

QRGS – Question Responses Generation *via* crowdsourcing

*Paweł Łupkowski*¹, *Jonathan Ginzburg*², *Ewelina Chmurska*¹,
*Adrianna Płatosz*¹, *Aleksandra Kwiecień*¹, *Barbara Adamska*¹,
*and Magdalena Szkalej*¹

¹ Adam Mickiewicz University

² Université Paris Cité, CNRS, Laboratoire de Linguistique Formelle

ABSTRACT

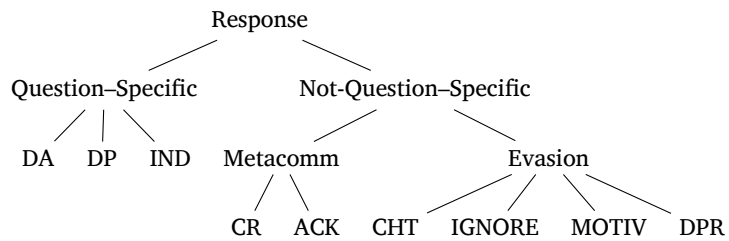
QRGS stands for the Question Responses Generation System. It is an online game-like framework designed for gathering various types of question responses. A QRGS user is asked to read a simple story and impersonate its main character. As the story unfolds the user is confronted with four questions and (s)he is expected to answer these in the way the main character would. In this way, we obtain responses to questions of a desired type. The data gathered *via* QRGS is a useful supplement to the linguistic data already present in language corpora – especially for languages for which such resources are sparse. As such, it opens the possibility for better understanding of the use of questions in natural language dialogues and analysing the response space of such questions. In this paper, we present the main idea of QRGS and the results of five studies (in Polish and in English) that test the framework. Our discussion addresses issues concerning the efficiency and accuracy of the proposed approach. We also discuss the availability of the QRGS and its potential future improvements.

Keywords:
gamification,
crowdsourcing,
questions,
responses,
language
resources

This paper describes how certain types of responses to questions (i.e. direct, indirect and evasive ones) may be gathered *via* a relatively simple and easy to use crowdsourcing framework. Question Responses Generation System (QRGS) is designed and implemented with the aim set for providing supplementary data for the study of the response space for questions (Ginzburg *et al.* 2019, 2022).

Ginzburg *et al.* (2019, 2022) present extensive corpus studies of the BNC (Burnard 2007), BEE (Rosé *et al.* 1999), Maptask (Anderson *et al.* 1991) and CornellMovie (Danescu-Niculescu-Mizil and Lee 2011) corpora for English (which include 607, 262, 460, and 911 question/response pairs respectively) and data for Polish using the Spokes corpus (Pęzik 2014; 694 question/response pairs) On this basis, a typology of responses to questions is proposed – see Figure 1.

Figure 1:
Typology
of responses
to questions.
Source: Ginzburg
et al. 2022, p. 86



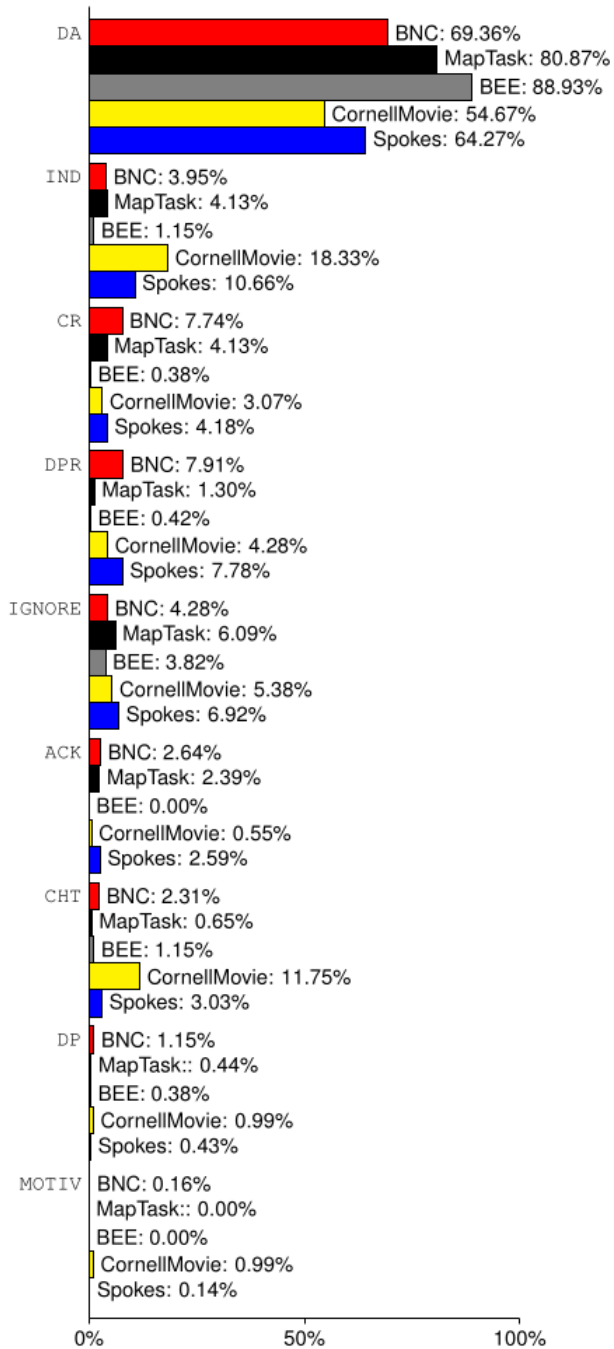
The two main categories of this typology are (1) *question-specific responses* (covering direct answers, dependent questions and indirect answers) and (2) *non-question-specific responses*. Direct answers (DA) provide an answer straightforwardly. For indirect answers (IND), one needs to infer an answer from the utterance. Dependent questions (DP) constitute a case where a question is provided as a response. What is more, the answer to the initial question (q1) depends on the answer to the query-response (q2). As for the non-question specific responses, we have: Clarification responses (CR) which address something that was not completely understood in the initial question (q1) and acknowledgements (ACK) wherein a speaker acknowledges that s(he) has heard and understood the question, e.g. *mhm*, *aha*, etc. Moving on to evasive question-responses, first we mention change-the-topic

(CHT). Instead of answering q1, the agent directly provides q2 and attempts to turn the table on the original querier. The original querier is pressured to answer q2 and put q1 aside. An IGNORE type of query-response appears when q2 relates to the situation described by q1 but not directly to the initial question. MOTIV is the type which addresses the motivation underlying asking q1. Whether an answer to q1 will be provided depends on a satisfactory answer to q2. DPR involves cases where the speaker states that it is difficult to provide an answer, points at a different information source, etc. or the speaker states that s(he) does not know the answer.

The corpus study revealed that for English the most frequent response classes in all four corpora are direct answers; the second most frequent class in the BNC is Difficult to Provide an Answer (DPR=7.91%), while in CornellMovie, the next biggest is indirect answers (IND=18.33%), whereas for the MapTask and BEE these are IGNORE (6.09% and 3.82% respectively). For Polish, the two most frequent classes of responses for Spokes are answers: direct ones (DA=64.27%) and – much smaller – indirect ones (IND=10.66%). The next two most frequent classes are DPR (stating that a person does not know the answer to the question, or it is difficult to provide one, DPR=7.78%) and utterances ignoring the question asked (questions and declaratives, IGNORE=6.92%). As illustrated in Figure 2 other classes are really rare – for MOTIV under 1% of the sample. This means that for certain response classes we have gathered very small numbers of examples. Such a result poses at least two challenges (as pointed out in the summary of Ginzburg *et al.* 2022). Firstly, how to collect more linguistic data for cross-linguistic testing? In the reviewed work, large English corpora were used but still certain classes of responses had small numbers of examples. The situation is even more challenging for languages lacking large or even hardly any speech corpora. Secondly, such a situation raises a serious difficulty when one thinks about potential applications of the corpus study with respect to dialogue interfaces. For such an application, machine learning should be used to acquire the response classification scheme (see Yusupujiang and Ginzburg 2022). This means that additional training and testing data are needed.

This brings us to a twofold motivation for designing QRGS. Firstly, to supplement the data from language corpora and open the way to

Figure 2:
 Response types frequency
 (BNC, n = 607;
 BEE, n = 262;
 MapTask, n = 460;
 CornellMovie, n = 911;
 Spokes, n = 694).
 Source: Ginzburg et al.
 2022, p. 93



apply machine learning approaches. Secondly, as not all languages have sizable linguistic corpora (see the disproportionate numbers for English and Polish in the aforementioned study) QRGS aims at closing this gap. This would pave the way for the cross-linguistic testing of the findings about the response space to questions (but not only).

The paper is structured as follows. Section 2 covers the main idea of QRGS and points at earlier work which it drew its inspiration. We also compare QRGS to selected, already existing crowdsourced solutions. Sections 3 to 6 present a series of QRGS evaluation studies. Starting from the pilot study where the effectiveness of the approach and correctness of the gathered data were checked, through questions concerning the non-native speakers' participation in QRGS, the role of game-like elements and the QRGS story theme. In Section 7, we describe a design of the crowdsourced evaluation module for QRGS. We end with the description of the part of QRGS data published as a part of the Erotetic Reasoning Corpus (Łupkowski *et al.* 2017). The summary gathers all the findings and points out aspects of QRGS that need further studies and improvements.

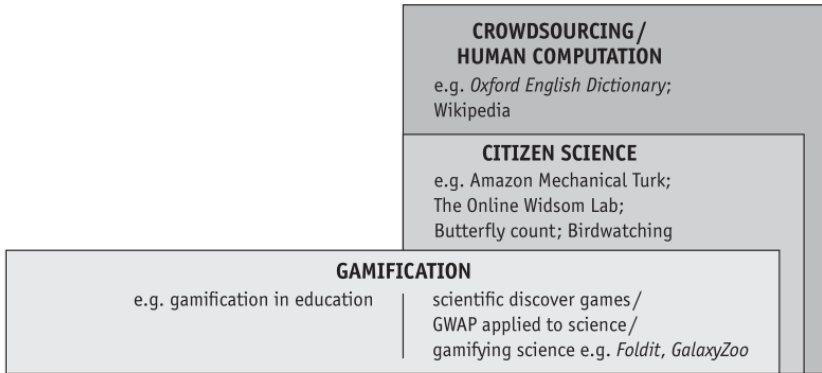
QUESTION RESPONSES GENERATION SYSTEM – THE IDEA

2

The idea behind QRGS is to use *crowdsourcing* for relatively easy and effective collection of specific linguistic data. As such it may be identified as an example of a *scientific discovery game* (Cooper *et al.* 2010). A game of this kind is intended to help in processing large amounts of data obtained in scientific research. Two main tasks performed by human players in this case are mainly intelligent data analysis and classification tasks.

