International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.852

(07/2022)

SERIES P: TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Methods for objective and subjective assessment of speech and video quality

Subjective quality evaluation of text-based chatbots

Recommendation ITU-T P.852

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	P.10–P.19
Voice terminal characteristics	P.30–P.39
Reference systems	P.40–P.49
Objective measuring apparatus	P.50–P.59
Objective electro-acoustical measurements	P.60–P.69
Measurements related to speech loudness	P.70–P.79
Methods for objective and subjective assessment of speech quality	P.80–P.89
Voice terminal characteristics	P.300–P.399
Objective measuring apparatus	P.500–P.599
Measurements related to speech loudness	P.700–P.709
Methods for objective and subjective assessment of speech and video quality	**P.800–P.899**
Audiovisual quality in multimedia services	P.900–P.999
Transmission performance and QoS aspects of IP end-points	P.1000–P.1099
Communications involving vehicles	P.1100–P.1199
Models and tools for quality assessment of streamed media	P.1200–P.1299
Telemeeting assessment	P.1300–P.1399
Statistical analysis, evaluation and reporting guidelines of quality measurements	P.1400–P.1499
Methods for objective and subjective assessment of quality of services other than speech and video	P.1500–P.1599

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T P.852

Subjective quality evaluation of text-based chatbots

Summary

Recommendation ITU-T P.852 describes methods and procedures for conducting subjective evaluation experiments for services that are based on text-based chatbots. Such chatbots enable a natural language-based dialogic interaction via text, and are used to offer customer care self-services, service selling, etc. The set-up and running of appropriate interaction experiments is described, and questionnaires for quantifying the relevant quality dimensions perceived by the user are given.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.852	2022-07-29	12	11.1002/1000/15078

Keywords

Chatbot, dialogue management, interaction parameter, natural language generation, natural language understanding, subjective evaluation.

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents/software copyrights, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the appropriate ITU-T databases available via the ITU-T website at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T P.852

Subjective quality evaluation of text-based chatbots

1 Scope

This Recommendation describes subjective evaluation methods providing information about the quality of services relying on text-based chatbots, as experienced by the users of such services. Text-based chatbots addressed by the Recommendation enable a text-based natural language interaction with a human user via a text interface on a turn-by-turn basis. They possess natural language understanding (NLU), dialogue management and natural language generation (NLG) capabilities. They may provide access to customer care or allow different types of transactions to be performed.

The evaluation methods described here address different aspects of quality from a user's point of view, taking the chatbot as a black box. Important quality aspects are the overall impression, the information provided by the system, the communication with the system, the system behaviour, the user's impression of the system and the overall acceptability. The described methods are based on laboratory experiments in which participants interact with the chatbot in order to perform a pre-defined, realistic task. The participant's opinion on perceptive quality dimensions are solicited with the help of questionnaires, and examples of such questionnaires are provided. The Recommendation describes the set-up and running of interaction experiments, relevant quality dimensions perceived by the user and methodologies that will provide information about these quality dimensions. Further guidance on subjective evaluation methods related to speech-based (in contrast to text-based) chatbots are given in [ITU-T P.851].

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T P.800] Recommendation ITU-T P.800 (1996), *Methods for subjective determination of transmission quality*.

[ITU-T P.851] Recommendation ITU-T P.851 (2003), *Subjective quality evaluation of telephone services based on spoken dialogue systems*.

[ITU-T P.910] Recommendation ITU-T P.910 (2022), *Subjective video quality assessment methods for multimedia applications*.

[ITU-T P.911] Recommendation ITU-T P.911 (1998), *Subjective audiovisual quality assessment methods for multimedia applications*.

[ITU-T P.1401] Recommendation ITU-T P.1401 (2020), *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*.

3 Definitions

None.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ACR Absolute Category Rating

ANOVA Analysis of Variance

NLG Natural Language Generation

NLU Natural Language Understanding

WoZ Wizard of Oz

5 Conventions

None.

