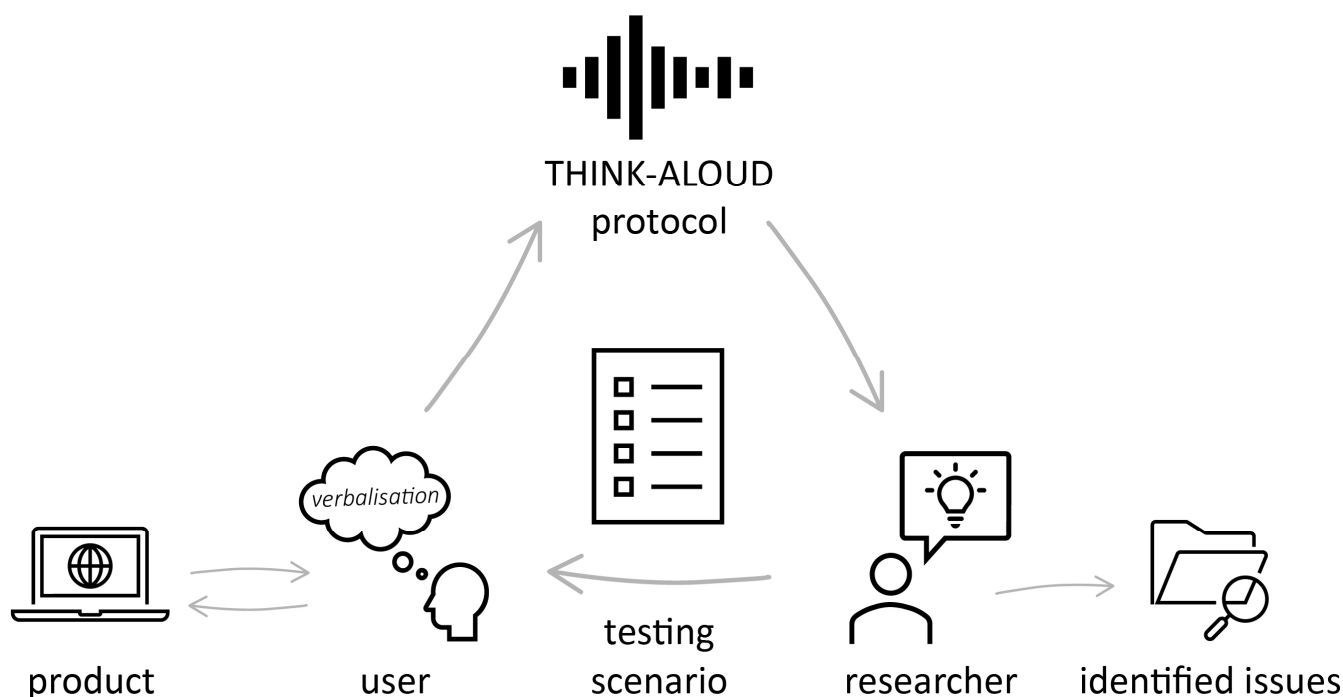
1.1. Principle of the Think-Aloud Method

The basic principle of the think-aloud method is that the participants verbalise all their thought processes aloud while working with a product. This principle allows the researcher to gain a deeper insight into the participant's cognitive processes. The output of the verbalisation is a verbal protocol (i.e., recorded verbal statements). By analysing the participants' statements, it is possible to obtain their attitudes toward a product's design and to find out their expectations for the product. In this way, a model of the participants' behaviour is established, and this can be used to modify the product [12]. For the process of verbalisation from working memory, only verbal information can be considered. Thus, it is important to think carefully about the test scenario (tasks given to participants) that directs the testing process. If a participant is given a task with an elevated level of cognitive load, then there may be a problem with the verbalisation of ideas and an inability to transfer information to working memory. Alternatively, if the participant is given a task that is too easy, it can be solved automatically, which is again difficult to put into words [13]. A model of the human cognitive system is given in Figure 1.
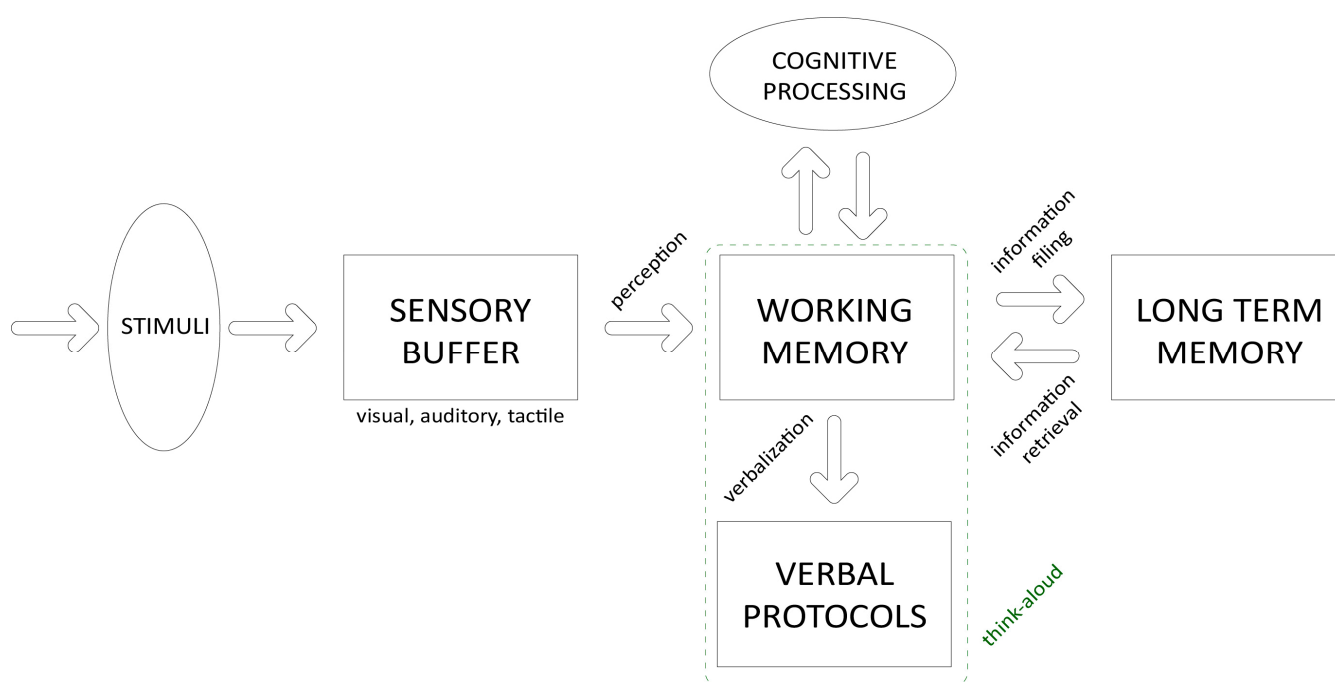
Figure 1. Model of the human cognitive system.

A cognitive process can be defined as a sequence of thought states that are stored primarily in working memory. To obtain a verbal protocol, participants must verbalise their thoughts. According to Ericsson and Simon [14], the sequence of verbalised thoughts from working memory is the same as in the cognitive process that is performed without any thinking aloud. Thus, the think-aloud method does not interfere with participants' thought processes during a task. Participants solve an assigned problem and concentrate on their task while verbalising their thoughts automatically. These are data direct from the participants' working memories, and there is no delay. Participants are not forced to supply any additional interpretation concerning their thoughts, nor are they obliged to transform them into a structured form, as is the case with various other methods. According to Someren [15], the think-aloud method has two main parts, consisting of (1) the research participants thinking aloud while solving a problem and (2) the researcher analysing

the verbal protocols obtained. The participants work with a product in a controlled environment and respond to a specific set of prepared questions and tasks while thinking aloud. They verbalise all their thoughts aloud continuously throughout the testing. These thoughts should only be verbalised by the participant without providing any additional interpretation, in order to maintain objectivity. The speech may also include various thoughts of the participants, such as what they are looking at, what they are thinking, what they are doing and how they are feeling. According to Bláha [16], participants express their emotions during the experiments, whether negative, such as confusion or frustration, or positive, such as joy at having mastered a task. The participants' statements are recorded in a recording device for the purpose of archiving the communication and subsequently analysing it. The audio/video recording of statements is then transcribed and analysed, both qualitatively and quantitatively. A usability evaluation of the product is thus generated. During the testing process, the observer takes notes on what the participants say and do without interfering with the experiment process. In particular, the observer notes the areas where the participant encounters difficulties, and these areas should be more focused on during the evaluation. In addition to audio, the testing is often videotaped so that the researcher can go back and refer to what the participants did and how they reacted [17]. A simplified principle of the think-aloud method is clearly shown in Figure 2.

Figure 2. The principle of the think-aloud method.

The most common option for analysing voice statements is to segment audio recordings into annotations or to transcribe the recordings into textual form. The text can then be used for linguistic analysis. The annotated segments correspond to individual words or sentences. The annotations are assigned to the segments based on a coding scheme designed in advance by the researcher. The individual annotations may correspond, for example, to the time taken in reading a question or the time taken in performing a task or to other actions, objects, and shapes, depending on the research. The annotation of the record is performed using special applications. One example is the ELAN tool. Once segments are annotated, their statistics can be used to calculate usability metrics.

According to Kuusela and Paul [18], the course of the think-aloud method can be split into two different types of experimental procedure. The first is the concurrent think-aloud protocol (CTA), collected during the solution of a task. The second is the retrospective think-

aloud protocol (RTA), collected after the task. In this second phase, the participants go back through the video recording and once again apply the think-aloud method. According to Guan et al. [19], the retrospective version of think-aloud is suitable in combination with the eye-tracking method and is often stimulated by visual reminders, typically video footage.

As with any usability assessment method, the think-aloud method has its advantages and disadvantages. One advantage is that participants' ideas and needs can be elicited during testing. In addition, ongoing verbalised information helps the evaluator to better locate the sources of potential problems. Thinking aloud helps some participants to concentrate on the task at hand and makes them aware of the facts as they think about the task through verbalisation. The course of the experiment is controlled by the test scenario, not by the individual. Moreover, the nature of the method allows the researcher to implement it distantly and/or to combine it with other methods. Testing can be conducted at any stage of product development, regardless of whether the product is in its final form or not. For certain participants, the process of testing may be natural because they are working with a product as they normally do, which may yield telling answers. On the other hand, for other participants, thinking aloud may be unnatural and distracting. The researcher also needs to choose participants carefully based on several criteria, and preference should be given to potential users of the product. At the same time, the participant must be able to fully verbalise their thoughts. Thus, people of an extrovert nature are most often chosen for testing. According to Alnashri [20], extroverts are able to identify a higher number of usability issues, have a higher success rate in completing tasks, and are more comfortable verbalising their thoughts. The output is an unstructured recording, and for a complete evaluation of a product, it is necessary to code it according to a prepared coding scheme. Longer testing can also be challenging and exhausting for participants. Thus, participants may not be as thorough in verbalising later tasks as they were at the beginning of the testing. If participants lose track of the task or have a question during the experiment, an observer must intervene, and this may affect the participants' normal behaviour. It is important to clearly identify this part of the recording during subsequent transcription and coding of the recording.