Scientific discovery games lie at the intersection of crowdsourcing, human computation and gamification – see Figure 3. Thus, we find methods and solutions known from these fields applied to solve given scientific problems. Typically, non-experts are employed to solve a given problem. As users perform the task in question in their free time and (usually) without gratification, it should be framed as relatively

Figure 3:
A conceptual
map of scientific
discovery games.
Source:
Łupkowski and
Dziedzic 2016,
p. 129



simple and not time-consuming. Using game elements in a design is aimed at providing additional fun to the task, and also to motivate a user (e.g. with the points, achievements or leader boards).

A notable example of such a solution is Galaxy Zoo (Lintott *et al.* 2008). Galaxy Zoo was designed as a result of the huge amounts of astronomical data obtained from the Sloan Digital Sky Survey (SDSS). The problem for astronomers was to provide visual morphological classifications for nearly one million galaxies extracted from SDSS. Such a task is extremely difficult for current algorithms, and the work performed by small groups of experts had low efficiency (cf. Lintott *et al.* 2008). The idea of Galaxy Zoo is to provide users with a simple and brief tutorial and then allow them to perform classifications, using a very intuitive (symbolic) interface. Galaxy Zoo users are provided with photos of galaxies' from SDSS (the players are additionally motivated by the fact that most of the pictures have not been seen by anybody before them). Galaxy Zoo was so successful that it served as a template for analogous solutions for classification problems from other fields which are now hosted on the Zooniverse¹.

Another interesting project of this kind is Foldit (Dsilva *et al.* 2019). Foldit is a perfect example of how a very difficult problem (3D modelling of protein structures) may be presented in the form of an easy to understand task – simple puzzle game.

From the field of linguistics it is worth mentioning such inspiring projects as PhraseDetectives (Chamberlain *et al.* 2008), which collects

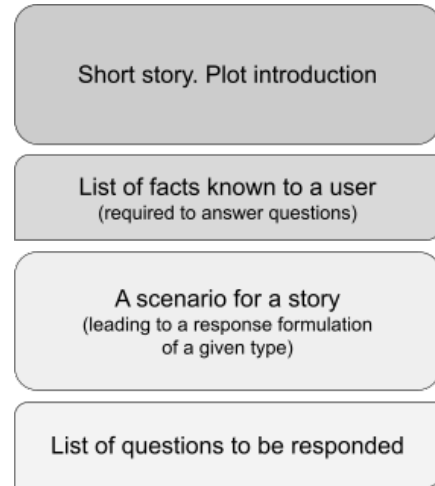
¹<https://www.zooniverse.org/projects>.

collaborative anaphoric decisions from online volunteers; Wordrobe (Venhuizen *et al.* 2013) which is a set of simple games developed to enable semantic annotation of the natural language data from the Groningen Meaning Bank (GMB); or RoboCorp (Dziedzic 2016) – the mobile game developed with the aim of annotation of the named entities retrieved from the Polish National Corpus (Przepiórkowski *et al.* 2011).

A direct inspiration for QRGS comes from the previous gamified solution related to questions and answers studies, which is the Quest-Gen described in Łupkowski and Wietrzycka 2015 and Ignaszak and Łupkowski 2017. QuestGen is a game-like system, in which players generate questions of a specific form while solving a detective game. In the game, two randomly chosen players are engaged in solving a detective puzzle. One of them plays as the Detective, while the other is called the Informer. The aim for the Detective is to solve the presented puzzle by questioning the Informer. Each story in the game has two formulations (one for the Detective and one for the Informer), containing all the additional data necessary to solve the puzzle. Each story should be solved within a given time limit. For each story the players switched roles, from the Detective to the Informer and *vice versa*. Players were not supervised in any way, they were just playing the game. Crucially, stories' plots were formulated according to erotetic search scenarios, a tool developed within Inferential Erotetic Logic (Wiśniewski 2013). Thanks to this, each story has only one correct solution and a normative way to reach it (pointed out by the underlying erotetic search scenario). Overall, 116 game transcripts from 40 players were collected. The general solution statistics for the study sample (all six stories) is the following: 91 solutions are correct, out of which 44 are normative, i.e. solved exactly according to the e-scenario underpinning a given story. In 18 cases, Detectives provided incorrect solutions and in 7 they did not provide any solution (mostly due to time constraints) – see detailed discussion in Ignaszak and Łupkowski 2017.

QRGS relies on a very similar schema. A QRGS user is asked to read the simple story and impersonate its main character. As the story unfolds, a user is confronted with four questions and (s)he is expected to answer them in the way the main character would do that. As the story unfolds, a user is confronted with questions related to the story

Figure 4:
A general QRGS schema



and (s)he is expected to answer them in a way the main character would in the given context – see Figure 4 for a general QRGS schema. Stories which are prepared for QRGS to this point are presented in detail in Section 3 and Appendices A, B and C.

Here we should also mention yet another crowdsourced solution for gathering question-response pairs. The motivation for the solution also comes from the corpus study presented earlier and it is aimed at addressing the challenge of characterising the response space to questions in a low-resource language – Uyghur. The early design is presented in Yusupujiang and Ginzburg 2020 and Yusupujiang and Ginzburg 2021. Initial studies and results are discussed in Yusupujiang and Ginzburg 2022. The paper presents a Uyghur dialogue corpus based on a chatroom environment (using the Rocket.Chat implementation). The Uyghur Chat-based Dialogue Corpus (UgChDial) is divided into two parts: (1) Two-party dialogues and (2) Multi-party dialogues. It consists of 25 chat sessions, with 6 participants, resulting in 1,581 question-response pairs. The sessions were based on different scenarios and topics. The analogue to QRGS are role-playing scenarios, which require participants to act according to certain situations (such as police vs. criminal, debtor vs. debtee, sales person vs. a customer with complaint). This is aimed at retrieving evasive or cooperative responses from users.

PILOT STUDY. PROOF OF CONCEPT

3

In this section, we present the pilot study of QRGS. The study was aimed at answering the following research questions.

1. How *effective* is QRGS in terms of data gathering – how many question/response pairs (Q-Rs) will be gathered and how long will it take?
2. How (linguistically) *interesting* are the gathered responses? Namely, will the responses generated to presented questions differ between subjects? Will they be comparable to responses that may be observed during a dialogue?
3. Are the gathered responses *correct*? I.e. are they of the type which is expected for a given scenario for the story?

Tools and materials

3.1

For the study, two stories were prepared: “The Bomb” and “The Party”. We describe them in detail below. For each story, we firstly present a user with the introductory plot including the facts known to the user. After that, four scenarios are presented to a user along with the questions (the same four questions are used for all scenarios). The task of a user is to immerse into the story and provide responses to the presented questions, which will be formulated in a manner appropriate to a given story and the current scenario).

The first story is entitled “The Bomb” and was adapted from the previous studies related to questions and question answering (Urbański *et al.* 2016a). The plot presented to a participant is the following.

A bomb was planted in the main train station of Nibyjunkcja. You are the chief of security at the train station where the bomb was planted. After checking the security cameras you have established the following facts:

1. The bomb was planted under the third pillar.
2. The bomb has the size of a shoe-box.

3. It was planted by a tall guy dressed in a red T-shirt.
4. It was planted between 8:00 and 8:30 A.M.

The first scenario for the story is such that a subject should provide a direct answer to the questions asked. It is entitled “The coordinator of the sapper unit”.

You are approached by the coordinator of the sapper unit who is trying to establish which wire to cut in order to disarm the bomb. You are *obliged to be truthful and give direct and precise answers* to his questions. Please answer the following:

1. Do you know where the bomb was planted?
2. How big is the bomb?
3. Can you describe the suspect?
4. Can you tell me when the bomb was planted?

As a result we should obtain four direct answers (DA) to the introduced questions.

The second scenario for the story is: “A trusted journalist”. For this we are expecting indirect answers (IA). To encourage a participant to provide such responses the following lead is used.

You are approached by Frank, a journalist for the local “Nibyjuncja Today”. You have known Frank for a long time and trust him. He wants to gather some news about the situation on the station. Given that the investigation is in progress you cannot give Frank direct information. Nonetheless, since you trust him, *try to provide truthful information but in an indirect manner*. Please answer the following questions of Frank. / Here the same set of questions is used as for the first scenario, p. 222. /

The next scenario is “A journalist you do not trust”, which is aimed at retrieving evasive and polite responses.

You are approached by a journalist you do not know. His ID indicates that he came from the capital and works for the big journal “NBJ News”. You do not trust him. However, you are obliged to answer his questions in order to avoid problems with the press. Please answer the journalist’s questions

in such a way that he will understand that you do not want to answer his questions (be polite...). Please answer the following. / Here the same set of questions is used as for the first scenario, p. 222. /

And the last one is entitled “A random guy” (for evasive and impolite responses).

You are approached by a random guy from the crowd surrounding the scene. He tries to ask you some questions. Please answer them in such a way that he will understand immediately that you *do not want to answer* his questions (you do not have to be extremely polite, however you should not lie, or simply answer using “no comments”). Please answer the following. / Here the same set of questions is used as for the first scenario, p. 222. /

“The Party”. The second story considers inviting people to a party. It also has four scenarios. The plot is introduced in the following paragraph.

Imagine that you are organising a party next Saturday. You want to invite just several close friends: Ann, John, Frank, Alice and Bill. The party is on Saturday and starts at 8 P.M. You would like it to end around midnight. You plan a barbecue and beer in the garden.

As in the previous case, four scenarios, each aimed at a different category of responses obtained were designed. “Alice” for direct answers (DA).

In a shop, you are approached by Alice. She is already invited to the party and has accepted the invitation, so you can *openly and directly* answer her questions. Please answer the following questions asked by Alice:

1. How many people will there be at the party?
2. Is Ann invited?
3. Will there be any alcohol at the party?
4. When do you want to start?

“Helen” for evasive answers (polite).