6 Chatbots

Chatbots are becoming more and more common as a way to interact between humans and computers. Instead of traditional chatbots that focused on maintaining a language-based informal interaction for its own sake (sometimes called chitchat bots) or non-task driven but still serious applications (e.g., Weizenbaum's ELIZA), most commercial chatbots now focus on question answering or on providing information through a multi-turn natural language interaction between user and computer. They are used by telecommunication service providers as tools for customer care self-services, pre-sorting in rather complex support requests or for selling new services. Although some such services allow for a spoken interaction and include speech recognition and speech synthesis, a majority rely on written text only.

In contrast to spoken interaction, written language interaction partly follows different rules, because the interaction can be asynchronous and the history of the interaction can more easily be reviewed by the user, and because recognition errors are bypassed (but may be replaced by spelling errors). For a spoken interaction, [ITU-T P.851] provides guidelines regarding quality aspects and influencing factors, as well as subjective evaluation methods, including experimental set-ups, questionnaires and result analysis. Whereas some aspects of text- and speech-based interactive systems are similar, many others require the specification of separate aspects to be considered in the evaluation.

This Recommendation provides guidance on how to evaluate text-based chatbots from a user-centred point of view. In addition, there is a need to specify parametric descriptions of text-based interactions, in a way similar to the parameters specified in [b-ITU-T P-Suppl.24]. These parametric descriptions are for further study.

6.1 Tasks and components of chatbots

A chatbot exchanges information between a user and a computer system by means of written language. It serves as a specific type of human-computer interface. The exchange of information normally serves a task the user would like to carry out with the help of the system; such tasks include answering a question or performing a transaction.

Written language is a common medium for human interaction. Thus, its use does not require any particular knowledge from the user, apart from the fact that the user needs to be sufficiently familiar with the language used (in case of non-native users), and that the user needs to read and type written messages without problems (such as illiteracy, or viewing or motor difficulties). In addition, as the interaction devices typically include a computer or on-screen keyboard and a pointing device (e.g., mouse, touchscreen), the user needs to be familiar with the operation of such devices.

In order to make use of written language as an interaction medium, a computer must have the ability to interpret written messages, i.e., to extract the semantics from the textual messages. In contrast to a

graphical user interface, where the semantics of controls is allocated in a fixed way, written language is flexible in its use, and may thus transport arbitrary meanings. Thus, chatbots first need a component to analyse the meaning of what was written by the user. This component is commonly called an NLU component. Depending on the task and language to be expected, an NLU component might extract keywords from a given text or perform more or less sophisticated pattern matching, using machine-learning techniques. Such techniques may consider the context of words in an utterance (e.g., to understand negations or relations), and may be able to extract several semantic concepts from one user utterance. The semantic concepts may be provided in terms of specific intents and attribute-value pairs, more abstract dialogue acts and potentially more complex knowledge representations (such as knowledge graphs).

On the basis of the extracted meaning, the chatbot decides on the next interaction step. This decision is performed by a so-called dialogue manager, which can be implemented either with the help of rules (following, e.g., a "dialogue grammar"), or with statistical transitions learned from training data, i.e., text or dialogue corpora. While a rule-based dialogue manager might be rather easy to implement, and has the advantage that it can easily be verified with respect to misleading dialogues, a corpus-based dialogue manager may be more flexible and may cover a wider range of interaction phenomena.

The next interaction step then needs to be put out to the user, again using written language. For this, an NLG module is necessary. It can be based on pre-defined templates that are filled and concatenated with textual information or again on machine-learning techniques.

In addition to the components already mentioned that directly serve the processing of written language, a chatbot is commonly equipped with a graphical output device, as well as an input device allowing for text input. The latter may be a separate keyboard that is physical or software implemented on a touchscreen, thus combining input and output within one physical device. A smartphone or tablet may be used for this purpose.

6.2 Interactions with chatbots

A task may require one or several interaction steps to be performed; these steps are commonly called "turns", and each may include one or several written utterances. Turns from the system normally alternate with turns from the user, but in some cases the system may take two successive turns (e.g., if the user does not respond within a given time period), or the user may take two subsequent turns (if the system does not show any perceivable reaction).