1.2. General Origins of the Think-Aloud Method

The think-aloud method has its roots in psychological research based on methods of introspection. The subject of introspection is the content of an individual's consciousness and how events in the external world are observed. Although the results of introspection have been highly successful, the approach was very complex and could only be carried out by very experienced psychologists. The solution was to simplify the processes using verbalisation and to assume that only the contents of an individual's working memory were subject to verbalisation, instead of dealing with the complete cognitive system. From an introspection perspective, the research data are the events that take place in a participant's consciousness. These events need to be analysed and explained by an experienced psychologist. However, the introspection data were fundamentally accessible only to the single observer who evaluated the study. This prompted researchers to develop a new approach that treated the statements as real data. Thus, the innovative approach would allow the output data to be archived and thus create the possibility of interpretation of the data by any person [15]. Based on these considerations, methods that operated on the principle of verbal protocol analysis were developed. The think-aloud method as it is known today for usability testing purposes originated in the 1980s [21]. In Ericsson and Simon's version [14,17,21], the interaction between the researcher and participant is minimal, and participants are not asked to filter, analyse, explain, or interpret their thoughts. This is the case even if their verbalised thoughts are difficult for the researcher to understand. However, according to Boren and Ramey [22], there are some significant differences among the designs for the think-aloud protocols that Ericsson and Simon [21] suggest. The differences arise from the specific needs and contexts of usability testing. Practitioners should be aware of these specific needs and should adapt their approach to

suit them while still collecting valid data. In some cases, it is necessary to ask additional questions of the test subject, which Ericsson and Simon [21] do not agree with, as this could influence the participant. The seminal publication by Van Someren [15] summarises the progress of the think-aloud method in the 1980s and identifies it as one of the most appropriate ways to uncover a participant's significant thought processes. Currently, the think-aloud method is accepted as useful by a significant part of the scientific community, not only those working in psychology [15]. The method is still used for the usability testing of many different products [9,11,23–46].

1.3. Applications in Cartography

Although many researchers have used the think-aloud method and published their results, this method is not so common in cartographic research. More complex methods such as eye-tracking or traditional methods such as questionnaires are used to evaluate the usability of cartographic products. Nevertheless, this method has been used several times by researchers in cartography. Popelka et al. [47] combined the think-aloud method with eye-tracking for the evaluation of weather forecast web maps. Knura et al. [37] addressed the question of whether the think-aloud method can provide comparable results to the more common eye-tracking method in the evaluation of maps in the context of the COVID-19 pandemic restrictions. In addition to participants' statements, screen and mouse movements were also recorded during the experiments and were part of the analysis. Using this information, the researchers were able to code the approximate location of a participant's attention on the map for each second of the interview. This allowed them to use the same visual methods used in eye-tracking studies, i.e., attention maps or trajectory maps. The results of this study showed that the think-aloud method can fully replace eye-tracking testing and can achieve comparable results when determining the usability of maps. Compared with eye-tracking, the main limitations of the think-aloud method are the lower accuracy in terms of which part of the product the participant is attending to and the longer time and labour-intensive nature of the manual coding process. In the field of cartography this method has been used by Nivala et al. [48] in combination with screen recording and questionnaires to assess potential usability issues in web map portals. The method was used as a complement to determine the participants' opinions of individual map environments. Furthermore, the method was used by Ooms et al. [49] in their study of participants drawing maps that had been previously shown to them. The aim was to find out what objects were drawn and in what order, depending on the participant's previous experience. The method's use in cartography was discussed in detail in Kulhavy's [50] study, where high school and university students were asked to study a thematic map while verbalising their thoughts. The aim was to find the differences in map orientation between these two groups of students. The study found that the cartographic experience of university students had a large effect on the outcome. They were more likely to use legends and map descriptions for easier orientation of each topic. Gołębiowska [51] focused on the distribution of legends for thematic maps. Her study aimed to understand how the legend influences the map reading process. Her research showed that participants preferred legends that were simple or familiar. The study used a mixed research design that employed both a quantitative (usability metrics) and qualitative (identifying problem areas) approach. This strategy allowed the analysis of a complex process that would not have been possible if only one approach had been used. The study identified four problem-solving strategies used by the participants and resulted in the definition of principles for the design of the legend. The think-aloud method for evaluating indoor navigation was also used by Viaene et al. [52]. The aim was to identify which of a number of landmarks were the most important. Based on the participants' statements and behaviour when moving inside the building, the study contributed to the formulation of adequate instructions for indoor wayfinding. Similarly, Kettunen et al. [53] used the method to describe outdoor landmarks to support wayfinding in combination with participants' drawings. Quaye-Ballard [54]

used the think-aloud method to test the usability of a prototype application developed for real estate agents in visualising buildings in a 3D environment (using virtual reality).

The range of areas mentioned in this literature search suggests that the think-aloud method is fully applicable in the field of cartography. Its advantage is that it can replace the more expensive eye-tracking testing, or it can be combined with eye movement recording.

1.4. Objectives of the Study

The primary aim of the study was to design an experiment to evaluate the usability of a regional atlas. The purpose of the study was not only to find any weaknesses in the atlas but also to prove if the think-aloud method was suitable for the evaluation of an atlas. This main goal was achieved through the completion of several tasks. The first task involved a comprehensive review of the method, based on previous studies. The second task focused on testing tools for data processing. The third task concerned the design of the experiment itself. The final task summarised the results of the evaluation of the atlas based on the responses obtained. Three approaches were used to process and analyse the recorded data, and their combination resulted in a comprehensive evaluation of the atlas.

2. Materials and Methods

2.1. Evaluated Product

The study focused on the evaluation of the Atlas of the Moravian-Silesian Region: People, Business, Environment [55]. The study aimed to determine the usability the atlas and identify potential issues concerning selected topics in the atlas, as well as collecting participants' opinions on the product. The atlas presents spatial information using cartographic methods in the form of thematic maps and graphic illustrations. The atlas is divided into several chapters containing maps, graphs, diagrams, and illustrations. Figure 3 shows examples of the pages from the atlas that were used in the study.

2.2. Testing Scenario

A test scenario was created before the actual experiment started. This consisted of questions and tasks for the participants, the wording of which would fundamentally influence the results of the think-aloud method. The atlas is a complex product, and testing all its parts would not be realistic. For this reason, only selected topics were included in the testing, and these were chosen in consultation with the authors of the atlas. Cognitive tests were not used to develop the questions because of the wide range of potential product users (and participants). On the other hand, the tasks created were (subjectively) judged to be relevant to the needs of the research and did not require any expertise on the part of the participants. Since the general public can be users of the atlas, the tasks have been designed so that the "average" person (regardless of age) can solve them. For the design of the tasks, Someren's strategy [15] was followed with two basic considerations.

The first consideration was whether the tasks were at an optimal level of difficulty given the appropriate cognitive load on the participant. Test participants needed to be able to solve tasks in a virtually automated manner, but at the same time, the tasks must be sufficiently difficult. For this reason, care was taken in the design of the tasks to ensure a reasonable level of difficulty.

The second consideration was whether a task was representative of the problem being solved and whether the participant could be asked to solve it. The questions had to be clearly defined, and participants were not allowed to customise them in any way. As a result, the tasks had to conform to an optimal form of complexity, and participants needed to be able to verbalise the course of their problem-solving from working memory. At the same time, tasks were created to be more open-ended in order to limit the strictness of the answers and allow the participants more space to express their ideas.
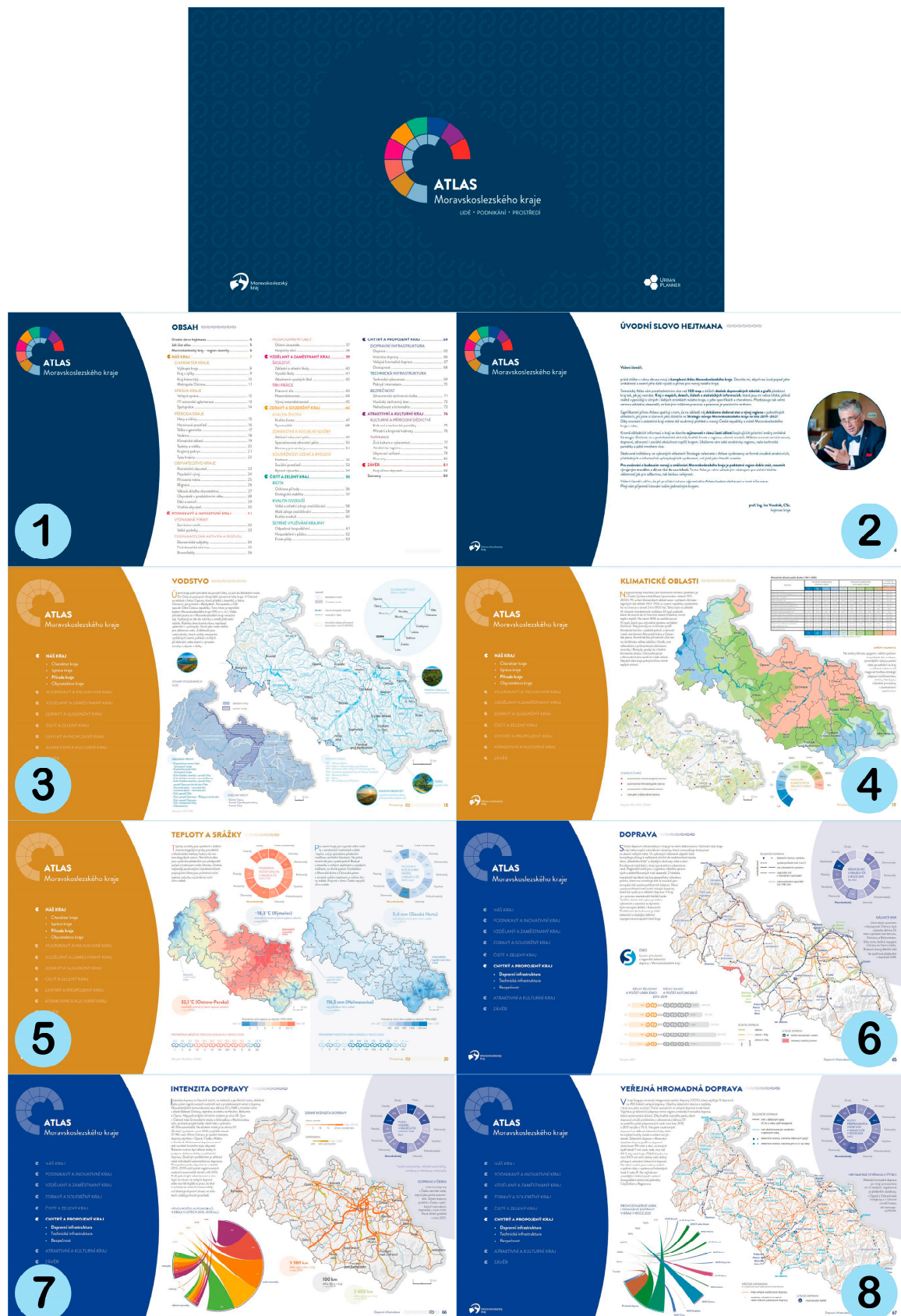
Figure 3. Examples of the content in the *Atlas of the Moravian-Silesian Region*.