In a shop, you are approached by Helen. She is your neighbour and somehow got to know about the party. You do not want to discuss any details with her so answer her questions in such a way that she will know that *you do not want to answer them* (still do be polite, she is your neighbour after all). Please answer the following. / Here the same set of questions is used as for the first “The Party” scenario. /

“Willy” for evasive answers (impolite).

While coming back from work you are approached by little Willy, your neighbours’ son. He tries to ask you some questions. Please answer them in such a way that he will understand immediately that you *do not want to answer* his questions (you do not have to be extremely polite, however you should not lie). Please answer the following. / Here the same set of questions is used as for the first “The Party” scenario, p. 223. /

And the last scenario is “John” for indirect answers (IA).

During the evening John calls you. You are in one room with your friend, who does not know about the party. John is asking some questions. Please answer them in an *indirect* manner so that your friend will not get any idea concerning the party. Please answer the following. / Here the same set of questions is used as for the first “The Party” scenario, p. 223. /

The summary of expected question responses types to the different formulations of stories is presented in Table 1.

3.2

The Procedure and Participants

Stories and questions were presented online with the use of the Google Forms platform (each scenario for a story separately). Only text was presented, no additional images were included to supplement stories. Instructions for the participants were the following:

Story	Scenario	Expected response
“The Bomb”	The coordinator	DA
	A journalist (trusted)	IA
	A journalist (untrusted)	Evasive (polite)
	A random guy	Evasive (impolite)
“The Party”	Alice	DA
	Helen	Evasive (polite)
	Willy	Evasive (impolite)
	John	IA

Table 1:
Expected question responses to the different formulations of stories

Below you will find a short story and 4 questions for it. Please try as best as you can to get into the character and write how you would answer the questions below in real life. The speed of completing the task will not be measured, so please take as much time as you need.

Invitations for participants (each participant for each variant of a story) were sent out *via* social media. No information was collected about the participants (which is a common practice for crowdsourcing tools), however the invitations were intentionally sent to people without experience in linguistics and with a high level of English language proficiency. 25 participants took part in the study. The data collection lasted from the 1st to the 5th May 2018.

Results and data validation

3.3

Effectiveness. Overall we gathered a sample of 100 Q-R pairs generated by 25 participants in just five days. The summary of generated responses is presented in Table 2. One may conclude that QRGS is effective when it comes to the numbers of gathered responses and the data collection time. This is mainly due to the fact that the task for a participant is not very demanding and the data collection itself does not require any supervision from the researcher.

Let us now take a closer look at the variety of the gathered data. In order to be useful for the intended use, the question responses retrieved for one question should have different formulations. The QRGS

Table 2:
Number
of responses
generated
for each
QRGS story

Story	Participants	Responses generated	Response type
Bomb 1	4	16	DA
Bomb 2	4	16	IA
Bomb 3	5	20	EAP
Bomb 4	4	16	EAI
Party 1	2	8	DA
Party 2	2	8	IA
Party 3	2	8	EAP
Party 4	2	8	EAI
Sum	25	100	–

data would not be interesting if we would obtain e.g., 50 “Yes” responses to the question “Do you know where the bomb was planted?”. Fortunately this is not the case. We observe a wide variety of the retrieved question responses. Consider the following examples.

For the “Bomb” history and scenario “Untrusted journalist” and question *Do you know where the bomb was planted?* we have responses such as the following (in all examples we preserve the original spelling):

- This information is available for me.
- Where would you plant such a bomb?
- All stations are being monitored. We have the data from the cameras – therefore we will be able to localise any unusual behaviour.
- Yes.
- There are some clues to figure out where the bomb is. It is probably somewhere nearby.

And for the same story, but the scenario “Trusted journalist” and question *How big is the bomb?*:

- The bomb could have been carried by a single person in a handbag
- Did you finally manage to reduce the size of space occupied by your precious collection by throwing out the unnecessary stuff?
- Let’s say that you can carry it in a shopping bag.

One may observe that the responses generated by our participants vary with the respect to complexity, length and style. Thus we are

Response category	Generated	Correct	(% corr)
DA	24	24	100%
IA	24	13	54%
EAP	28	18	64%
EAI	24	22	92%
All	100	77	77%

Table 3:
Summary of responses' correctness with respect to categories

gathering responses which are close to the natural language dialogue outcomes. This also suggests that, to a large extent, our participants were able to immerse into the storyline presented and answer questions suitable to the plot.

Correctness. Naturally the most important question is whether these generated responses were of an expected type – i.e. were they correct? This aspect is very important as the data gathering with QRGS is not supervised. A high percentage of correct answers is needed for the data to be useful for future applications. To answer this question, the responses were manually evaluated by two researchers. The aim was to check whether the actual responses given by participants fit into the expected categories.

Each response was tagged independently by two annotators using the following tagset: DA (direct answers), IA (indirect answers), EAP (evasive polite), EAI (evasive impolite), OTHER (for cases not matching the listed categories). Inter-annotator agreement was then calculated with the use of Kappa statistics. In what followed, a final tag was assigned to each response as a result of discussion between annotators. This tag was then compared with the intended category for a given response.

For the reported study the agreement between both raters was measured using Cohen's kappa coefficient. This was established using the R programming language (version 3.5.0) and the irr library. Cohen's kappa was 0.775 (which indicates the substantial agreement between raters, see Viera and Garrett 2005).

The manual evaluation shows that 77% of responses are in line with the predictions – see details in Table 3.

Error analysis. Participants of this study had no problems with providing direct answers. All of the gathered DA responses were

correct. As for indirect answers (IA) we observe a common mistake, which is providing a DA instead of an IA, like in the following example:

A: Do you know where the bomb was planted?

B: *Yes, somewhere in the station.*

This constitutes 10 of the 11 observed errors. Only one error was that instead of an IA an evasive answer was provided.

A: How many people will there be at the party?

B: *I really enjoy spending time with my close friends.*

Let us now take a closer look at evasive responses. All the mistakes in this case were that instead of an evasive answer a direct one was provided (however, these were partial answers). This is exemplified in the following:

A: Can you tell me when the bomb was planted?

B: *Certainly today.*

Interestingly, more errors for evasive responses were observed for the polite condition than for the impolite condition – this suggests that for the participants the impolite condition was easier to formulate such responses.

Summary. QRGS proved to be a simple and effective crowdsourcing tool for gathering interesting data. The task is not demanding for a user and is thus very quick to complete. QRGS needs no supervision on the level of data collection. Also, the data correctness in our study is satisfactory. It is worth stressing that incorrect responses (i.e. the ones that do not match the expected type for a given scenario) are not useless for future applications. Manual re-annotation leads to their classification to the appropriate type.

The pilot study results presented in this section led to further research questions and potential improvements for QRGS. Firstly, we have not gathered any information concerning our participants. For a QRGS evaluation it would be useful to learn whether language proficiency matters for data generated with QRGS; in particular, whether we would observe differences between native and non-native speakers. Another question addresses the level of game-like elements involved in QRGS – would it be better to supplement QRGS stories with

graphics? Last but not least, it is an open question whether the type of story plot for QRGS matters for the results. We address these questions in the following sections.

NATIVE VS NON-NATIVE ENGLISH SPEAKERS 4

In this section, we present a study focused on the research question whether we would observe any differences in QRGS outputs for native and non-native English speakers groups. Given that in the pilot study we did not gather any data concerning participants, this question remains open. The answer is important for potential QRGS applications.

Materials 4.1

For the purpose of this study, two previously written QRGS stories were used (“Bomb” and “Party”). Also, two new ones were prepared. These were “The Epilepsy” and “The Secret Santa”. The first one is a story that your co-worker, Anna, has just had an epilepsy attack and you helped her and called an ambulance. The second story revolves around you and your friends having decided to organise a Secret Santa event this year and you considering different ideas for presents. New stories were prepared exactly in line with the first two. After a short introduction of the situation and of the known facts, a user is presented with four scenarios along with questions – each scenario formulated in such a way that it leads to different types of responses (direct ones, indirect ones, evasive polite and evasive impolite). The complete stories along with their corresponding scenarios are presented in Appendix A.

Procedure 4.2

The study was conducted *via* the Internet using the Google Forms platform. Participants were presented with one short story each and asked to answer 4 questions. Participants were asked to “enter into” the situation and empathise with the assigned role and provide written answers as if they were responding directly to the character from the

story. It was made clear in the instructions that no time limit for the task completion was assumed. After answering all the four questions, the participant was asked to answer several demographic questions (covering age, gender, education, native language and for non-native English speakers their English proficiency level).

4.3

Study group

The group consisted of 49 participants, of which 28 were female, 19 were male and 2 preferred not to reveal their gender information. Participants were recruited *via* social media. The average age was 32.02 (SD=11.67, min=15, max=57). The declared education level was the following: doctoral degree: 7; university degree: 27; high school diploma or equivalent degree: 9; less than high school diploma: 6. Most importantly 31 participants were native speakers of English and 18 were non-native speakers (12 Polish; 2 Czech; 1 Spanish; 1 Swedish; 1 Azerbaijani; 1 Arabic). The declared English proficiency level for the non-native speaker group was the following: A2 (Elementary): 1; B1 (Intermediate): 2; B2 (Upper Intermediate): 4; C1 (Advanced): 7 and C2 (Proficient): 4.

4.4

Results and data validation

Effectiveness. The data were collected during December 2018 and March–May 2019. Overall participants generated 196 responses. 124 in the native speakers group and 72 in the non-native group – see details in Table 4.

Table 4:
Summary
of responses'
correctness
with respect
to groups
and categories

Response type	Native	(% corr)	Non-native	(% corr)
DA	48	96%	20	100%
IA	12	42%	12	58%
EAP	36	36%	28	28%
EAI	28	50%	12	12%
All	124	63%	72	62%

Variety. Firstly, we observe that the generated responses are interesting and differ between participants. Examples are provided below. For the “Bomb” story (scenario the unit coordinator) and question: *Do you know where the bomb was planted?* we have for example:

- Under the third pillar in the Nibyjunkcja main train station.
- In the main train station of Nibyjunkcja, at the base of the third pillar.
- In the main train station of Nibyjunkcja.

For the “Secret Santa” story (DA scenario) and question: *What are we giving to Joe?*

- Craft Beer Brewing Kit.
- We’re giving him the Craft Beer Brewing Kit.
- Craft Beer Brewing Kit.
- the craft beer brewing kit.

We also find responses which are carefully prepared and much longer than the presented ones.