Turns may consist of complete sentences, phrases or single words. Commonly, chatbots try to provide full sentences as output, whereas they accept also text fragments or text commands as input. In addition, chatbots may provide suggestions to the user about which type of input they expect at a certain stage of the dialogue; in that case, the textual output of the system might be augmented by radio buttons, selection boxes, etc. While these input options might be helpful to guide the user and to lead the dialogue to an expected goal, they are rather similar to options in a graphical user interface, and may require other evaluation techniques that lie outside the scope of this Recommendation.

6.3 Quality aspects and influencing factors

Chatbots mainly serve the result or solution of a given task. Thus, from a functional point of view, effectiveness and efficiency are important quality aspects. They are both related to performance in achieving the task goal the chatbot-based service has been built for. Effectiveness is an absolute index that describes the accuracy and completeness with which specified users can achieve specified goals in particular environments. Measures of effectiveness that are reported in literature are, for example, success in the task. Efficiency, on the other hand, is a relative measure of goal achievement in relation to the resources used, i.e., the resources expended in relation to the accuracy and completeness of goals achieved. Commonly used metrics are, for example, the interaction duration, the number of turns written by the system or by the user or the cognitive demand put on the user to perform the interaction.

Both effectiveness and efficiency are components of the usability of a chatbot, see the taxonomy of quality aspects of spoken dialogue services in [b-Möller] as well as [ITU-T P.851]. Usability, however, is generally conceived in a much broader sense, and describes the capability of the service to be understood, learned and used by specified users under specified conditions. It indicates the suitability of the service to fulfil user requirements, includes effectiveness and efficiency of the system and results in user satisfaction. User satisfaction is an indicator of the service's perceived usefulness and usability for the intended user group. It includes whether a user gets the desired information, is comfortable with the service and gets the information within an acceptable elapsed time [b-Möller].

The quality of the written interaction largely determines its functional quality. This includes the input quality (i.e., whether the user feels understood by the system), the output quality, the cooperativity of system behaviour and the symmetry of the written dialogic interaction. The system should also behave cooperatively with the user, in the sense that it supports the user in reaching the task goal. Cooperativity can be conceived in the sense of applying the cooperative principles of conversational communication, as described in [b-Grice]. It includes the aspects of informativeness, truth and evidence, relevance, manner, background knowledge and meta-communication handling (i.e., confirmation, clarification, repair and recovery from communication errors); see [b-Bernsen].

The mentioned quality aspects result in a (more or less) efficient interaction, and in an efficient solution of the task to be carried out. Interaction efficiency is related to the speed or pace of the interaction, to dialogue conciseness, and to dialogue smoothness. Task efficiency, on the other hand, is linked to task success and task ease. Two additional quality aspects are important: the "personality" of the chatbot (politeness, friendliness, naturalness of behaviour) and the effort required from the human user for the interaction (ease of communication, stress or fluster, etc.). These aspects can been subsumed under the term comfort.

Communication efficiency, task efficiency and comfort all contribute to service usability, for which user satisfaction can be seen as an indicator. Service efficiency, on the other hand, is influenced by both task efficiency and contextual factors. It is important for the adequacy of the service (for fulfilling the desired task), and for the added value attributed to the service (e.g., in comparison to similar methods for obtaining the same information, like a web interface or a news ticker). Usability, service efficiency and economic benefit result in utility of the service, and finally in its acceptability.

The chatbot exercises an influence on the mentioned quality aspects in two ways, regarding: a) the task a user is able to carry out with its help (task factors such as how well the chatbot captures the task it has been designed for, the complexity of the task); and b) the factors that influence the dialogic interaction (agent factors). In addition, the usage environment may be an influencing factor (physical environment such as light or moving conditions, social environment such as other persons being present during the interaction). Finally, the user with their characteristics (aims, experience, expectations, typing behaviour, linguistic background, etc.) influence perceived quality. The quality of the service finally results from user perception, in relation to expectations of or desires from the service. It is highly dependent on the situation in which perception and judgement take place. This fact has to be taken into account when carrying out subjective quality evaluation experiments, namely by creating a sufficiently natural test situation and a realistic test user motivation.