In total, participants were asked 19 specific questions under ten headings. The first three questions introduced participants to the atlas and analysed how well the participant could manipulate and search the atlas. The remaining questions focused on specific topics described in the atlas. Within each topic, sub-tasks focused on the atlas's characteristic elements. As the questions were related to attributes or spatial information (location of an object), this part of the research considered the full range of information conveyed by the map. In addition, answering two questions such as "what?" and "where?" requires the activation of different parts of the brain [56]. The last question asked participants to find information in the atlas that was of personal interest to them. This task made it possible to find the participants' preferences for a given product. Before the study began, one training task was conducted to habituate the participants to thinking aloud. As the atlas is in Czech, all tasks were also formulated in Czech. Translations of these tasks are given in Table 1. Participants worked through the questions one by one (from 1 to 10). Participants strictly followed the test scenario, which was an advantage because the researcher did not have to interfere in the testing process. At the same time, there was no time limit for completing the task. The first three questions served primarily to get the participants used to working with the atlas. More important were the subsequent topics (starting with question 4).

Table 1. The test scenario in the study for the *Atlas of the Moravian-Silesian Region*. The related page of the atlas is specified in the third column.

| Mark	Task	Image Number (see Figure 3)
1	What page is the CONTENTS of the atlas on?	1
2	How many maps does the atlas contain in total?	2
3	What topic is marked in brown?	2
4	Find the topic WATER in the atlas and describe the maps in this section.	3
4a	Which basin is numbered 411?	3
4b	Which sea do most of the rivers in the Moravian-Silesian Region flow into?	3
4c	What is the longest river in the Moravian-Silesian Region?	3
5	Find the topic CLIMATE AREAS in the atlas and describe the maps in this section.	4
5a	What do the colours indicate on the main climate map?	4
5b	What type of climate area has the largest area in the county?	4
5c	What type of climate can be found around Rýmařov?	4
6	In the atlas, find the topic TEMPERATURES AND RAINFALL and describe the maps in this section.	5
6a	Which parts of the region are the warmest and where are the coldest?	5
6b	What was the AVERAGE ANNUAL TEMPERATURE in the Moravian-Silesian Region in 2020?	5
6c	In which part of the territory can you see the highest annual rainfall?	5
7	Find the topic TRANSPORTATION in the atlas and describe the map in this section.	6
7a	What is the predominant mode of transport in the region?	6
7b	In what part can we find the NORTHEAST track?	6
8	Go to the topic TRANSPORT INTENSITY and describe the map in this section.	7
8a	What parts of the region have the highest and lowest traffic volumes?	7
8b	What is the rate of growth in the NUMBER OF PASSENGER CARS in the region?	7
9	Go to the topic PUBLIC TRANSPORTATION and describe the map in this section.	8
9a	What types of transport can I use to get to Ostrava Airport?	8
9b	Which company is the largest operator of public transport lines in the region?	8
10	Find any information in the atlas that interests you about the region (i.e., think of a question and find the answer).	X

The test scenario was printed on A4 paper and placed in a plastic sleeve, providing the participants with a visible reference for the questions. The questions were spaced widely and colour-coded to facilitate navigation for the participant.

2.3. Participants

Thirteen participants from different age groups (7 × 18–30 years, 1 × 30–45 years, 3 × 45–60 years, and 2 × seniors) were selected for testing. The key criterion for selecting participants was their experience in map reading. Participants with a cartographic background are better than non-professionals at focusing on different aspects of a map when solving tasks. For this reason, seven people without cartographic education and six people with at least some basic cartographic knowledge were selected. Some of the participants were selected because the geographical location of their residence was in the study area, and this ensured that the participants had knowledge of the local environment. This strategy was chosen to ensure the relevance of the test results to potential users of the product; those who are familiar with the area depicted can therefore provide more information on its participant-friendliness and effectiveness. To follow the principle of the method, it is preferable to select participants with an extroverted nature [20,57], and extroverts were preferred for this study. However, this characteristic is difficult to predict in advance, so before the testing began, participants were given a logic task to find how easily they could verbalise their thoughts.

2.4. Experiment Design

The experiment was conducted in a controlled and comfortable environment, specifically a soundproof room with a comfortable chair. The researcher interfered minimally, if at all, during the experiment to avoid impacting on the participants' thought processes. To achieve this, the observer was seated at a suitable distance behind the participants. The testing room was equipped with a high-quality Trust GXT 232 microphone with a filter (with a 48 kHz sampling rate), and the researcher had note paper, a pencil, and the test scenario (as depicted in Figure 4).
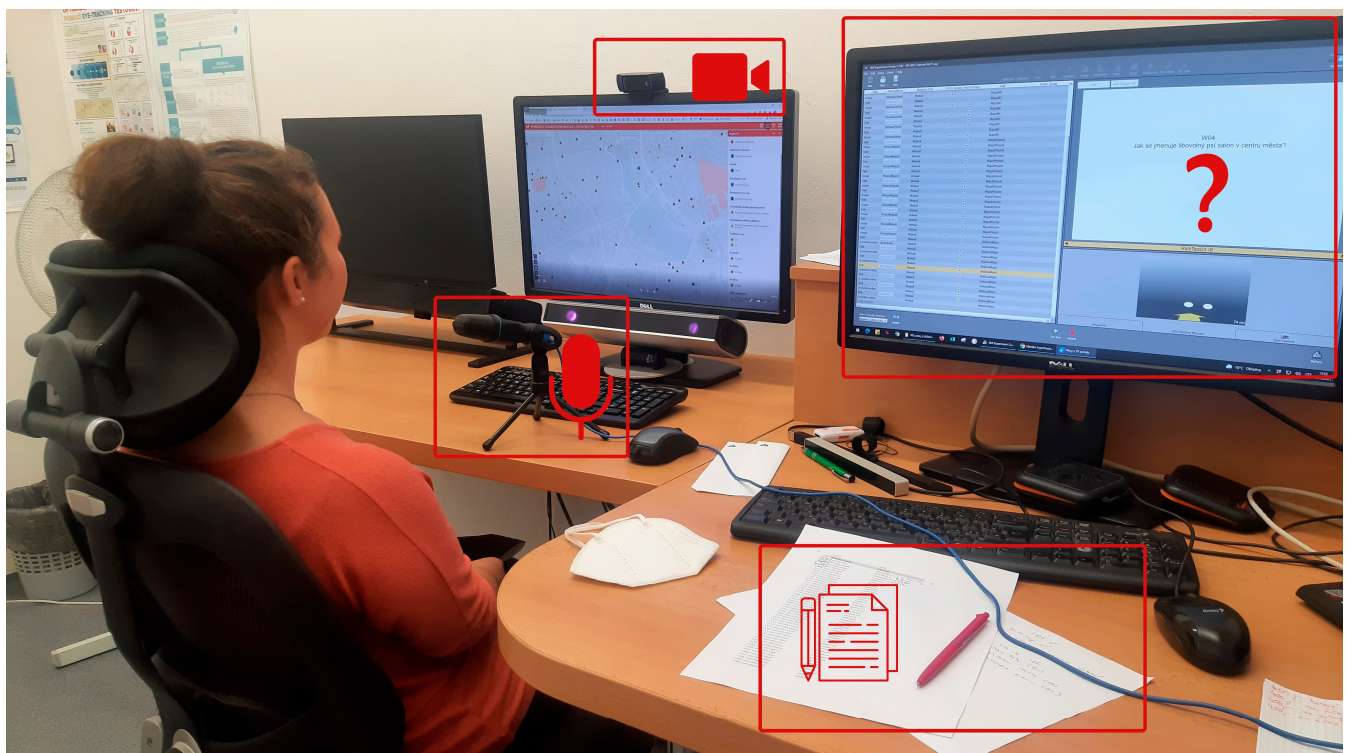
Figure 4. Testing controlled environment in the (pilot) study.

The instructions for the think-aloud method were straightforward, and they directed the participants to simply "do the task and say out loud whatever comes to mind." To enhance the clarity of the task assignments, the instructions were written down and read out to the participants sequentially. As noted by Someren [15], it was deemed inappropriate to use phrases such as "tell me what you think" as this may prompt participants to consider their specific opinions or to evaluate their thoughts, and this could reduce their comfort level during the experiment.

Upon the participants' arrival, the first step was to greet them and provide an overview of the testing and its objectives and to provide them with information on the confidentiality of the data collected. Participants were also reassured that "any answer is good" and that there was no need to fear giving incorrect answers.

Next, the participants were introduced to the product and the test scenario and informed that the researcher would not interfere with the testing process. The document consisted of two parts: the test instructions on one side and the introductory task on the other. This logic task was presented to the participants prior to the actual testing, with the aim of familiarising them with the practice of thinking aloud and enabling the observer to assess the ease with which the participant could verbalise ideas. This task was easily completed by all participants.

The testing was then initiated, with the observer maintaining stillness and avoiding speech and noise while taking notes to minimise interference with the participants' thought processes. The observer only prompted the participants with the phrase "Please continue talking" in instances where a participant had ceased speaking completely. The testing was recorded using a microphone, and the audio recording was stored on a computer. Upon completion of the testing, participants were given the opportunity to ask any questions. The participants' satisfaction with the evaluated atlas was assessed through an interview in which they shared their impressions of working with the atlas. This information was subsequently incorporated into a subjective analysis.