What are we giving to Joe?: I don’t know much about home brewing, so this would be a bit difficult. I’d try to reference something that ‘hops’. My best idea so far is to talk about a trip to the zoo. My family went to the zoo last week. We watched the kangaroos for ages and my daughter insisted on hops, hops, hops to get around after that. We found a nice kit at the gift store that will allow us to make our own hoppy creature. It’ll be especially good for taking with us to enjoy at BBQs.

How much is the contribution rate?: If I were in the US, I’d say something about Hamilton, the musical. (Alexander Hamilton is on the 10 Dollar bill.) If I were in Australia, I would say something about the Wattle tree or quote The Man from Iron Bark (Both the Wattle and Banjo Patterson are on the A 10 Dollar note).

As in the case of the pilot study we may conclude that QRGS data are interesting and reminiscent of responses provided in spontaneous conversation.

Table 5:
Length of responses
(number of characters)
generated in QRGS

Group	Mean	SD	Median	Min	Max
Native	38.34	57.75	26.50	1	420
Non-native	30.56	29.79	19.50	2	152

Correctness. For the data evaluation a procedure analogous to the one described in Section 3.3 was applied.

Two annotators were engaged in the evaluation. The agreement between both raters was measured using Cohen's kappa coefficient (established using the R programming language (version 3.5.0) and the irr library). Kappa for the native speakers group = 0.717, and for the non-native speakers group = 0.639. Both results indicate substantial agreement between raters (see Viera and Garrett 2005).

The general correctness of the respondents in the group of native speakers was 63% and for the non-native speakers group it was 62%. One may conclude that for the correctness factor of the gathered data these groups do not differ. The detailed summary is presented in Table 4. As expected, providing a Direct Answer to a question was the easiest task, with almost 100% accuracy in both groups.

We also decided to take a closer look at the length of the generated responses in order to check whether they differ between groups. The intuition behind this step is that the length of a response (in the numbers of characters used) provides a rough (quantitative) indication of how elaborate the response is. One may expect that the native group would provide longer, more elaborate responses.

The length of responses for the groups is presented in Table 5 and Figure 5.

The Wilcoxon Test shows that there are no statistically significant differences between the groups ($W = 4504.5$, $p = 0.9168$). The responses provided by QRGS users do not differ between groups of native and non-native English speakers. Their correctness is at a similar level. Also, the average number of characters per response indicates that responses were similarly complex when it comes to formulation. Such a result is promising for future QRGS implementation for popular languages (such as English). QRGS may be used to gather data for such languages even if the access to the group of native speakers is limited. Naturally, this may not be easily generalised for other languages and needs further testing.

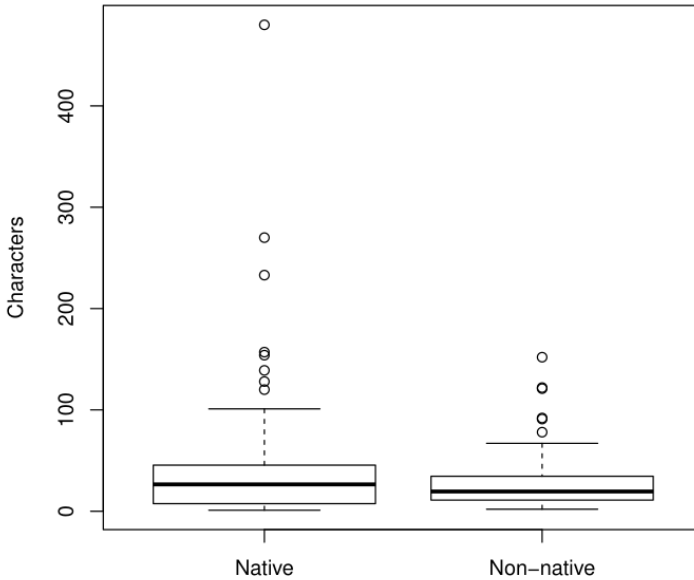


Figure 5: Comparison of number of characters used in responses generated by group of native and non-native English speakers

GRAPHICAL VS TEXTUAL VERSION

5

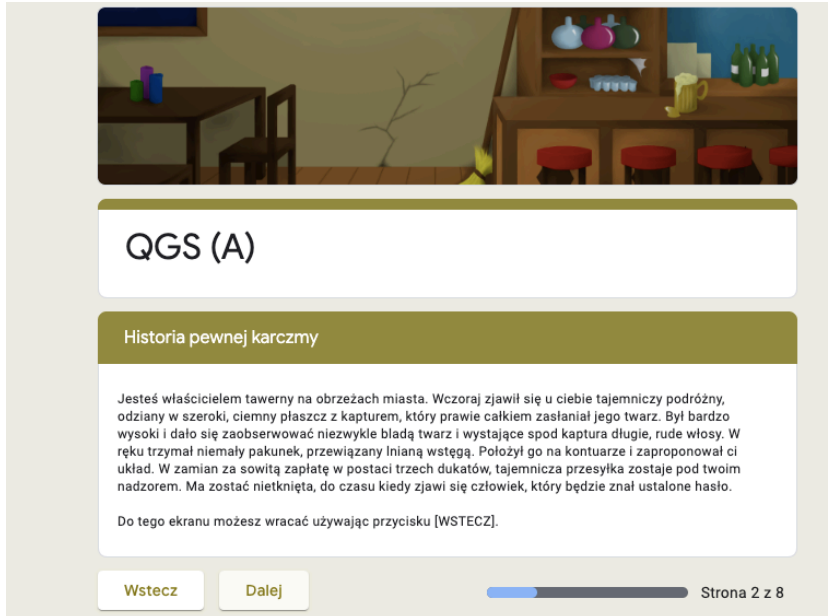
The following section describes a study where we asked the research question whether QRGS should involve more game-like elements, especially graphical ones. The intuition would be that a more game-like system will stronger immerse participants. The more immersed the participant, the better (i.e. more natural and correct) responses provided for QRGS stories. Thus, we designed a graphical version of QRGS for the experimental group, while the control group used the already tested textual one. We studied the differences in outcomes in terms of correctness of the data, response length and the self-declared engagement of users.

Materials

5.1

For the purpose of this study, a new QRGS story was prepared in Polish. It is entitled “The Tavern” and tells a story of a tavern owner who is asked for a favour – storing a mystery object for an unknown person. The complete story along with its corresponding scenarios is presented in Appendix B.

Figure 6:
Textual version
of QRGs. English
translation
of the story
in Appendix B



As mentioned above, two versions of QRGs were prepared. First, the traditional one, i.e. textual (as presented in Figure 6). The header of a questionnaire was supplemented with one simple graphic presenting the inside of a tavern. The second version was a graphical one with the style inspired by RPG games. The story was presented step by step with the appropriate illustrations (Figure 7). Also, the characters from the story were presented in the visual form (Figure 8). It is worth stressing that the text presented in both versions was identical.

In order to assess the engagement level of the participants, we employed the shortened version of the IMUW questionnaire. IMUW (Wasielewska and Łupkowski 2022) is a questionnaire based on the Polish adaptation (Strojny and Strojny 2014) of the immersion questionnaire (Jennett *et al.* 2008). It measures self declared engagement into task performance. The full IMUW consists of 25 items. For the purpose of the study we, prepared a 10 item short version (as the IMUW reliability study reports that it is a one factor questionnaire). Below, we present this IMUW version with the English translation of items.

1. W jakim stopniu zadanie podtrzymywało Twoją uwagę? / *To what extent did the task hold your attention?*

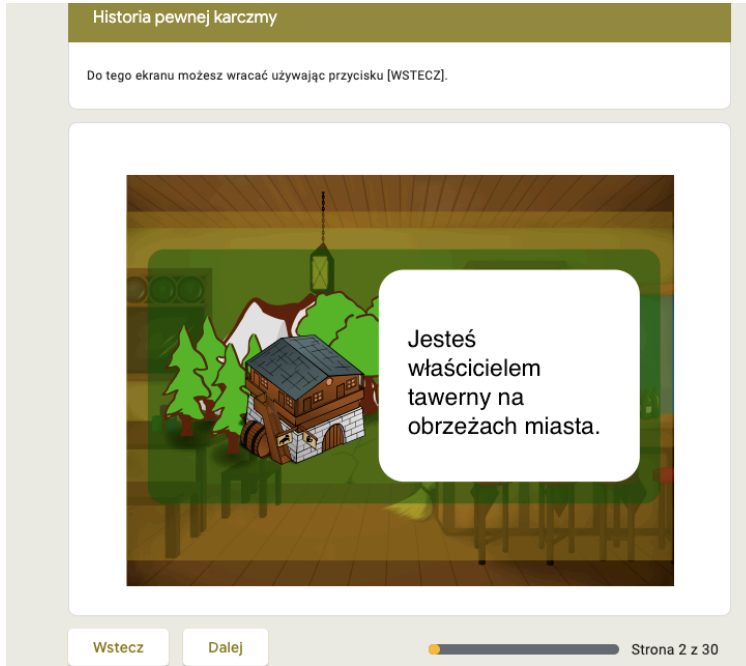


Figure 7: Graphical QRGS version. The story unfolds step by step and is illustrated. The panel says “You are the owner of a tavern in the suburbs”. Full story in English in Appendix B

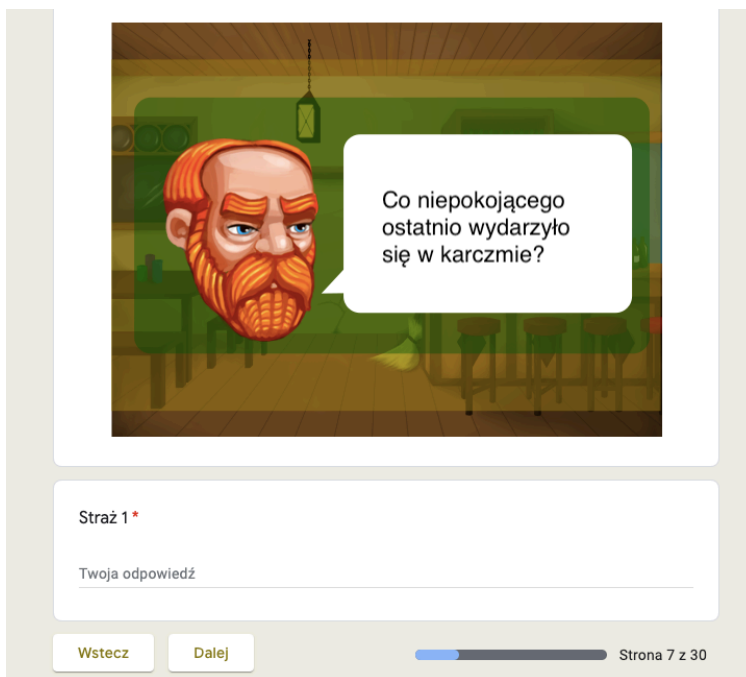


Figure 8: Graphical QRGS version. Characters from the story are presented, and dialogues are simulated. The panel says “What was the worrying thing that happened at the tavern?”. Full story in English in Appendix B