6.4 Subjective evaluation methods

Chatbots can be assessed on a component level (e.g., with respect to the NLU or the NLG component), or with respect to the overall (integrated) system. Analytical assessment on the component level is a valuable source of information in describing how the individual parts of the system fulfil their task. It may, however, sometimes miss the relevant contributors to the overall quality of the service, as perceived by the user. For this reason, subjective experiments with real or test users interacting with the chatbot are indispensable when the quality of a chatbot-based service is to be determined.

In order to evaluate different aspects of the quality of a chatbot-based service, subjective experiments with human users have to be carried out. These experiments serve two main purposes as follows.

1. During the interaction, instrumentally measurable system parameters are collected, and the messages of the system and of the user are logged. The log files are submitted to an expert evaluation, the outcome of which is a set of parameters describing specific aspects of the interaction on the turn, dialogue and task level, from a system developer's point of view.
2. After the interaction, test participants are given a questionnaire to collect information about the perceptive quality features that are relevant to the formation of the overall quality impression of the human user. Such experiments can be performed with fully functional systems or those that are still in the development phase and in which parts of their modules have to be simulated. Details of the experimental set-up, the questionnaires and usability evaluation methods are given in clauses 7 and 8.

In laboratory experiments, both types of information can be obtained in parallel. In a field test situation with real users, however, instrumentally logged interaction parameters are often the unique source of information for the service provider in order to monitor the quality of the system. The amount of data that can be collected from an operating service may become very large. In this case, it is important to specify a core set of metrics that describe system performance, and to have tools at hand that automate a large part of the data analysis process. The task of the human evaluator is then to analyse and interpret this data, and to estimate the effect of the collected performance measures on the quality that would be perceived by a (prototypical) user. Some general considerations about the analysis and interpretation of test results are given in clause 9.

7 Experimental set-up

Subjective interaction experiments with a chatbot should be set up according to the general rules for subjective quality tests that are generally laid out in [ITU-T P.800], [ITU-T P.910] and [ITU-T P.911]. However, under certain circumstances. such as operating chatbots from mobile devices, it may be more appropriate to test in an environment that resembles typical usage conditions.

Classical multimedia laboratory tests are commonly carried out in "neutral" environments, such as sound-shielded rooms with daylight imitation. While these environments create controlled conditions for each participant, it is obviously not representative of real-life usage situations. In particular, when operating a chatbot from a portable device (e.g., smartphone, tablet), such an environment may generate misleading results with respect to the impact of device and display size on quality aspects. In such a case, it might be better to allow for more realistic – but less controlled – usage situations, in order to reach a better ecological validity.

Subjective experiments can either be carried out with fully working chatbots, or with the help of a human experimenter simulating missing parts of the chatbot. In order to obtain valid and reliable results, the (simulated) system, the test users, and the experimental task have to fulfil several requirements, see clauses 7.1 to 7.3. The interactions are usually logged and afterwards annotated by a human expert, so that interaction parameters can be calculated. After each interaction and after the whole test session, questionnaires have to be completed by test participants. These questionnaires allow different aspects of the quality of a chatbot-based service to be quantified. The design of such questionnaires is discussed in clause 8.

7.1 System set-up and Wizard-of-Oz simulation

In order to carry out interaction experiments with human users, a set-up providing the full functionality of the chatbot in real time has to be implemented. The exact nature of the set-up depends on the availability of system components and thus on the system development phase. If system components have not yet been implemented, or if an implementation is unfeasible (e.g., due to lack of data) or uneconomic, simulation of the respective components is required. The simulation of the

interactive system by a human being (the so-called Wizard-of-Oz (WoZ) simulation), is a well-accepted technique in the system development phase. At the same time, it serves as a tool for evaluation of the system in the loop. If a WoZ simulation instead of the fully implemented system is used, any deviations from the real system (e.g., a different response time of a human WoZ compared to an NLU component) should be reported and taken into account in the interpretation of the results.