2.5. Data Analysis

After testing, the output was in the form of audio data (verbal logs). During the study, three directions were defined on how to use, process, and analyse the measured data (see Figure 5). The first and most fundamental option was the production and analysis of annotations, which is the main essence of the think-aloud method. The audio data were separated into segments, the participants' activity was logged manually for these segments, and the segments were annotated. The statistics of annotations were then used to calculate usability metrics. The other two proposed analysis options were additional options for the final evaluation. These were linguistic or retrospective analyses. All three analysis options were used to evaluate the usability of the atlas. These methods are described in more detail below.
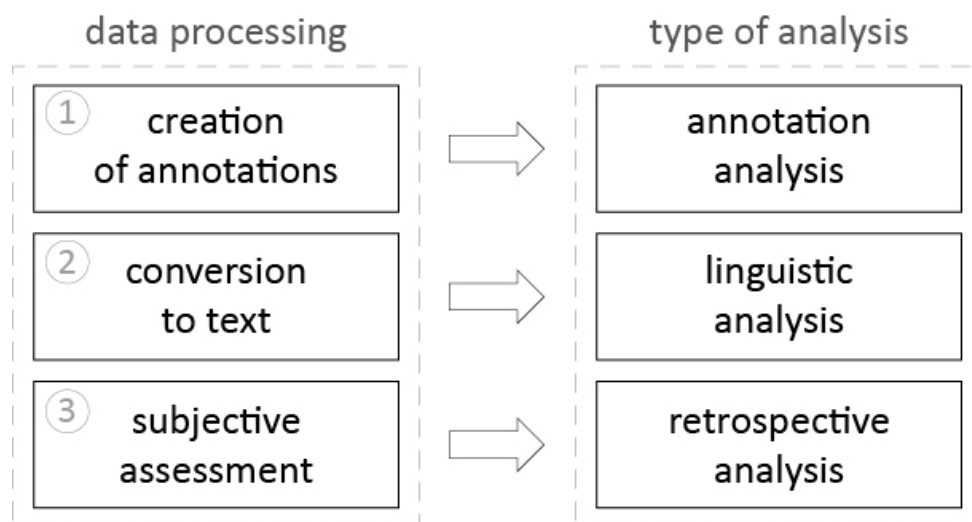
Figure 5. Possibilities for audio data processing and analysis.

2.5.1. Evaluation: Annotation Analysis

The think-aloud method is used to determine the usability of the product. Measured audio recordings are an unstructured source of information and are unsuitable for calculating usability metrics. For this reason, the recordings had to be segmented into different lengths and annotated according to a coding scheme. The annotation statistics were then used to calculate the metrics. Given the nature of the method, this principle was the most valid choice among the options for processing the audio data and carrying out the subsequent analyses.

Before the annotation of the recording, a coding scheme was designed. The categories were designed according to a defined research question that aimed to identify potential problems with the topics in the respective areas of interest (e.g., map field, title, etc.), in combination with a test scenario. Since the atlas is in Czech, the coding scheme was also formulated in Czech. The translation of the scheme is given in Table 2.

Table 2. The coding scheme used for audio processing (annotation).

Category	Description	Example
Reading the question/task	The participant is reading the question or parts of it	"Find the water theme in the atlas and describe the maps in this section." " . . . water theme and describe . . . "
Handling the atlas, reading the contents	The participant scrolls through the atlas, searches for relevant topics, reads the content	" . . . I'm going back to the content, topic to the right, I'm going back to the top. To the right, I am looking for. To the right it is page 127, I'm going to page 127."
Reading a map description, text, or title	The participant reads the description around the map	"Public transport is operated in 13 cities in the region."
Orientation in the map and legend	The participant searches for information and navigates in the given thematic map	" . . . the narrow-gauge regional line is marked with dashed lines and it is on the map between Osoblaha and Třemešná . . . "
Reading graphic and infographic elements	The participant searches for information in additional infographics, graphs, and images	" . . . here's an infographic on that, with it, directly broken down into auto bike buses and trucks . . . "
Finding information and answers	The participant has found the requested information	" . . . so, it's probably the Oder River, just by pure reasoning from the map."
Other	Other comments and actions of the participant when thinking aloud and the participant's opinion	"Might be a good idea to check the list to see if there's anything there." "I see I was wrong . . . "

A problem with the participants' statements can be the unintelligibility of the recording, as people may have a habit of not finishing words, or they may mumble. It also happened that participants interrupted their statement in the middle of a word, whispered, or made various unintelligible sounds. It was therefore not always possible to understand everything that was said. Such parts of the recording were labelled with the segment "unintelligible". According to Someren [15], in psychological research, every statement is relevant, and so all parts of a recording where the participant or even the observer spoke should be annotated, i.e., annotate everything recognisable in the recording.

The ELAN program, which is the most common annotation tool and has been used in several think-aloud studies [51,58], was used to create and analyse annotations. ELAN is an annotation and transcription tool for audio and video recordings. Using this tool, an unlimited number of textual annotations can be added to the audio recordings. An annotation can be a sentence, a word, a comment, a translation, or a description. They can be created on multiple layers, called tiers, and linked hierarchically [59]. ELAN provides several ways to display annotations. Each view is attached to the timeline of the recording and synchronised. By segmenting the recording, it is structured (see Figure 6). Already annotated segments can be compared with each other, and basic statistics about

their number or length can be collected to calculate usability metrics. A great advantage of the program is the segmentation into multiple levels. Various levels of annotation can serve different purposes, such as transcribing a portion of a record, classifying it into categories, analysing the attention paid to a specific part of a map, describing a participant's activity, an observer's note, etc. In this study, three levels of annotation were created: categorising by the coding scheme, identifying by question number, and recording correct or incorrect answers.
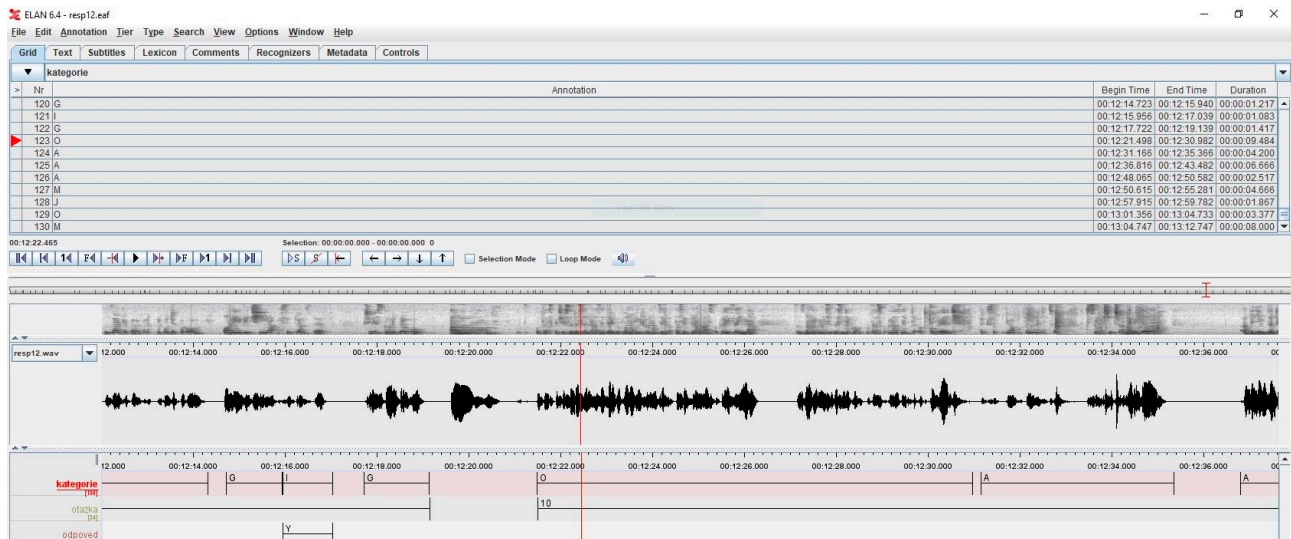

Figure 6. ELAN annotation environment.

Participants were asked nineteen questions under ten headings. Basic usability metrics (effectiveness, efficiency, and satisfaction) were used to assess the usability of specific parts of the atlas. These metrics are often used in cartographic research into map effectiveness. Using structured audio recordings from the annotation analysis, these metrics can be easily applied. The first metric was the calculation of success scores at three levels—participant success scores, success scores for individual tasks, and overall atlas success scores. Next, an error matrix was created and evaluated for the error rate analysis, and the average task completion time was also collected. The findings from these calculated metrics were applied in the overall conclusion of the atlas evaluation.

2.5.2. Evaluation: A Linguistic Analysis

Linguistic analysis was an additional option for the analysis of the audio recordings. The emphasis is on qualitative, not quantitative analysis, i.e., determining the meaning-forming elements contained in the acquired data and interpreting their meaning [60]. In the final stage, interesting results can be obtained from the analysed corpus, especially when determining the frequency of occurrences of significant words, keywords (in the case of comparison with reference data), the emotional meaning of words, and other characteristics of the text set. To perform an effective and meaningful linguistic analysis, a word corpus must contain at least several thousand words. For small text corpora, linguistic analysis is meaningless. The growing interest in exact-based research in linguistics and other humanities disciplines is creating a demand for analogous applications that make quantitative methods accessible to a wider range of participants [61]. Many tools exist for linguistic analysis, the majority of which are created and programmed directly by linguists. In total, sixteen tools for linguistic analysis were tested—AntConc, ATLAS.ti, Sketch Engine, QuitaOnline, QuitaUp, WaG, KonText, SyD, Morfio, KWords, Treq, Lists, InterText, Corpus Calculator, UDPipe, and Orange. Based on the analysis of the tools, Sketch Engine [62] was chosen as the tool that was able to analyse the textual data most efficiently and comprehensively. This tool has also been used by other researchers to

analyse questionnaires on the topic of school atlases [63]. It is a corpus (speech sample) management and text analysis software that has been improved since 2003. Its purpose is to enable people studying linguistic behaviour to search large text collections for complex and linguistically motivated queries. Sketch Engine [62] is an exemplary tool for exploring how language works. Its algorithms analyse authentic texts of billions of words (text corpora) and instantly identify what is typical, rare, unusual, or frequently occurring in a language. It is also designed for text analysis or text mining applications. All the functions of the tool are performed through a browser in an online environment.

To perform the linguistic analysis, the audio recordings of each participant were converted into text form and merged into a single corpus. There are several speech-to-text transcription tools, but most of them can only work effectively with English. The entire study was conducted in Czech. This brought with it some limitations in the form of the challenge of finding a way to convert audio data efficiently and automatically into text. In total, fifteen tools were assessed. The Speech-to-Text tool under Google Engine, which is part of the Google Cloud platform, was used for the conversion. However, the converted recording always had to be manually edited. The resulting corpus was then subjected to analysis and contained a total of 17,006 tokens and 13,986 words in 1576 sentences from thirteen participants. The object of the linguistic analysis for the present study was to determine the frequencies of words (primarily verbs) and to compare them in a broader context (why they were said so many times). If an important phrase with a frequent word was found that was meaningful to the evaluators, the sentence was compared with the quantitative results from the previous annotation analysis. The individual sub-annotation results were formulated and interpreted into an overall conclusion. In this way, linguistic analysis served as another avenue for drawing qualitative conclusions.

2.5.3. Evaluation: A Retrospective Analysis

Retrospective analysis is a research method and a subjective form of information analysis. In a sense, it is the opposite of a prospective study. Based on their professional experience, researchers make assumptions in advance about what outcome can be expected. During an experiment they refute or confirm their assumptions. They record all findings in their notes. Thus, at the end of the experiment, both verbal protocols and observers' notes are collected. During the analysis, the researcher goes back in time and, based on all the data and notes, identifies product issues and suggests possible solutions. The basis for defining retrospective analysis can be found predominantly in the medical field. However, the method is also used by ecologists, historians, and criminologists and in other fields. However, the method has received much criticism from many scientists as it is a "quick and dirty" method of answering a question, and the downside is the sheer subjectivity that can lead to bias and increased systematic error in determining any real problems with a product. The advantage however, is that the information is gathered quickly and inexpensively.