2. W jakim stopniu odczuwałeś(aś), że jesteś skupiony(a) na zadaniu? / *To what extent did you feel you were focused on the task?*
3. Jak dużo wysiłku włożyłeś(aś) w wykonanie zadania? / *How much effort did you put into playing the game?*
4. Czy odczuwałeś(aś) w którejkolwiek chwili potrzebę przerwania wykonywania zadania i zobaczenia, co się dzieje wokół? / *Did you feel the urge at any point to stop performing the task and see what was happening around you?*
5. W jakim stopniu odczuwałeś(aś), że zadanie jest czymś, czego raczej doświadczasz niż po prostu czymś, co robisz? / *To what extent did you feel that the task was something you were experiencing, rather than something you were just doing?*
6. W jakim stopniu czułeś(aś) się emocjonalnie zaangażowany(a) w zadanie? / *To what extent did you feel emotionally engaged in the task?*
7. W jakim stopniu byłeś(aś) zainteresowany(a) tym, jak potoczy się fabuła czytanego przez Ciebie tekstu? / *To what extent were you interested in seeing how the presented story plot would progress?*
8. W jakim stopniu podobał Ci się poziom artystyczny tekstu? / *To what extent did you enjoy the presented text?*
9. Jak dużą czerpałeś(aś) przyjemność z wykonywania zadania? / *How much would you say you enjoyed performing the task?*
10. Czy chciałbyś(aś) wykonać zadanie jeszcze raz? / *Would you like to perform the task again?*

5.2

Procedure

The study was conducted online with the use of Google Forms. Participants were invited to take part in the study via a link on the social media pages. The link led to the page where a participant was randomly assigned to one of the groups. Participants received necessary information about the study and provided their agreement to take part. After that, they were presented with the story followed by four scenarios with questions. Next, they filled out the IMUW questionnaire and provided basic demographic data.

Study group

5.3

70 participants took part in the study. 35 in group A (textual QRGS version), aged 18-31 (mean 22.43; SD = 3.19), 62.9% women. 35 in group B (graphical QRGS version), aged 19-41 (mean 24.85; SD = 5.67), 54,3% women. All participants were native Polish speakers.

Results and data validation

5.4

The data was collected from the 10th of March 2019 till the 22nd of March 2019. Overall 1,120 responses were collected. The variety of responses was satisfactory, as observed for previous studies in English.

Correctness. To assess response correctness, we randomly chose 100 Q-R pairs from group A and 100 from group B. A procedure analogous to the one described in Section 3.3 was applied. For this study, each response was tagged independently by three annotators, thus inter-annotator agreement was controlled for with the use of the Fleiss kappa coefficient (established using the R programming language, version 3.5.0, with the irr package). Fleiss' kappa was 0.504 for group A and 0.575 for group B. As for percentage of correct answers, we got 49% for group A and 59% for group B – details are presented in Tables 6 and 7. We observed a small advantage in the case of the graphical QRGS when it comes to providing responses according to the expected type.

Length. As in the case of native/non-native speaker study we decided to check the length of responses provided in both groups. The length of the responses for the groups is presented in Table 8 and Figure 9.

Response type	Generated (A) ^{a)}	Correct (A)	(% corr)
DA	25	20	80%
IA	25	13	52%
EAP	25	7	28%
EAI	25	9	36%
All	100	49	49%

Table 6:
Summary of responses' correctness with respect to categories for group A (textual)

^{a)} Subset of the whole sample.

Table 7:
Summary of responses' correctness with respect to categories for group B (graphical)

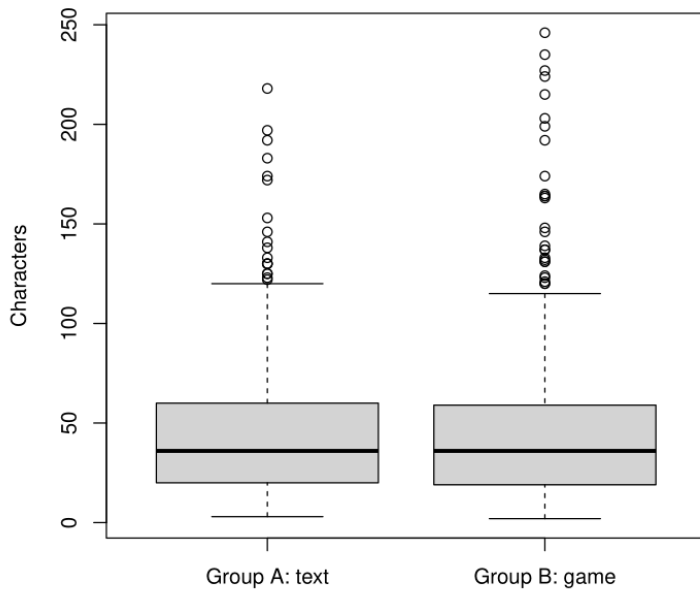
Response type	Generated (B) ^{b)}	Correct (B)	(% corr)
DA	25	23	92%
IA	25	8	32%
EAP	25	10	40%
EAI	25	18	72%
All	100	59	59%

^{b)} Subset of the whole sample.

Table 8:
Length of responses (number of characters) generated in QRGs

Group	Mean	SD	Median	Min	Max
A	45.10	33.91	36.00	3	218
B	44.98	37.53	36.00	2	246

Figure 9:
Comparison of the number of characters used in responses generated by groups A (textual QRGs) and B (graphical QRGs)



The Wilcoxon Test shows that there are no statistically significant differences between the groups when it comes to the length of the responses ($W = 160198$, $p = 0.5302$).

Engagement. The Cronbach alpha of IMUW for this study was 0.86 for group A and 0.83 for group B. Hence, we can confirm that

Group	Mean	SD	Median	Min	Max
A	37.17	7.39	38	20	50
B	31.80	6.68	32	20	45

Table 9: IMUW results (declared engagement in the task) for groups A (textual QRGS) and B (graphical QRGS)

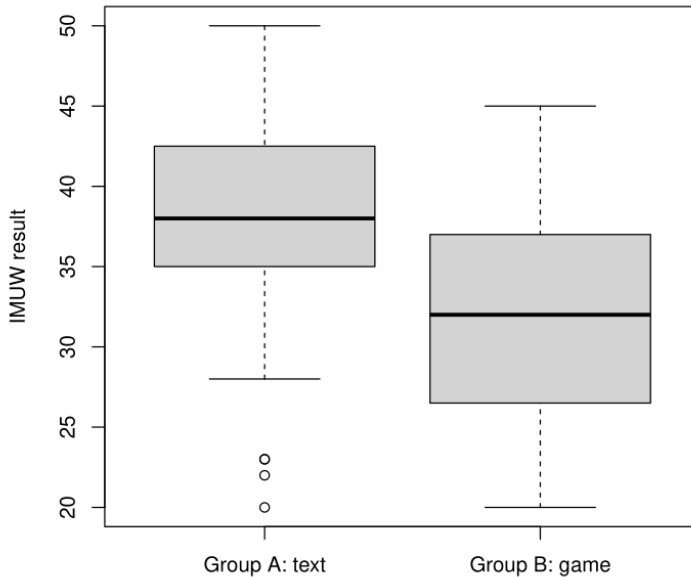


Figure 10: Comparison of IMUW results (declared engagement in the task) for groups A (textual QRGS) and B (graphical QRGS)

the reliability of the tool used was high. IMUW results are presented in Table 9 and Figure 10.

The Wilcoxon Test shows that the difference between group A and group B is statistically significant ($W = 874.5, p = 0.002106$). We may conclude that the textual version of QRGS was more engaging for our participants than the game-like, graphical one.

This study indicates that a step towards more game-like solutions for QRGS is not necessary. In terms of correctness of the gathered data and response length, we do not observe any apparent differences between groups. What is interesting is the result related to the self-reported engagement into the task. Smaller engagement for the graphical version of QRGS may suggest more distracting factors exist for this version. This may be also the result of the fact that in this version the story unfolds more slowly and the whole task takes more time.

Definite answers on these issues require further investigation. However, on the basis of the results obtained already we can say that the textual version of QRGs is still a good option to be used – especially due to the simplicity of the design and implementation.

6 PLOT FORMULATION

In the following study, our aim was to test QRGs in yet another respect. Namely, whether the style of the plot of the story used matters for correctness. The intuition behind this question is that certain types of stories may be more immersive or more appealing to users and thus result in more correct responses being generated. That is why we decided to design two new QRGs stories in Polish, one of which is a detective story in which a participant is lured into the crime-solving plot. The second one is more neutral as it concerns organisation of an engagement surprise party.

6.1 *Materials*

For the sake of the study two QRGs stories were prepared: “Jewellery theft” and “Engagement”. The first one tells the story of a bold theft of old jewellery from a nobleman’s home. A participant takes part in the interrogations to find the culprit. Thus, the questions to be responded to in QRGs concern the following: what did the thief look like? What did he use to carry the stolen goods? How did the thief manage to escape the home? What time did the theft take place? The second story concerns an engagement surprise party. A participant plays the role of a friend asked to book the restaurant. Questions to be responded to cover the time of the party, number of people to be invited or planned surprises. Stories and their scenarios are presented in Appendix C.

6.2 *Procedure*

The study was conducted online using Google Forms. Two separate forms were prepared for the two stories. Participants were invited to take part in the study *via* a link on the social media platforms. The link

led to the page where a participant was randomly assigned to one of the stories. Participants received the information about the study and provided their agreement to take part. After that, they were presented with the story followed by four scenarios with questions. At the end, they provided basic demographic data.

Study group

6.3

Overall, 199 participants took part in the study. The “Engagement” story form (group A) was filled out by 101 participants, including 90 women and 11 men. The participants were between 17 and 45 years, and the mean was 21.84 years (SD = 5.13). The version with the story of jewellery theft (group B) was filled out by 98 people, of which 88 of individuals were women and 10 are men. The age of the participants ranged from 17 to 45 years with an average of 22.48 years (SD = 5.53).