7.2 Test scenarios

Because of the lack of a real motivation, laboratory tests often make use of experimental tasks that the participants have to carry out. The experimental task provides an explicit goal, which should not be confused with one that a user would like to reach in a real-life situation. Because of this discrepancy, valid user judgements on system helpfulness and acceptability cannot easily be obtained in a laboratory test set-up.

In a laboratory test, the experimental task is specified by a scenario description, which sets out a particular task that the subject has to perform through interaction with the system, e.g., to collect information about a specific product or to search for a specific tariff. Using such a scenario gives maximum control over the task carried out by the test participants, while at the same time covering a wide range of possible situations (and possible problems) in the interaction. Scenarios can be intentionally designed to test specific system functionalities (so-called development scenarios), or to cover a wide range of potential interaction situations which is desirable for evaluation. Thus, development scenarios are usually different from evaluation scenarios.

Scenarios help to identify different weaknesses in a dialogue, and thereby to increase the usability and acceptability of the final system. They establish user goals in terms of the task and the sub-domain addressed in a dialogue, and are a prerequisite to determine whether users achieve their goal. Without a pre-specified scenario, it is extremely difficult to compare results obtained in different dialogues, because user requests could differ and fall outside the system domain knowledge. If the influence of the task is a factor that has to be investigated in the experiment, the experimenter needs to ensure that all users execute the same tasks and on the basis of equal information. This can only be achieved by pre-specified scenarios.

Unfortunately, pre-specified scenarios can have some negative effects on user behaviour. Although they do not provide a real-life goal for test participants, scenarios prime users on how to interact with the system. Written scenarios may invite test subjects to imitate the language given in the scenario, leading to copying of the scenario text instead of the use of the user's own language.

Test participants carrying out pre-specified scenarios are usually not particularly concerned about the response of the system, as they do not really need the information. As a result, task success may not show a substantial effect on the usability judgements of test participants. In addition, it is possible that test participants do not always read instructions carefully, and may ignore or misinterpret key restrictions in the scenarios. The priming effect on user language can be reduced with the help of graphical scenario descriptions.

7.3 Test participants

The general rule for evaluation experiments is that the choice of test participants should be guided by the purpose of the test. For example, analytic assessment of specific system characteristics is only possible for trained test participants who are experts of the system under consideration. However, this group will not be able to judge overall aspects of system quality in a way that would not be influenced by their knowledge of the system. Valid overall quality judgements can only be expected from test participants who match as closely as possible future service users.

The factors expected to influence user behaviour and perception should be taken into account in selecting test participants in a representative way. Some of these factors are related to the vision, motor capacity and language usage that may be expected, such as age, individual body characteristics, physical status and native language. Because these factors may be critical for the NLU performance

and the appropriateness of the intended interaction device, quality judgements obtained from a user group differing in these characteristics might not reflect the quality that can be expected for the target user group. However, user groups are variable and ill defined. A service that is open to the general public will sooner or later be confronted with a large range of different users. Testing with specified users outside a narrowly defined target user group will therefore provide a measure of system robustness with respect to user characteristics.

A second group of user factors is related to experience and expertise with the system (including the input and output devices), the task and the domain. It can be expected that these factors will impact language usage, and thus interaction with the system. Users seem to develop specific interaction patterns, so called practices, when they get familiar with a new system. These practices may reflect a "cognitive model" the user develops of the system and depend on the user's former technology acquisition processes. Such a model is partly determined both by messages given to and coming from the system. The user generally assumes that their utterances are well understood by the system. If there are misunderstandings, the user gets confused and dialogue flow problems are likely to occur.

8 Questionnaires

In order to obtain information about quality features perceived by the user, subjective judgements have to be collected. Typically, these experiments are carried out using pre-defined scales on which test participants have to provide a quantitative judgment. In addition, open text answer options allow to collect subjective experiences which are not covered by the rating scales.