In this study, a retrospective analysis was applied from the researcher's perspective without interaction with participants. However, it would also be possible to apply the retrospective think-aloud (RTA) method from the participants' perspective. The participants would think back through the resulting recording and think aloud about their decisions during the experiment.

To create as complete an analysis as possible, all the material generated during the experiment was used—observer notes, audio recordings, and recorded videos. The material sources collected were subjected to subjective analysis. Typically, questions such as what the participant had problems with or what could be improved based on the statements in the map were addressed in combination with the observer's recollection of specific situations.

3. Results

The objective of the study was to evaluate the usability of the *Atlas of the Moravian-Silesian Region* through the implementation of an experiment utilising the think-aloud method. Three methods of data analysis were identified and applied: annotation analysis,

linguistic analysis, and subjective analysis. The majority of the analysis was based on annotation analysis. The tasks assigned to the participants were grouped according to the topics covered in the atlas or the type of work performed with the atlas. Linguistic and retrospective analyses were also conducted, and the findings from these three forms of analysis were integrated into the overall conclusion.

3.1. Effectiveness

The primary metric utilised in the evaluation was the success rate score, which was calculated to determine the failure rate of a given aspect of the product. If a significant number of participants answered questions incorrectly, it indicated a need for improvement in that particular aspect of the product's usability. The success rate reflects the participants' ability to complete the task at hand and is calculated as follows:

$$Success\ Score = successfully\ completed\ questions\ /\ all\ questions \qquad (1)$$

The success score should fall between 0 and 1 (or 0 and 100%). The value indicates the relationship between a task's success or failure. A partial success is considered a failure. However, to gain a more nuanced understanding, a group of "partially successful" responses was also defined for a specific type of task. For instance, a task could be to determine the modes of transportation available to reach Ostrava airport. A partial success could be defined as identifying the primary modes of transportation (e.g., rail and bus). If the researcher found that the participant only made logical assumptions (e.g., car or plane) without fully navigating the map, this response was marked as a failure. The results of the success rate were presented in a table format, as illustrated in Figure 7.

	1	2	3	4a	4b	4c	5a	5b	5c	6a	6b	6c	7a	7b	8a	8b	9a	9b	s. part.
P01	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	0	0.72
P02	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	0.83
P03	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0.89
P04	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0.89
P05	1	0	0	1	0	1	0	1	1	1	0	1	1	1	1	0	0	0	0.56
P06	1	0	0	0	1	-	1	1	1	1	0	1	0	1	1	0	0	0	0.50
P07	0	0	1	1	0	1	1	1	1	1	0	1	1	-	-	0	1	0	0.56
P08	0	0	1	1	0	0	1	1	1	1	1	1	0	1	1	0	0	1	0.61
P09	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0.89
P10	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	1	0.78
P11	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0.89
P12	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0.89
P13	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0.89
	11	2	11	12	9	9	11	12	13	13	7	12	10	12	12	8	7	7	s. total
s. tasks	0.85	0.15	0.85	0.92	0.69	0.69	0.85	0.92	1.00	1.00	0.54	0.92	0.77	0.92	0.92	0.62	0.54	0.54	0.71

Figure 7. Task success rate and success score.

In addition to the success rate score, the accuracy of the participants' responses was evaluated using the error analysis method. The researcher conducted an assessment, either through annotation or linguistic analysis, to determine the level of difficulty experienced by the participants in completing the assigned tasks. A scale ranging from zero to two points was used to rate the difficulty of each task from each participant's perspective, where zero represented easy task completion, one point indicated partial success or difficulty in completion, and two points signified failure to complete the task. The table of error scores (refer

to Figure 8) was compared with the task completion and success scores (refer to Figure 7), and any correlation between the two was identified as an area requiring improvement.

	1	2	3	4a	4b	4c	5a	5b	5c	6a	6b	6c	7a	7b	8a	8b	9a	9b
P01	0	2	0	0	2	0	0	0	0	0	2	0	0	0	0	1	2	2
P02	0	0	0	0	1	1	1	0	0	0	2	0	0	0	0	0	2	2
P03	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	2
P04	0	2	0	0	0	0	2	0	0	0	0	0	1	0	0	1	1	1
P05	0	2	2	0	2	0	2	0	0	0	2	0	0	0	1	2	2	2
P06	1	2	2	2	1	-	1	0	0	0	2	0	2	0	1	2	2	2
P07	2	2	0	0	2	0	0	0	0	0	2	0	1	-	-	2	1	2
P08	2	2	0	0	2	2	0	0	0	0	0	1	2	0	0	2	2	0
P09	0	2	0	0	0	0	0	0	0	1	2	0	1	0	0	1	1	0
P10	0	2	0	0	1	0	0	2	0	0	1	1	1	0	0	2	2	1
P11	0	2	0	0	0	2	1	0	0	0	0	0	0	0	0	1	0	1
P12	0	2	0	0	0	1	0	0	0	0	0	2	0	0	0	1	1	0
P13	0	2	0	0	0	2	0	1	0	0	1	0	1	0	0	1	0	0
	5	23	4	2	11	8	7	3	0	1	14	4	11	0	2	17	16	15

Figure 8. Error rate.

The results of the experiment indicated that some of the tasks presented challenges for the participants, leading to incorrect responses. With the exception of two cases, participants struggled to accurately identify the number of maps in the atlas. Additionally, the overall performance was limited, with the highest success rate being 89% and the lowest success rate being 50–56%. This low success rate was observed in both senior participants. A correlation was observed between the success score for each task and the number of errors made. Task 2 (number of maps in the atlas) was the least successfully completed task, and with the other tasks the participants achieved a minimum success rate of 54%.

Based on the chosen scale, tasks 8b (development of the number of cars in the county, see Figure 3, #7), 9a, and 9b (questions related to public transport in the county, see Figure 3, #8) presented a high error rate. This was attributed to difficulties encountered by the participants in understanding and interpreting the infographic sections. The task of identifying the modes of transport to Ostrava airport also posed challenges as most participants were unable to accurately identify the various transportation lines. Despite the presence of a direct railway line to the airport, participants did not mention this option.

3.2. Efficiency

The use of the task execution time as a metric was also employed in the evaluation process. The ability to quickly and successfully retrieve information is a hallmark of a product with high usability. The design of the metric used to measure task completion time is straightforward, but it must be implemented effectively by the evaluator. The average time spent by participants on each task was calculated based on the structure of the test scenario, offering a comprehensive assessment of performance:

$$Average\ Task\ Time = \frac{time\ of\ the\ 1^{st} + 2^{nd} + \ldots + n^{th}\ user}{number\ of\ all\ participants} \tag{2}$$

The result of the calculation is the average time spent by the participants on the task. However, it is not possible to determine whether the result is good or bad without a

benchmark. Unlike other metrics, there are no industry standards for task-solving time, but a reference can be established by selecting a competent and knowledgeable participant. The selected participant, ideally with a background in cartography, should provide the time spent on the tasks as a reference. This metric of task time can also be used for future comparisons with updated versions of the product, with a decrease in the average time indicating improvement. Quick information delivery can greatly impact the participants' overall experience with the product and determine whether they return to the product or seek alternatives due to a negative experience (especially for websites).

The task (see Table 1, #10) of generating their own questions on any topic in the atlas took the longest as it required more critical thinking. The second-longest task was finding the number of maps in the atlas, with some participants counting each map page by page, which took a longer time. Task 9a (public transport information) also had a high average time of 70 s, despite being a simple map information task. The overall average task time is visualised in Figure 9.
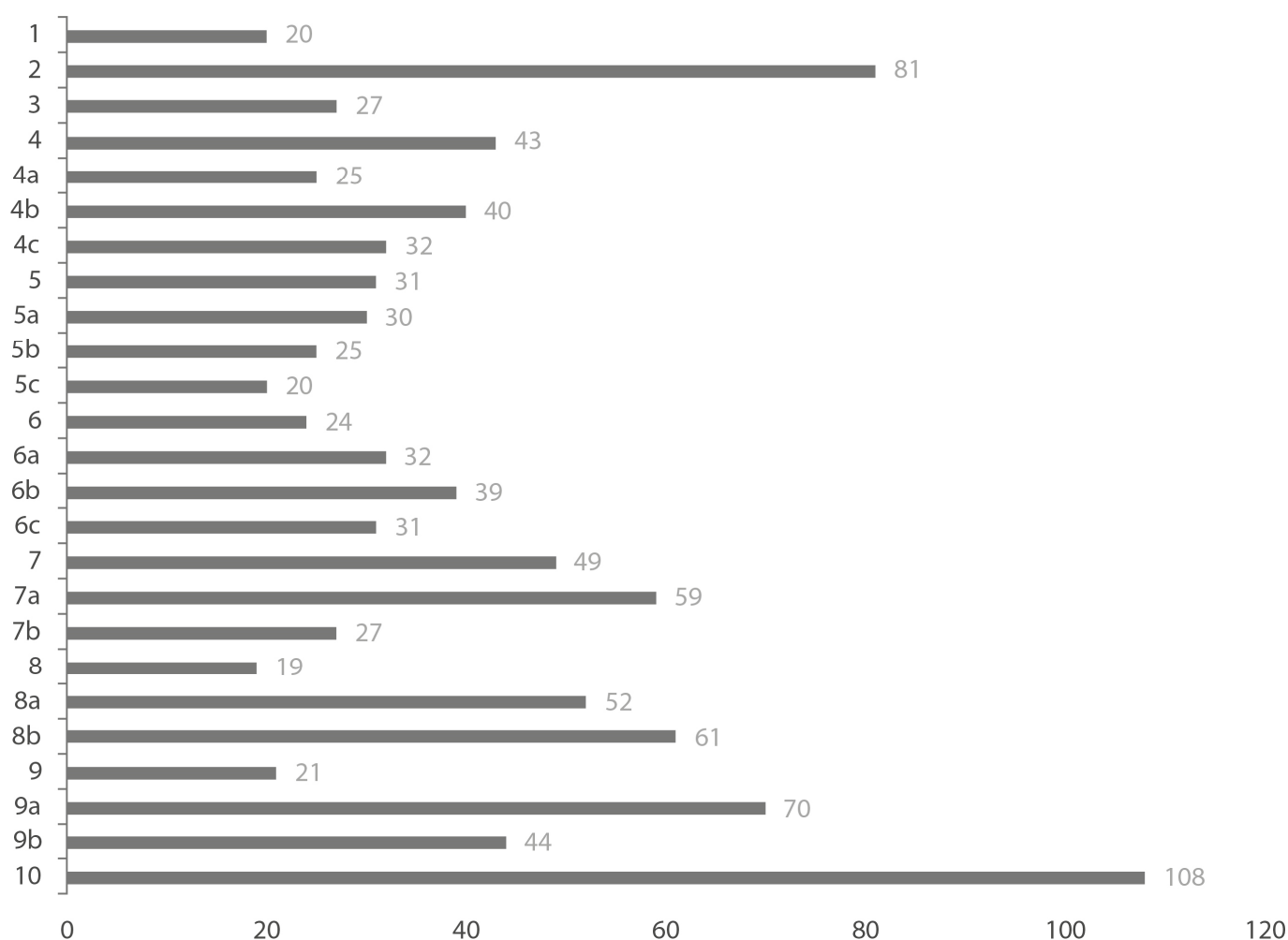

Figure 9. Average task completion time.