Results and data validation

6.4

Data was gathered from May 2019 till January 2020. We collected 3,184 responses: 1,616 responses to the first story and 1,568 to the second one. The variety of responses was satisfactory, as observed for previous QRGS studies.

Correctness. The correctness check covered the whole gathered sample. We used the same procedure as in previous studies. Fleiss’s kappa (for three annotators) was 0.502 for group A and 0.527 for group B. As for the percentage of correct answers, we got 54% for group A and 57% for group B – see the details in Tables 10 and 11. Thus, when it comes to correctness the plot formulations are very similar.

Length. As in the case of previous studies, we decided to check the length of responses provided in both groups. The length of responses for the groups is presented in Table 12 and Figure 11.

The Wilcoxon Test shows a statistically significant difference between groups ($W = 1525410$, $p < 0.001$). Responses gathered for the detective-like story were significantly longer than the ones for the story about the surprise party.

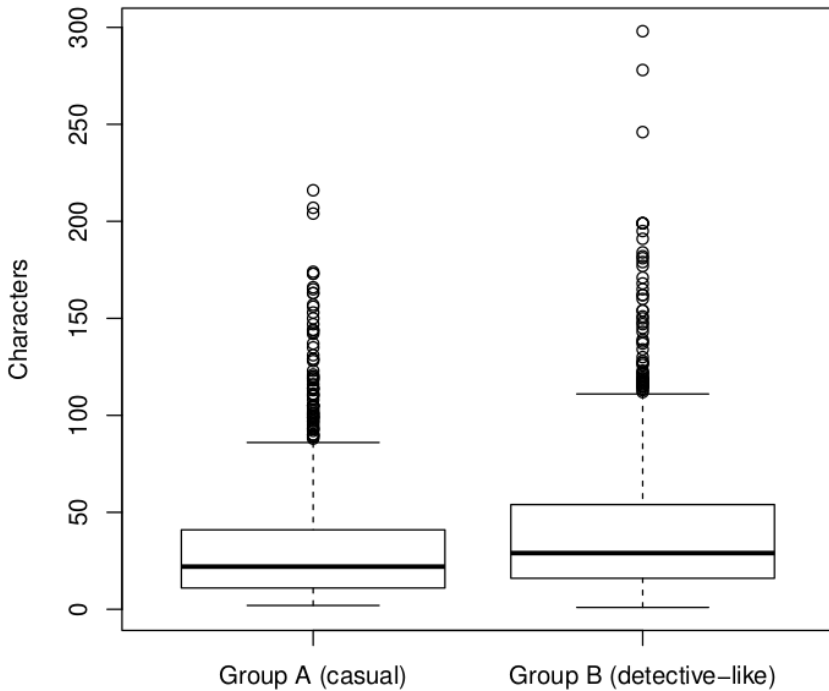
Table 10:
Summary
of responses'
correctness
with respect
to categories
for group A
(*Kradzież
biżuterii*)

Response type	Generated (A)	Correct (A)	(% corr)
DA	392	388	99%
IA	392	78	20%
EAP	392	225	57%
EAI	392	154	39%
All	1,568	845	54%

Table 11:
Summary
of responses'
correctness
with respect
to categories
for group B
(*Zaręczyny*)

Response type	Generated (B)	Correct (B)	(% corr)
DA	404	410	99%
IA	404	144	36%
EAP	404	199	49%
EAI	404	186	46%
All	1,616	930	57%

Figure 11:
Length
of responses
(number
of characters)
generated
in QRGS



Group	Mean	SD	Median	Min	Max
A	30.09	28.65	22	2	216
B	39.65	33.75	29	1	298

Table 12:
Length of responses
(number of characters)
generated in QRGS

The results indicate that no apparent differences are observed between two different story topics when it comes to the correctness factor. The difference is observed in terms of length. For the detective-like stories, provided responses were longer. In consequence we may conclude that detective-like stories are recommended for QRGS if one wishes to obtain longer responses. These also seem to be easier to plan and write. However, as the differences in correctness are very small, the choice of topics to be used for QRGS stories is open.

QRGS EVALUATION MODULE

7

The studies described in previous sections revealed a potential weakness of the QRGS framework, namely the need for a manual data check (after they have been gathered). The process of data gathering needs no supervision. Data correctness is also satisfactory (especially for selected types of responses, like DA). However when one thinks about the potential use of the QRGS for supplementing carefully collected corpus data, it certainly requires additional control.

To deal with this issue, we decided to design, implement and check the evaluation module for QRGS which also uses a crowdsourcing mechanism. In this scenario, a user’s task is not only to generate new responses but also to evaluate selected responses previously provided by other players. As the generation phase is not demanding and is rather short, we believe that adding the evaluation phase to QRGS would not be troublesome for a user.

Two evaluation module designs

7.1

For the evaluation module, we assume that a user has previously read the story and generated responses to provide questions. In the evaluation module, the user is asked to perform only a simple matching

task. The task has two versions: (EM-A) matching story characters to responses, and (EM-B) matching one of the response categories to a provided response. After completing the task, a user is asked to assess how certain s(he) is about the solution proposed (the higher the assessment, the more certain the user is). This is heavily inspired by the Wordrobe (Venhuizen *et al.* 2013) – in this system it was observed that it results in better user performance. For future QRGs applications EM-A or EM-B may be used separately or together (to add more task diversity).

EM-A. A user is presented with the instruction that s(he) should match four characters to the four responses given and for each choice declare how certain s(he) is about the match. Each character has a short description of the type of response it provides (according to the story plot) – see Figure 12.

EM-B. A user is presented with a question/response pair (in a form known from chat applications) below which the one-choice list of response types is presented. The user chooses one of the answers and declares how certain (s)he is about it – see Figure 13.

The results of testing both described designs are presented below.

7.2

Evaluation module test

The evaluation system was designed and tested in Polish. For the test, we used the “Tavern” story and the data gathered and checked in the study described in Section 5.





For the EM-A (see Figure 12) the following instruction was provided to the user.

Poniżej są cztery odpowiedzi na pytanie “Co niepokojącego ostatnio wydarzyło się w karczmie?”. Każdą z nich przyporządkuj do postaci i typu odpowiedzi. Tylko jedna odpowiedź pasuje do każdej z postaci, nie będą się powtarzać. Zaznacz też, jaką pewność, że Twoja odpowiedź jest poprawna. Odpowiedz w skali od 1 do 3. Im wyższa liczba, tym większa pewność.

Below you will find four responses to the question: “What was the worrying thing that happened at the tavern?” Take each response and match them with the characters from the story and

Połącz odpowiedź z postacią

Poniżej są cztery odpowiedzi na pytanie "Co niepokojącego ostatnio wydarzyło się w karczmie?". Każdą z nich przyporządkuj do postaci i typu odpowiedzi. Tylko jedna odpowiedź pasuje do każdej z postaci, nie będą się powtarzać. Zaznacz też, jaką pewność, że Twoja odpowiedź jest poprawna. Odpowiedz w skali od 1 do 3. Im wyższa liczba, tym większa pewność.


Straż grodowa	Wspólnik	Pacholek	Żona
			
Twoja odpowiedź była bezpośrednia i zgodna z prawdą	Twoja odpowiedź była wymijająca, ale uprzejma.	Twoja odpowiedź była wymijająca i niezbyt uprzejma.	Twoja odpowiedź była prawdziwa, ale nie bezpośrednia.

Sam sprzątałeś przed chwilą wychodek, to sam sobie odpowiedz na to pytanie. Donieśłeś w końcu te beczi z piwem?


Ach, wiesz jak zwykle to bywa w tawernach – trochę szalonych ludzi, trochę pijanych, lecz raczej nic groźnego. Zresztą, przecież pracujesz tu nie od wczoraj.

W karczmie pojawił się tajemniczy mężczyzna, który chciał za zapłatą zostawić mi do przechowania pakunek z nieznaną zawartością, który miałbym następnie przekazać dalej.


Był jeden typ, ale nie obawiaj się.


Pacholek


Twoja pewność - 3 +


Wybierz postać

Twoja pewność - 1 +


Wybierz postać

Twoja pewność - 1 +


Wybierz postać

Twoja pewność - 1 +

Figure 12: Evaluation system A – description in the text

Figure 13:
Evaluation
system B –
description in
the text

Jaka jest ta odpowiedź?

Był tu ostatnio podejrzany człowiek... Jak on wyglądał?

Był wysoki, miał długą ciemną szatę i kaptur na głowie. Miał dość bladą twarz i długie rude włosy.

Ta odpowiedź jest:

- Bezpośrednia i zgodna z prawdą
- Wymijająca, ale uprzejma
- Wymijająca i nieuprzejma
- Pośrednia i zgodna z prawdą

Na ile masz pewność, że odpowiedź trafi do dobrej kategorii?
Odpowiedz w skali od 1 do 3. Im wyższa liczba, tym większa pewność. - 1 +

the answer type. Each character can be matched with only one response. Choose how certain you are in regard to the correctness of your answer. Answer on the scale of: 1 to 3. The more certain you are, the higher your answer should be.

Four matching tasks were prepared for the EM-A. The data retrieved from the study described in Section 5 used for this module is presented in Table 13. For each question, a user is presented with four different responses. Each time, the order of response types is different. Table 13 presents this order.

For the second annotation module design, EM-B, the instruction was simply: What kind of response is that? The data used for the EM-B is presented in Table 14. Analogously to design A, here four tasks were prepared.