Scaling tasks can be carried out relatively easily, and untrained participants often prefer this method of judgement. They will yield valid and reliable results when two main requirements are satisfied: the items to be judged have to be chosen adequately and meaningfully, and the scaling measurement has to follow well-established rules. Scaling methods are described in detail in the psychometrics literature, e.g., in [b-Guilford], [b-Dunn-Rankin], [b-Borg]. For rating transmission quality, [ITU-T P.800] recommends absolute category rating (ACR), degradation category rating and comparison category rating methods. For rating the quality of spoken-dialogue-system-based services, judgements on continuous rating scales or on different ACR scales are usually solicited from the test subjects, see [ITU-T P.851]. An ACR scale consists of a number of discrete categories, one of which has to be chosen by the test subject. The categories are displayed visually and may be labelled with attributes for each category or for the extreme (left- and right-most) categories only. Examples of continuous rating scales are given in [ITU-T P.851].

The rating task on both continuous or category scales is often described in terms of a statement (e.g., "The system was easy to understand"), and test participants have to express their agreement with the statement by marking the appropriate box associated with a category on the scale. This method is based on early proposals made in [b-Likert], and an exemplary scale is depicted in Figure 1. Numbers are attributed to the categories or to the scale positions, depending on whether the statement is positive (from 1 for "strongly disagree" to 5 for "strongly agree") or negative (from 5 for "strongly disagree" to 1 for "strongly agree"), and the individual ratings are summed up for all subjects. Another possibility is to define self-explainatory labels for each category, as proposed, for example, in [ITU T P.800] for speech transmission quality experiments.

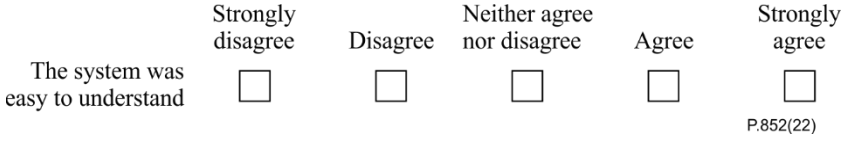

Figure 1 – Judgement on a statement in a way proposed in [b-Likert]

8.1 Questions related to user background

A number of questions should be answered at the beginning of the test session in order to describe users and their background, which is relevant to the experiment. These questions address the following items:

– personal information – age, gender, area or birth, current residence, language proficiency, visual impairments, motor impairments;

– task-related information – frequency of task, usual approach when resolving the task (alternative interfaces), motivation, other important task- and domain-related aspects;

– system-related information – experience with; chatbots; speech technology devices (speech recognition, synthesized speech, etc.); haptic input devices (smartphone, tablet, laptop, etc.).

8.2 Questions related to the individual interaction

Example questions may include those listed in Table 1 (question format and answer options are still to be determined).

Table 1 – Example questions related to the individual interaction

Quality aspect	Item
Overall quality	Overall impression
Information provided by the system	The system provided the desired information. The answers and solutions proposed by the system were clear. You would rate the information as wrong/true. The provided information was complete/incomplete
Communication with the system	You were always well understood by the system. You had to concentrate in order to understand what the system expected from him. The system responses were well readable. You had to put high effort into typing your messages. You were able to interact efficiently with the chatbot.
System behaviour	You knew at each point of the interaction what the system expected from you. In your opinion, the system processed your specifications correctly. The system behaviour was (not) always as expected. The system often makes mistakes (in understanding the user). The system reacted appropriately. The system reacted flexibly/inflexibly. You were able to control the interaction in the desired way. The system reacted too fast/too slowly. The system reacted in a polite way. The system responses were short/long. You perceived the dialogue as natural/unnatural. The course of the dialogue was clear/confusing. The dialogue was too short/too long. The course of the dialogue was smooth/bumpy. Misunderstandings could be cleared easily. You would have expected more help from the system.