The next step involved comparing the average time for each type of annotation. Figure 10 shows the results of this comparison. On average, the longest type of annotation was searching for information in a map or legend (148 s). Participants often revisited some of the questions they had been asked to read out loud, taking an average of 139 s. The third longest part of the test was manipulating the atlas, taking an average of 114 s. On the other hand, participants spent the least amount of time reading the descriptions of the maps,

simply skimming over them without delving into the text, and this took an average of only 27 s for all tasks combined.
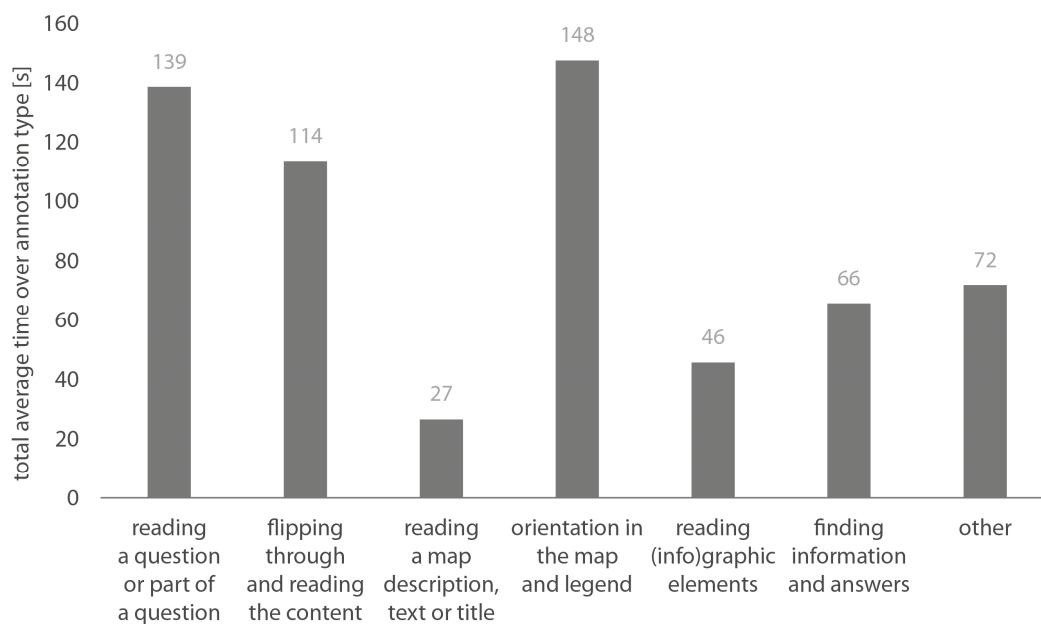
Figure 10. Average time of each annotation for all participants.

3.3. Satisfaction

To measure the efficiency or effectiveness of a product, there are various metrics available. However, the last characteristic of determining satisfaction from the participants' perspective has a different approach. Satisfaction is usually evaluated through a simple questionnaire at the end of a usability test, either in the form of a single ease question (SEQ) or system usability scale (SUS). In this study, satisfaction was determined through post-test interviews where participants verbally shared their opinions and experiences of the product. The researcher recorded these impressions in notes, which were later used for subjective retrospective analysis.

During the testing, participants reported difficulties in handling the atlas due to its size and weight. Finding specific topics in the atlas content (see Figure 3, #1), such as transportation, was time-consuming for some participants. They suggested that key topics should be made more prominent. One participant suggested adding an introductory page with statistical information about the atlas. When asked to identify seas (see Figure 3, #3), participants often referred to the map description where two names appeared: Baltic and Black. Unfortunately, almost a third of the participants answered incorrectly, identifying the Black Sea. The longest river was identified with the help of an infographic. For climatic regions (see Figure 3, #4), participants referred to either the table with climatic characteristics or the infographic showing the area of climatic regions, which they used as a legend. To determine the largest climatic area, participants compared different areas on the map or found the relevant infographic and read the information. When asked about the average annual temperature in the county (see Figure 3, #5), participants were confused by the values of annual and daily averages, leading to many incorrect answers. Half of the participants easily found the appropriate infographic with the value. The predominant type of traffic (see Figure 3, #6) was estimated from another infographic, not from the map. The biggest challenge was in calculating the rate of change in the number of cars (see Figure 3, #7). The correct answer was hidden in one of the infographics, but it was too complex for the participants to understand, leading to frustration with the cluttered and difficult-to-understand design. A similar issue was encountered with public transportation (see Figure 3, #8), where the complex infographic was described as overwhelming and irrelevant. Some participants preferred the information to be described in the text. When asked about

the types of public transportation to Ostrava airport, participants had difficulty finding the airport and the types of transportation due to the mixture of lines. The participants felt that the atlas had too many modern graphics and insufficient visual representations of the information, making it difficult to find what they were looking for. They described the photos on the left side of the map as unnecessary and suggested splitting the maps into two parts. The maps were small and not very prominent, so most participants looked for information around the map first. One participant had to bring the atlas close to their eyes to see the features. Some participants found the text hard to read because the letters were too small. There were also long pauses during some tasks. Some participants appreciated the clear division of topics. Figure 11 shows an example of the problems with the topic of public transport.
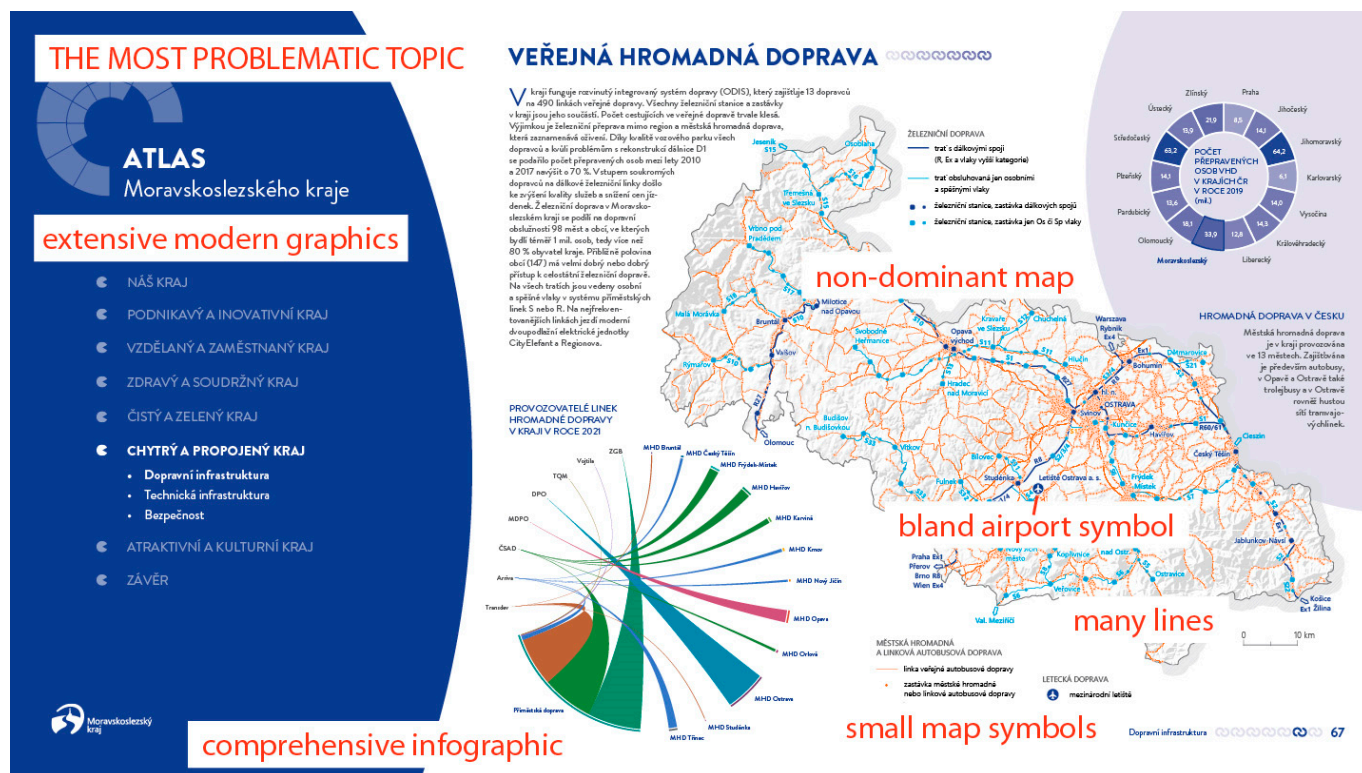

Figure 11. Issues found for the topic of public transport by participants.

Linguistic analysis can also be considered as part of the satisfaction analysis. In the case of the frequency of nouns, the words *region* (occurrence in the title of the atlas and the questions), *transport* (the most discussed topic), *map*, *part*, and *area* (part of the questions) appeared most frequently in the statements. Among the most frequent verbs, the forms *have*, *describe*, *find*, and *see* could be found. These frequent verbs were due to the questions asked and read aloud by the participants. The verb to see was said by participants when they were describing the area of occurrence of information or, on the other hand, when they could not find something, i.e., not see it. One of the initial questions asked about the number of maps in the atlas. The follow-up sentence suggested that participants tended to count the maps by hand immediately after reading the question. However, some of them subsequently realised that "*it could be stated somewhere*". Those who chose to count the maps counted the number of maps incorrectly. Another topic was temperature and precipitation, and participants were asked to find the value of the average annual temperature in the county. Participants tried to read this information from the map where daily temperatures were given. The correct answer was given in the corresponding infographic. Often, parts of the description such as annual, daily, minimum, maximum, and average were pronounced. The next topic was traffic volume. Participants found it difficult to find the right answer to

the question about the evolution of the number of cars in the region. The correct answer was in the map description and in the corresponding infographic. According to the statements, the infographic in question was described as weird, horrible, confusing, and not clear. The last topic defined was the most controversial and focused on public transport (see Figure 11). Participants were asked to identify the types of transport to Ostrava Airport. The information was searched for by the participants in the text around the main map and sentences such as "*I can't see it anywhere*" or "*I'm blind*" appeared. Subsequently, participants switched to searching for information on the map. Finding Ostrava airport was also a problem as participants looked for it in the city of Ostrava, not near the village of Mošnov (where the airport is located). Therefore, many answers gave the form of transport as public transport or on foot. Another problematic question was identifying the biggest operator of public transport lines in the region. This information was provided in a similar type of infographic. After the previous experience, participants were both disappointed and unpleasantly surprised by this finding. There were expressions such as " . . . there's that horrible graph again". To avoid searching the graph, some participants tried to find the information in the text they were reading aloud but without success.