7.3

Procedure

The user study was conducted with the use of a dedicated website, and the answers were gathered online. Before starting the study, the users

Table 13: The data used for the first evaluation module EM-A

The original version	English translation
<p>Q1: Co niepokojącego ostatnio wydarzyło się w karczmie?</p> <p>(DA) Odwiedził mnie tajemniczy człowiek.</p> <p>(EAI) Nic. Przecież cały dzień tu siedzisz, to ciebie powinniśmy zapytać.</p> <p>(EAP) Jak to w karczmie, codziennie jakieś przygody.</p> <p>(IA) Nic, czym z czym już sobie nie poradziłem, był tu taki jeden</p>	<p>Q1: What was the worrying thing that happened at the tavern?</p> <p>I was visited by a mysterious man.</p> <p>Nothing. You sit here all day, we should be asking you.</p> <p>As usual in the tavern, new adventures every day.</p> <p>Nothing I couldn't deal with, some guy stopped by</p>
<p>Q2: Był tu ostatnio podejrzaný człowiek. Jak on wyglądał?</p> <p>(EAI) a czy Ja muszę wszystkich pamiętać.</p> <p>(EAP) Był typowym wędrowcem, niczym się nie wyróżniał</p> <p>(IA) Trochę jak Twój kuzyn, Edmund, tylko wyższy.</p> <p>(DA) Nie widziałem wiele z powodu kaptura ale miał dość bladą cerę i rude włosy oraz był bardzo wysoki.</p>	<p>Q2: A suspicious man came by recently. What did he look like?</p> <p>is it Me who has to remember everyone.</p> <p>He was a typical vagabond, there was nothing special about him</p> <p>A little like your cousin, Edmund, just taller</p> <p>I couldn't see much because of his hood but he was quite pale, red haired, and very tall.</p>
<p>Q3: Co od niego dostałeś i jak to wyglądało?</p> <p>(EAP) Zamknięta, nie wiadomo co w środku, ale to taka przysługa tylko, powinniśmy być mili dla klientów jeśli chcemy mieć większy utarg.</p> <p>(EAI) Nie wiem o co ci chodzi. Zajmij się swoją pracą</p> <p>(DA) Dużą paczkę przewiązana lnianym sznurem</p> <p>(IA) Wyglądało jak pościel pod łóżkiem w naszym pokoju</p>	<p>Q3: What did you get from him and how did it look?</p> <p>It was closed, hard to tell what was inside, but it was just a favour we should be nice to clients if we want to have a bigger turnover</p> <p>I don't know what you're talking about. Get back to work</p> <p>A big package with a linen ribbon</p> <p>It looked like the sheets we keep under the bed in our room</p>
<p>Q4: Co ci za to zaferował?</p> <p>(DA) 3 dukaty.</p> <p>(IA) tak ze sześć razy tyle, co nasze całe wesele nas wyszło, a skromne to było wesele, a skromne (mruga okiem).</p> <p>(EAI) Co mi zaferował, to mi zaferował.</p> <p>(EAP) Nie wspominał konkretnie</p>	<p>Q4: What did he offer you?</p> <p>3 ducats.</p> <p>about six times as much as we paid for our wedding reception, and it was a modest one, definitely modest (winks).</p> <p>What he offered me, he offered me.</p> <p>He didn't mention anything specific</p>

Table 14: Q-R pairs for the second annotation mode EM-B

The original version	English translation
Q1: Co niepokojącego ostatnio wydarzyło się w karczmie? Pojawili się kilku podejrzanych typów, ale to nic szczególnego (EAP)	Q1: What was the worrying thing that happened at the tavern? A few suspicious guys came here, but it's nothing out of the ordinary.
Q2: Był tu ostatnio podejrzany człowiek. Jak on wyglądał? Trochę jak Twój kuzyn, Edmund, tylko wyższy. (IA)	Q2: A suspicious man came by recently. What did he look like? A little like your cousin, Edmund, just taller
Q3: Co od niego dostałeś i jak to wyglądało? Po co ci te wszystkie informacje? Szpiegujesz nas? (EAI)	Q3: What did you get from him and how did it look? Why do you need all this information? Are you spying on us?
Q4: Co ci za to zaoferował? 3 złote dukaty (DA)	Q4: What did he offer you? 3 golden ducats

could read information about it, then they had to agree to take part. Having done that, users were presented with a series of eight tasks in the two evaluation modes. The structure of the study was the following. First, the introduction and instructions were displayed. Once the user had expressed their agreement, the story “The Tavern” was introduced. This was followed by four tasks in EM-A and afterward by four tasks in EM-B. At the end, users provided elementary demographic data.

7.4

Study group

32 participants took part in the study, aged 29–70 (mean = 28.03, SD = 11.27). 26 participants were female, 5 were male.

7.5

Results

The results were gathered on April 4–8, 2021. User solutions were compared with the predefined answers (see Tables 15 and 16) to establish the evaluation correctness.

Question / Correct response	Correctness (%)	Average certainty	SD for average certainty
Q1 / DA	72	2.47	0.80
Q1 / EAI	72	2.50	0.76
Q1 / EAP	63	2.34	0.83
Q1 / IA	41	2.09	0.86
Q2 / EAI	81	2.53	0.76
Q2 / EAP	75	2.16	0.85
Q2 / IA	78	2.03	0.86
Q2 / DA	90	2.44	0.84
Q3 / EAP	81	2.28	0.81
Q3 / EAI	78	2.44	0.84
Q3 / DA	81	2.41	0.87
Q3 / IA	90	2.50	0.80
Q4 / DA	81	2.56	0.80
Q4 / IA	94	2.53	0.72
Q4 / EAI	81	2.34	0.90
Q4 / EAP	72	2.19	0.93

Table 15:
The correctness of responses provided by users of the evaluation module EM-A

For the evaluation module EM-A users were requested to perform 16 matchings of responses to characters who would provide these responses (i.e. to one of four response types). The lowest correctness in this task was IA response in the first question (41%) and the highest was also for IA but for the fourth question (94%). Detailed results are presented in Table 15.

A closer look at the data correctness presented in Table 15 suggests that a form of training example or a training session would be needed for the evaluation module. The somewhat surprising lowest and highest correctness percentage for IA may be better understood in light of an overall low correctness percentage for the first question presented to users.

As for EM-B, the correctness of users' identification of responses types is presented in Table 16. Here, we also observe that the highest correctness level was observed for the IA responses. This needs further exploration as it indicates interesting user behaviour – IA is the most

Table 16:
The correctness
of responses
provided
by users
of the evaluation
module EM-B

Question / Correct response	Correctness (%)	Average certainty	SD for average certainty
Q1 / EAP	97	2.53	0.84
Q1 / IA	100	2.62	0.75
Q1 / EAI	72	2.00	0.72
Q1 / DA	67	2.03	0.86

difficult response type to generate but the easiest to identify (as this preliminary data suggest).

The overall inter-annotator ($N=32$) agreement established with the Fleiss Kappa measure (with the use of R programming language and irr library) was slightly higher for EM-B (0.666) than for the EM-A (0.503).

We believe that the correctness of evaluations provided by the users is satisfactory. The proposed designs naturally need further study. The correctness may be further improved by implementing training mechanisms known from the scientific discovery games, like the aforementioned Galaxy Zoo (Lintott *et al.* 2008). Training examples should be presented before the target responses and provide instant feedback for the user.

The evaluation module design presented in this section offer a promising addition to QRGS. It is worth stressing that a researcher may still rely completely on the manual check of the data performed by expert(s). We can imagine different QRGS usage scenarios which depend on the main purpose of the gathered linguistic data.

8 QRGS DATA AS A PART OF THE EROTETIC REASONING CORPUS

We decided to publish part of the data gathered during our QRGS evaluation studies. As a platform to do this, we decided to use the Erotetic Reasoning Corpus (ERC; Łupkowski *et al.* 2017).

ERC is a data set for research on natural question processing. The basic intuition is that we are dealing with question processing in a situation when a question is not followed by an answer but with a new

question or a strategy of reducing it into auxiliary questions. Usually, such a situation takes place when an agent wants to solve a certain problem (expressed in a form of an initial question) but is not able to reach a solution using his/her own information resources. Thus, new data, collected *via* questioning is necessary. The corpus consists of the language data collected in previous studies on the question processing phenomenon. The outcomes of three research projects are employed here. These are: Erotetic Reasoning Test (Urbański *et al.* 2016a), Quest-Gen (Ignaszak and Łupkowski 2017) and Mind Maze (Urbański *et al.* 2016b). All the data are in Polish, but the tagging schema is in English to make it more universal to use.

The tagging schema for the ERC has three layers:

1. *Structural* – representing the structure of tasks used for the aforementioned studies. Here we distinguish elements like: instructions, justifications, different types of questions and declaratives.
2. *Inferential* – which allows for recognising normative elements related to the logic of questions used.
3. *Pragmatic* – representing various events that may occur in the dialogue, like e.g. long pauses. It also contains tags that allow expression of certain events related to the types of tasks used (like e.g. when a forbidden question is used).

QRGS data preparation

8.1

The data to be added to ERC were retrieved in the study described in Section 5 (the study in Polish checking textual vs graphical QRGS version).

The data generated by the first 20 participants was used. Each participant provided responses to all four scenarios of “The Tavern” story (see the whole story in Appendix B):

1. Guard: DA
2. Business partner: EAP
3. Minion: EAI
4. Wife: IA.

Each solution was saved into a separate file. Overall, we have 80 files (20 per scenario), with 17,426 words. Each file started with “The Tavern” story followed with the paragraph introduction a scenario. Then, we have questions and user-generated responses formatted in a dialogue-like fashion.

These 80 files were manually annotated with the appropriately modified and extended ERC tagset.

8.2

ERC tagset extensions and modifications

When it comes to the structural layer, the QUESTION tag has been extended with the AQ-ANSWER to cover cases where a question is responded with a question.

An example is presented below:

P: Co od niego dostałeś i jak to wyglądało? / *What did you get from him and what did it look like?*

K: Kogo masz na myśli? / *Who do you mean?*

Pachołek: <QUESTION A1="AUXILIARY" A3="OTHER" A4="3">Co od niego dostałeś i jak to wyglądało?</QUESTION>

Karczmarz: <QUESTION A1="AQ-ANSWER" A4="3"><EAI>Kogo masz na myśli?</EAI></QUESTION>

Query-response “Kogo masz na myśli / *Who do you mean?*” is identified with the tag AQ-ANSWER. The attribute of A4 links question and response in a given data file. (Arguments of A4 are the consecutive numbers of question-response pairs in a given file, see Figure 14.)

The pragmatic layer received one extension and one new tag, which is required to address the type of data from QRGs. The already existing tag KEY-INFO was extended with the attributes characteristic to the story, i.e. *character*, *package* and *payment*. This allows for identification of key information appearing in the story and user-generated responses.

S: Co od niego dostałeś i jak to wyglądało? / *What did you get from him and how did it look?*

K: Czarny pakunek, szczelnie zamknięty. / *Black package, it was tightly wrapped.*


```
Pachołek: <QUESTION A1="AUXILIARY" A3="OTHER" A4="3">Co od niego
dostałeś i jak to wyglądało?</QUESTION>
Karczmarz: <DECLARATIVE A1="AQ-ANSWER" A4="3"><DA><KEY-INFO A1="package"
A4="2">Czarny pakunek, szczelnie zamknięty.</KEY-INFO></DA></DECLARATIVE>
```

The response: “Czarny pakunek, szczelnie zamknięty. / *Black package, it was tightly wrapped*” is identified as a declarative answer to the question above and also as a key-info from the point of view of the story plot.