Table 1 – Example questions related to the individual interaction

Quality aspect	Item
User's impression of the system	Overall, you were satisfied with the dialogue. The dialogue with the system was useful. It was easy for you to obtain the information you wanted. You perceived the dialogue as pleasant/unpleasant. During the dialogue, you felt relaxed/stressed. Using the system was frustrating/fun.
Acceptability	In the future, you would use the system again. You would advise your friends to use the system. You were satisfied with the solution offered by the system.

8.3 Questions related to the user's overall impression of the system

Example questions may include the following ones (question format and answer options are still to be defined):

Table 2 – Example questions related to the user's overall impression of the system

Quality aspect	Item
Overall quality	Overall impression
System behaviour	The system's way of expression was clear/unclear. The system reacted politely/impolitely. You would have expected more help from the system. The system was able to answer all of your questions. Misunderstandings could be cleared easily. The system controlled the flow of the dialogue. You were able to handle the system without any problems.
User's impression of the system	You enjoyed the dialogues.
Usability	The handling of the system was easy/complicated. The system appropriately informed you about its capabilities. The interactions with the system were worthwhile. You perceived this possibility to obtain information as helpful/not helpful. You rate the system as reliable/unreliable.
Acceptability	You prefer to use another source of information. You prefer a human operator. In the future, you would use the system again. Which characteristics of the system did you like most? Which characteristics of the system disturbed you mostly? Do you have any proposals for system improvement?

9 Analysis and interpretation of the collected information

In general, the guidelines given in [ITU-T P.1401] should be followed. The judgements that are obtained on closed rating scales can, for example, be analysed by means of bar charts or cumulative distributions. Although the distributions are not necessarily gaussian, it is common practice to calculate arithmetic mean values (and not medians) across all ratings obtained with a specified system configuration, see [ITU-T P.800]. For the mean values, confidence limits are evaluated and significance tests performed by conventional analysis of variance (ANOVA). The assumptions underlying an ANOVA (gaussian distribution and homogeneity of variances) are not always satisfied; nonetheless, this method seems to be robust enough to provide reasonable results if there are departures from statistically ideal conditions. In the case of a statistically significant effect of one of the variates (system configuration or voice, test subject, scenario, order of conditions in the experiment, test session, etc.), a *post-hoc* test can be used to perform pairwise comparisons in between the means or medians of pairs. Tukey's honestly significant difference is recommended as the critical value to use for multiple comparison tests. When the assumptions underlying a parametric statistic are not satisfied, analysis of ratings in terms of a median or mode, and the use of non-parametric tests for comparison, are advised.

Bibliography

[b-ITU-T P-Suppl.24] ITU-T P-series Recommendations – Supplement 24 (2005), *Parameters describing the interaction with spoken dialogue systems*.

[b-Bernsen] Bernsen N.O., Dybkjær H., Dybkjær L. (1998). *Designing interactive speech systems: From first ideas to user testing*. Berlin: Springer. 276 pp.

[b-Borg] Borg. I., Staufenbiel, T. (1993). *Theorien und Methoden der Skalierung: Eine Einführung* [Theories and methods of scaling: An introduction]. Bern: Hans Huber. 243 pp.

[b-Dunn-Rankin] Dunn-Rankin, P. (1983). *Scaling methods*. Hillsdale, NJ: Lawrence Erlbaum. 429 pp.

[b-Grice] Grice, H.P. (1975). Logic and conversation. In*:* Cole, P., Morgan, J.L., editors. *Syntax and semantics*, Vol. 3: *Speech acts*, pp. 41-58, New York, NY: Academic Press.

[b-Guilford] Guilford, J.P. (1954). *Psychometric methods*, 2nd edition. New York, NY: McGraw-Hill. 597 pp.

[b-Likert] Likert, R. (1932). A technique for the measurement of attitudes. *Arch. Psychol*. **22**(140**)**, pp. 1-55.

[b-Möller] Möller, S. (2000). *Assessment and prediction of speech quality in telecommunications*. Boston, MA: Kluwer. 244 pp.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality, telephone installations, local line networks**
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2022