4. Discussion

In this study, thirteen participants from different age groups were selected for usability testing. The purpose of selecting participants from different age groups was to ensure potential users of the product were in the study. Although the number of participants may seem low, it is considered a high number for the think-aloud method as, according to Nielsen [64], testing with five participants can reveal up to 80% of the product's usability problems. The optimal number of participants for the think-aloud method, according to Fiderici et al. [65], is six, which revealed up to 84% of usability problems in their research. Nielsen and Norman's group [64,66] found no noticeable increase in findings with more participants in their 83 usability studies [64]. The participants in this study included product experts, such as cartographers, to obtain their expert opinions. However, verbalising their knowledge and attitudes toward the product can pose a challenge as experts may not be able to explain how they arrived at the correct answer. To overcome this, the experts were instructed to provide natural and direct statements, and any unintelligible parts could be clarified later. Verbalisation skills vary among individuals, with extroverts typically speaking more easily. Selecting participants with good verbalisation skills is ideal, but it is difficult to determine in advance. Children may find it challenging to think aloud, but this can be addressed through a pilot experiment.

In the study, several tools for processing the recorded data were assessed, including annotation tools, speech-to-text tools, and linguistic analysis tools. The most commonly used annotation tool is ELAN, but other tools such as Audino, Prodigy, Label Studio, Audio-annotator, and Audio-labeler were also tested. Ultimately, ELAN was selected as the final annotation software. The quality of the recording and the sensitivity of the microphone greatly impact the quality of transcribing audio recordings to text. Choosing a recording device that filters out noise and gently registers the participant's voice is essential. The study found the Trust GXT 232 filter microphone to be the best quality microphone (of those tested).

During the research, three directions were defined for processing the recorded verbal protocols. The method only applies the approach of separating the audio recording into annotations. In this study, the method was extended to include two other approaches, linguistic analysis and retrospective subjective analysis. It was verified that these approaches can provide additional information about the usability of the product. In several studies, only the method of subjective analysis is applied. Thus, it can be concluded that only established tracks were used, but they had not been named up to now; therefore, a retrospective subjective analysis of verbal protocols was defined in this study. Thus, in contrast to many studies, this study offers multiple workable solutions. Applying the simplest variant of analysis by simply collecting opinions is performed because of the difficulty

in annotating the records. Therefore, the ordinary researcher opts for simple subjective analysis, the application of which, as the only option, degrades the think-aloud method. Streamlining the annotation principle thus offers the potential for future improvements and simplification of this method. Other studies, such as Knura et al. [37] or Prokop et al. [67], have compared the think-aloud method with the eye-tracking method. In their studies, it was found that the think-aloud method is greatly beneficial and can fully replace the eye-tracking method. Nevertheless, this is still a less explored area, and comparing the think-aloud method with other usability testing methods still holds potential for the future.

5. Conclusions

The aim of the study was to use the think-aloud method to evaluate a regional atlas. In the first stage, the researchers familiarised themselves with the principles and history of the method through the literature review. Then, they tested the tools used to process the recorded data. Three approaches were proposed for analysing the audio recordings: annotation, linguistic analysis, and retrospective analysis. Thirteen participants with varying ages and expertise levels participated in the experiment. The atlas was evaluated based on the annotation analysis, resulting in the calculation of usability metrics such as success score, error analysis, and average task completion time. Additionally, linguistic analysis and retrospective analysis were conducted for supplementary data.

The study found that only a small portion of participants were able to correctly find information about the number of maps in the atlas. The most error-prone topic was public transport, with many participants having difficulty finding information due to the large number of lines on the map and unclear infographics. Overlapping less relevant topics made it difficult to search for key information, and some participants had trouble identifying the climatic regions and temperatures shown in the atlas. The small maps, complex information, and excessive graphics were also criticised, leading to the conclusion that the atlas was unsuitable for senior users. The overall success score was 0.71, and recommendations were made to improve the atlas, including adding a page with statistical data, correcting the topic of public transport, changing infographics, simplifying the visual appearance, and enlarging maps and descriptions. These findings were shared with the authors of the atlas.

In conclusion, the think-aloud method was shown to be a valuable tool for evaluating cartographic products, specifically atlases and the maps within them. The method is dependent on the researcher's ability to create an optimal test scenario and accurately interpret collected data. However, it is important to carefully select representative participants to ensure that the results are relevant. As was proven by the research, the think-aloud method could be among the first choices for evaluating the usability of cartographic products in the testing phase. Its ability to provide valuable insights into participants' behaviour and their opinions on cartographic products makes it an important method to be used in the evaluation of these products.

Author Contributions: Conceptualization, Tomas Vanicek and Stanislav Popelka; investigation, Tomas Vanicek; methodology, Tomas Vanicek; supervision, Stanislav Popelka; visualization, Tomas Vanicek; writing—original draft, Tomas Vanicek; writing—review and editing, Stanislav Popelka. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: Recorded audio data and their transcriptions (in Czech) and files from ELAN software are available at Mendeley Data repository—Vanicek, Tomas; Popelka, Stanislav (2023), "Think-Aloud Usability Assessment of a Regional Atlas", Mendeley Data, V1, doi: 10.17632/y6f9ykp4dt.1.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1.   *I.O.f. ISO 9241-11*; 2018—Ergonomics of Human-System Interaction—Part 11: Usability: Definitions and Concepts. HCI International: Los Angeles, CA, USA, 2018.
2.   Dumas, J.S.; Redish, J. *A Practical Guide to Usability Testing*; Intellect Books: Bristol, UK, 1999.
3.   Nielsen, J. *Usability Engineering*; Morgan Kaufmann: Burlington, MA, USA, 1994.
4.   Chen, W.P.; Lin, T.; Chen, L.; Yuan, P.S. Automated comprehensive evaluation approach for user interface satisfaction based on concurrent think-aloud method. *Univers. Access Inf. Soc.* **2018**, *17*, 635–647. [CrossRef]
5.   Heller, D. Thinking aloud as a research method. *Ceskoslovenska Psychol.* **2005**, *49*, 554–562.
6.   Duarte, N.P.; Korelo, J.C. The use of the think aloud verbal protocol for tracking processes in research on the consumer decision-making. *Rev. Bras. De Mark.* **2017**, *16*, 317–333. [CrossRef]
7.   Habibi, M.R.M.; Khajouei, R.; Eslami, S.; Jangi, M.; Ghalibaf, A.K.; Zangouei, S. Usability testing of bed information management system: A think-aloud method. *J. Adv. Pharm. Technol. Res.* **2018**, *9*, 153–157. [CrossRef]
8.   Chang, C.C.; Johnson, T. Integrating heuristics and think-aloud approach to evaluate the usability of game-based learning material. *J. Comput. Educ.* **2021**, *8*, 137–157. [CrossRef]
9.   Smith, J.G.; Alamiri, N.S.; Biegger, G.; Frederick, C.; Halbert, J.P.; Ingersoll, K.S. Think-Aloud Testing of a Novel Safer Drinking App for College Students During COVID-19: Usability Study. *JMIR Form. Res.* **2022**, *6*, e32716. [CrossRef]
10. Ishaq, K.; Rosdi, F.; Zin, N.A.M.; Abid, A. heuristics and think-aloud method for evaluating the usability of game-based language learning. *Int. J. Adv. Comput. Sci. Appl.* **2021**, *12*, 311–324. [CrossRef]
11. Fan, M.M.; Li, Y.; Truong, K.N. Automatic Detection of Usability Problem Encounters in Think-Aloud Sessions. *Acm Trans. Interact. Intell. Syst.* **2020**, *10*, 1–24. [CrossRef]
12. Jaspers, M.W.; Steen, T.; Van Den Bos, C.; Geenen, M. The think aloud method: A guide to user interface design. *Int. J. Med. Inform.* **2004**, *73*, 781–795. [CrossRef]
13. Charters, E. The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Educ. J.* **2003**, *12*. [CrossRef]
14. Ericsson, K.A.; Simon, H.A. Verbal reports on thinking. In *Introspection inSecond Language Research*; Faerch, C., Kasper, G., Eds.; Multilingual Matters: Bristol, UK, 1987.
15. Van Someren, M.; Barnard, Y.F.; Sandberg, J. The think aloud method: A practical approach to modelling cognitive. *Lond. Acad.* **1994**, *11*, 29–41.
16. Bláha, D. Vliv Kognitivní Zátěže Na Použitelnost Uživatelských Rozhraní Vybraných Internetových Bankovnictví. Master Thesis, Masarykova univerzita, Filozofická fakulta, Brno, 2015.
17. Ericsson, K.A.; Simon, H.A. Verbal reports as data. *Psychol. Rev.* **1980**, *87*, 215–251. [CrossRef]
18. Kuusela, H.; Paul, P. A comparison of concurrent and retrospective verbal protocol analysis. *Am. J. Psychol.* **2000**, *113*, 387–404. [CrossRef] [PubMed]
19. Guan, Z.; Lee, S.; Cuddihy, E.; Ramey, J. The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In Proceedings of the Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 22–27 April 2006; pp. 1253–1262. [CrossRef]
20. Alnashri, A.; Alhadreti, O.; Mayhew, P.J. The Influence of Participant Personality in Usability Tests. *Int. J. Hum. Comput. Interact.* **2016**, *7*.
21. Ericsson, K.A.; Simon, H.A. *Protocol Analysis: Verbal Reports as Data*; The MIT Press: Cambridge, MA, USA, 1993.
22. Boren, T.; Ramey, J. Thinking aloud: Reconciling theory and practice. *IEEE Trans. Prof. Commun.* **2000**, *43*, 261–278. [CrossRef]
23. May, J. YouTube Gamers and Think-Aloud Protocols: Introducing Usability Testing. *IEEE Trans. Prof. Commun.* **2019**, *62*, 94–103. [CrossRef]
24. Fan, M.M.; Lin, J.L.; Chung, C.; Truong, K.N. Concurrent Think-Aloud Verbalizations and Usability Problems. *Acm Trans. Comput. -Hum. Interact.* **2019**, *26*, 1–35. [CrossRef]
25. Fan, M.M.; Shi, S.; Truong, K.N. Practices and Challenges of Using Think-Aloud Protocols in Industry: An International Survey. *J. Usability Stud.* **2020**, *15*, 85–102.
26. Fan, M.M.; Tibdewal, V.; Zhao, Q.W.; Cao, L.Z.; Peng, C.; Shu, R.X.; Shan, Y.J. Older Adults' Concurrent and Retrospective Think-Aloud Verbalizations for Identifying User Experience Problems of VR Games. *Interact. Comput.* **2023**, iwac039. [CrossRef]
27. Alhadreti, O. Comparing Two Methods of Usability Testing in Saudi Arabia: Concurrent Think-Aloud vs. Co-Discovery. *Int. J. Hum.-Comput. Interact.* **2021**, *37*, 118–130. [CrossRef]
28. Carter-Roberts, H.; Antbring, R.; Angioi, M.; Pugh, G. Usability testing of an e-learning resource designed to improve medical students' physical activity prescription skills: A qualitative think-aloud study. *BMJ Open* **2021**, *11*, e042983. [CrossRef] [PubMed]
29. Crawford, N.D.; Josma, D.; Harrington, K.R.V.; Morris, J.; Quamina, A.; Birkett, M.; Phillips, G. Using the Think-Aloud Method to Assess the Feasibility and Acceptability of Network Canvas Among Black Men Who Have Sex With Men and Transgender Persons: Qualitative Analysis. *JMIR Form. Res.* **2021**, *5*, e30237. [CrossRef]

30. Doi, T. Usability Textual Data Analysis: A Formulaic Coding Think-Aloud Protocol Method for Usability Evaluation. *Appl. Sci.-Basel* **2021**, *11*, 7047. [CrossRef]

31. Giang, W.C.W.; Bland, E.; Chen, J.; Colon-Morales, C.M.; Alvarado, M.M. User Interactions With Health Insurance Decision Aids: User Study With Retrospective Think-Aloud Interviews. *JMIR Hum. Factors* **2021**, *8*, e27628. [CrossRef]

32. Hanghoj, S.; Boisen, K.A.; Hjerming, M.; Elsbernd, A.; Pappot, H. Usability of a Mobile Phone App Aimed at Adolescents and Young Adults During and After Cancer Treatment: Qualitative Study. *JMIR Cancer* **2020**, *6*, e15008. [CrossRef]

33. Hendradewa, A.P.; Yassierli. A Comparison of Usability Testing Methods in Smartphone Evaluation. *Ind. Eng. Manag. Syst.* **2019**, *18*, 154–162. [CrossRef]

34. Irlitti, A.; Hoang, T.; Vetere, F.; Acm. Surrogate-Aloud: A Human Surrogate Method for Remote Usability Evaluation and Ideation in Virtual Reality. In Proceedings of the CHI Conference on Human Factors in Computing Systems, Electr Network, Hamburg, Germany, 8–13 May 2021.

35. Jabbour, J.; Dhillon, H.M.; Shepherd, H.L.; Sundaresan, P.; Milross, C.; Clark, J.R. A web-based comprehensive head and neck cancer patient education and support needs program: Usability testing. *Health Inform. J.* **2022**, *28*. [CrossRef]

36. Khajouei, R.; Farahani, F. A combination of two methods for evaluating the usability of a hospital information system. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 14604582221087128. [CrossRef] [PubMed]

37. Knura, M.; Schiewe, J. Map Evaluation under COVID-19 restrictions: A new visual approach based on think aloud interviews. *Proc. ICA* **2021**, *4*, 60. [CrossRef]

38. Kusumawati, R.E.; Muslim, E.; Nugroho, D. Usability Testing on Touchscreen Based Electronic Kiosk Machine in Convenience Store. In Proceedings of the 16th International Conference on Quality in Research (QiR)/International Symposium on Advances in Mechanical Engineering (ISAME), Padang, Indonesia, 21–24 July 2019.

39. Nielsen, L.; Salminen, J.; Jung, S.G.; Jansen, B.J. Think-Aloud Surveys A Method for Eliciting Enhanced Insights During User Studies. In Proceedings of the 18th IFIP TC 13 International Conference on Human-Computer Interaction (INTERACT), Bari, Italy, 30 August–3 September 2021; pp. 504–508.

40. Sin, J.; Woodham, L.A.; Henderson, C.; Williams, E.; Hernandez, A.S.; Gillard, S. Usability evaluation of an eHealth intervention for family carers of individuals affected by psychosis: A mixed-method study. *Digit. Health* **2019**, *5*. [CrossRef]

41. Soure, E.J.; Kuang, E.; Fan, M.M.; Zhao, J. CoUX: Collaborative Visual Analysis of Think-Aloud Usability Test Videos for Digital Interfaces. *IEEE Trans. Vis. Comput. Graph.* **2022**, *28*, 643–653. [CrossRef]

42. Steeb, T.; Brutting, J.; Reinhardt, L.; Hoffmann, J.; Weiler, N.; Heppt, M.V.; Erdmann, M.; Doppler, A.; Weber, C.; Schadendorf, D.; et al. One Website to Gather them All: Usability Testing of the New German SKin Cancer INFOrmation (SKINFO) Website-A Mixed-methods Approach. *J. Cancer Education.* **2022**, 1–7. [CrossRef]

43. Taylor, S.; Allsop, M.J.; Bennett, M.I.; Bewick, B.M. Usability testing of an electronic pain monitoring system for palliative cancer patients: A think-aloud study. *Health Inform. J.* **2019**, *25*, 1133–1147. [CrossRef]

44. Sucha, L.Z.; Bartosova, E.; Novotny, R.; Svitakova, J.B.; Stefek, T.; Vichova, E. Stimulators and barriers towards social innovations in public libraries: Qualitative research study. *Libr. Inf. Sci. Res.* **2021**, *43*, 101068. [CrossRef]

45. Špriňarová, K.; Juřík, V.; Šašinka, Č.; Herman, L.; Štěrba, Z.; Stachoň, Z.; Chmelík, J.; Kozlíková, B. Human-Computer Interaction in Real-3D and Pseudo-3D Cartographic Visualization: A Comparative Study. In *Cartography-Maps Connecting the World, Proceedings of the 27th International Cartographic Conference 2015-ICC2015, Rio de Janeiro, Brazil, 23–28 August 2015*; Robbi Sluter, C., Madureira Cruz, C.B., Leal de Menezes, P.M., Eds.; Springer International Publishing: Midtown Manhattan, NY, USA, 2015; pp. 59–73.

46. Šašinka, Č.; Stachoň, Z.; Sedlák, M.; Chmelík, J.; Herman, L.; Kubíček, P.; Šašinková, A.; Doležal, M.; Tejkl, H.; Urbánek, T.; et al. Collaborative Immersive Virtual Environments for Education in Geography. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 3. [CrossRef]

47. Popelka, S.; Vondrakova, A.; Hujnakova, P. Eye-tracking Evaluation of Weather Web Maps. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 256. [CrossRef]

48. Nivala, A.-M.; Brewster, S.; Sarjakoski, T.L. Usability Evaluation of Web Mapping Sites. *Cartogr. J.* **2008**, *45*, 129–138. [CrossRef]

49. Ooms, K.; De Maeyer, P.; Fack, V. Listen to the Map User: Cognition, Memory, and Expertise. *Cartogr. J.* **2015**, *52*, 3–19. [CrossRef]

50. Kulhavy, R.W.; Pridemore, D.R.; Stock, W.A. Cartographic Experience And Thinking Aloud About Thematic Maps. *Cartogr.: Int. J. Geogr. Inf. Geovisualization* **1992**, *29*, 1–9. [CrossRef]

51. Gołębiowska, I. Legend Layouts for Thematic Maps: A Case Study Integrating Usability Metrics with the Thinking Aloud Method. *Cartogr. J.* **2015**, *52*, 28–40. [CrossRef]

52. Viaene, P.; Vanclooster, A.; Ooms, K.; De Maeyer, P. Thinking Aloud in Search of landmark characteristics in an Indoor Environment. In Proceedings of the 2014 Ubiquitous Positioning Indoor Navigation and Location Based Service (UPINLBS), Corpus Christi, TX, USA, 20–21 November 2014; pp. 103–110.

53. Kettunen, P.; Irvankoski, K.; Krause, C.M.; Sarjakoski, L.T. Landmarks in nature to support wayfinding: The effects of seasons and experimental methods. *Cogn. Process.* **2013**, *14*, 245–253. [CrossRef]

54. Quaye-Ballard, J. Usability testing: Using "think aloud" method in testing cartographic product. *J. Sci. Technol.* **2007**, *27*, 141–149. [CrossRef]

55. Burian, J. *Atlas Moravskoslezského Kraje: Lidé, Podnikání, Prostředí*; Moravskoslezský kraj: Ostrava, Czech Republic, 2021.

56. Lloyd, R.; Patton, D.; Cammack, R. Basic-Level Geographic Categories. *Prof. Geogr.* **1996**, *48*, 181–194. [CrossRef]

57. Burnett, G.; Ditsikas, D. Personality as A-Criterion for Selecting Usability Testing Participants. In Proceedings of the IEEE 4th International Conference on Information and Communications Technologie, Cairo, Egypt, 10–12 December 2006. [CrossRef]

58. Meier, M.; Mason, C.; Putze, F.; Schultz, T. Comparative Analysis of Think-Aloud Methods for Everyday Activities in the Context of Cognitive Robotics. In Proceedings of the 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15–19 September 2019; pp. 559–563.

59. Wittenburg, P.; Brugman, H.; Russel, A.; Klassmann, A.; Sloetjes, H. ELAN: A Professional Framework for Multimodality Research. In Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy, 24–26 May 2006.

60. Buhajová, L. Ověření Využitelnosti Softwaru ATLAS. Ti Pro Literární Vědu. Diploma Thesis, Palacký University, Olomouc, Czech Republic, 2010.

61. Cvrček, V.; Čech, R.; Kubát, M. QuitaUp-Nástroj Pro Kvantitativní Stylometrickou Analýzu. Available online: https://korpus.cz/quitaup (accessed on 1 October 2022).

62. Kilgarriff, A.; Baisa, V.; Bušta, J.; Jakubíček, M.; Kovář, V.; Michelfeit, J.; Rychlý, P.; Suchomel, V. The Sketch Engine: Ten years on. *Lexicography* **2014**, *1*, 7–36. [CrossRef]

63. Beitlova, M.; Popelka, S.; Voženílek, V.; Fačevicová, K.; Janečková, B.A.; Matlach, V. The Importance of School World Atlases According to Czech Geography Teachers. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 504. [CrossRef]

64. Nielsen, J. How Many Test Users in a Usability Study? *Nielsen Norman Group* **2012**, 1. Available online: https://www.nngroup.com/articles/how-many-test-users/ (accessed on 1 October 2022).

65. Federici, S.; Borsci, S.; Stamerra, G. Web usability evaluation with screen reader users: Implementation of the Partial Concurrent Thinking Aloud technique. *Cogn. Process.* **2010**, *11*, 263–272. [CrossRef]

66. Nielsen, J. Usability 101: Introduction to Usability. *Nielsen Norman Group* **2012**, *1*, 3246–3252.

67. Prokop, M.; Pilař, L.; Tichá, I. Impact of Think-Aloud on Eye-Tracking: A Comparison of Concurrent and Retrospective Think-Aloud for Research on Decision-Making in the Game Environment. *Sensors* **2020**, *20*, 2750. [CrossRef] [PubMed]