Another additional tag is RRT (required response type) with the attributes related to four response types generated by QRGS users: DA, IA, EAP, EAI and OTHER (for possible responses not fitting the expected categories).

80 QRGS files were annotated by two annotators with the updated ERC tagset. A sample annotated file is presented in Figure 14.

The annotated files were all checked in accordance with the procedure for the ERC described in Łupkowski *et al.* 2017. Firstly, all files were checked for the syntactic correctness of the XML tags with the Emacs editor (version 26.3) and Vacuous XML schema². All identified errors were eliminated. In the next step, 50 files were chosen randomly and intra- and inter-annotator studies were performed. Kappa values were established with the use of the R programming language (version 3.5.0) and irr package. Results were satisfactory, as Cohen’s kappa for the intra-annotator study was 0.819 (with 84% agreement) and for the inter-annotation study was 0.791 (with 82% agreement).

The annotated QRGS data are now available as a part of the Erotetic Reasoning Corpus.³

²<https://www.w3.org/TR/xmlschema11-1/>.

³Erotetic Reasoning Corpus homepage is: <https://ercorpus.wordpress.com/>. The latest version of ERC is available there along with documentation describing the tag-set used, and ERC tools: Search & Browse Tool (for browsing ERC files with and without annotation visible, as well as searching for particular ERC tags); XML/L^AT_EX Parser (easy transformation of XML files into L^AT_EX files); and ERC XML Schema (which allows for validating the annotation of ERC files).

```

1 <KORPUS A1="QRGS" A2="straz12">
2
3 <INSTRUCTION>
4 Historia pewnej karczmy: Straż grodowa
5 Jesteś właścicielem tawerny na obrzeżach miasta. Wczoraj zjawił się u ciebie <
KEY-INFO A1="character" A4="1"> tajemniczy podróżny, odziany w szeroki, ciemny
płaszcz z kapturem, który prawie całkiem zasłaniał jego twarz. Był bardzo wysoki
i dało się zaobserwować niezwykle bladą twarz i wystające spod kaptura długie,
rude włosy.</KEY-INFO> W rękę trzymał <KEY-INFO A1="package" A4="2">niemały
pakunek, przewiązany lnianą wstęgą.</KEY-INFO> Położył go na kontuarze i
zapropnował ci układ. W zamian za <KEY-INFO A1="payment" A4="3">sowitą zapłatę
w postaci trzech dukatów</KEY-INFO>, tajemnicza przesyłka zostaje pod twoim
nadzorem. Ma zostać nietknięta, do czasu kiedy zjawi się człowiek, który będzie
znał ustalone hasło.
6
7 Następnego dnia do twojej tawerny wkracza straż grodowa. Wygląda na to, że dziś
nie przyszli na ciepły posiłek po służbie. Chcą zadać ci kilka pytań.
8 <RRT A1="DA">Chociaż pytają o tajemniczego wędrowca, uznajesz że mądrze będzie
odpowiadać im bezpośrednio i zgodnie z prawdą. Nie chcesz przecież popaść w
konflikt z władzą.</RRT>
9 </INSTRUCTION>
10
11 Straż: Witaj karczmarzu! Dziś przybывamy w sprawie służbowej. Mamy kilka pytań.
12
13 Straż: <QUESTION A1="AUXILIARY" A3="OTHER" A4="1">Co niepokojącego ostatnio
wydarzyło się w karczmie?</QUESTION>
14
15 Karczmarz: <DECLARATIVE A1="AQ-ANSWER" A4="1"><EAP>Nic szczególnego. Kilku
podpitych gości wszczęło bójki, ale to w zasadzie norma. </EAP></DECLARATIVE>
16
17 Straż: <QUESTION A1="AUXILIARY" A3="OTHER" A4="2">Był tu ostatnio podejrzany
człowiek. Jak on wyglądał?</QUESTION>
18
19 Karczmarz: <QUESTION A1="AQ-ANSWER" A4="2"><DA>Podejrzany człowiek?</DA></QUESTION>
<DECLARATIVE A1="AQ-ANSWER" A4="2"><DA> A no był taki jakiś dziwak.<KEY-INFO A1="
character" A4="1"> Chudy, blady, rude włosy. Nie widziałem twarzy, bo zakrywał ją
kaptur.</KEY-INFO></DA></DECLARATIVE>
20
21 Straż: <QUESTION A1="AUXILIARY" A3="OTHER" A4="3">Co od niego dostałeś i jak to
wyglądało?</QUESTION>
22
23 Karczmarz: <DECLARATIVE A1="AQ-ANSWER" A4="3"><DA><KEY-INFO A1="package" A4="2">
Dał mi jakiś spory pakunek na przechowanie dla kogoś innego.</KEY-INFO></DA></
DECLARATIVE>
24
25 Straż: <QUESTION A1="AUXILIARY" A3="OTHER" A4="4">Co ci za to zaoferował?</QUESTION
>
26
27 Karczmarz: <DECLARATIVE A1="AQ-ANSWER" A4="4"><DA><KEY-INFO A1="payment" A4="3">
Dał mi trzy dukaty. A to sporo za jakiś tam pakunek.</KEY-INFO></DA></DECLARATIVE>
28
29 </KORPUS>

```

Figure 14: An exemplary QRGS file annotated with the ERC tagset

This paper presents the concept of the Question Responses Generation System, a crowdsourced framework for gathering linguistic data of a specific form. QRGS allows for relatively simple and efficient retrieval of various responses to questions.

QRGS requires a simple story and follow up scenarios to the main plot which lead a user to provide responses of the required type. As such, it is a very universal framework. The stories are relatively simple and easy to write. The whole schema of the framework is also simple and – crucially – easy to implement. One does not need any special programming skills. As presented in the paper, even Google Forms (or any other similar platform) is enough to implement QRGS and gather data.

We presented a series of evaluation studies of QRGS. Seven stories in total were tested so far (and are available as appendices for this paper). Four of them are in Polish, three in English.

During our evaluation studies QRGS appeared to be effective in terms of the amount of data gathered.⁴ Altogether, 4,304 responses to questions have been generated for Polish and 296 Q-R pairs have been generated for English. Also the correctness of the data is satisfactory, as summarised in Table 17. Correctness is understood here as compliance with the type of the response expected from a given story scenario. Our findings indicate that the most unproblematic response type in an unsupervised crowdsourced data generation are direct answers (DA). The most difficult for QRGS users are indirect answers (IA).

As the reported results show, the correctness level varies between studies and does not reach 100%. This indicates that the data gathered *via* QRGS cannot be straightforwardly used for certain applications, e.g., training data for language models. Such data needs to be evaluated first. That is why we also propose a promising and effective crowdsourcing solution that allows for data evaluation. Using one

⁴However, as pointed by the anonymous reviewer, the amount of data gathered may be dependent on many parameters, not only the framework supporting the acquisition, such as: availability of participants or the interval of time allocated for the crowdsourcing activity.

Table 17:
The summary
of the
correctness
of the data
gathered with
the use of QRGS

Study	Correctness (%)	
Pilot (Section 3)		77
Native vs non-native (Section 4)	63	62
Textual vs graphical (Section 5)	49	59
Casual vs detective-like story (Section 6)	54	57

(or two) proposed evaluation modules for additional data correctness checks. Naturally, an expert manual check of the data is still possible (and recommended for certain future applications). After the evaluation phase we envisage two potential scenarios: 1) eliminating non-correct responses (as the relative cost of generating data with QRGS is not high, we see this as an acceptable option); 2) reusing non-correct responses for which correct labels are added during the evaluation stage (this leaves a researcher with the complete generated dataset).

In line with scenario 2, part of QRGS generated data was formatted, manually annotated, thoroughly checked and incorporated into the Erotetic Reasoning Corpus and is now publicly available.

The series of QRGS studies resulted also in several findings useful for future QRGS development and implementations.

1. No difference between native and non-native English speakers for correctness and the response length were observed. At least for English, we may expect valuable data as long as we gather users with good knowledge of the language. Naturally, this observation needs further study for other languages.
2. Very small differences were observed *vis a vis* correctness for the graphical vs text and casual vs detective-like stories; similarly, no difference between the text condition and the graphical condition for the response length. This suggests that the simple, textual version is enough to effectively use QRGS.
3. Participants in the text condition were more engaged in the task (than in the graphical condition). This is an interesting and somewhat surprising finding suggesting (as in the case of 2) that the text-only version of QRGS may be a better solution.
4. Observed differences in the response length for the casual vs detective-like stories. This effect suggests that the detective-like stories may result in more extended responses. This needs further

study – especially quantitatively, where users’ experiences would be evaluated.

QRGS offers a promising framework for gathering large amounts of various types of responses to questions. We believe that it needs further testing with other languages, especially those which have lower spoken language corpora coverage. There are also open questions which may be addressed when it comes to the QRGS idea, e.g. how to increase the data correctness level, especially for IA? Or how to add more scenarios to the stories, such that more response categories would be generated (and the whole QRGS task would not get boring and time-consuming for users)?

ACKNOWLEDGEMENTS

This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the program “Investissements d’Avenir” (reference: ANR-10-LABX-0083). It contributes to the IdEx Université de Paris – ANR-18-IDEX-0001. Work on the Erotetic Reasoning Corpus was supported by the National Science Centre, Poland (DEC-2013/10/E/HS1/00172 and DEC-2012/04/A/HS1/00715).

We would also like to give our thanks to three anonymous reviewers for the Journal of Language Modelling for their insightful comments on this article.

APPENDICES

A STORIES FOR THE NATIVE / NON-NATIVE STUDY

Story I. EPILEPSY. Your co-worker Anna just had an epilepsy attack. You are aware this happens sometimes, as for the safety reasons she has informed you some time ago. You also know it has been 6 months since her last seizure event. Today was just an ordinary day and nothing uncommon preceded the attack. When she lost consciousness and fell to the floor you were standing next to her. You have assisted Anna making sure she does not hurt herself during the convulsions. You measured the length of the attack – it took about 5 minutes. After that, she did not regain consciousness, so you have decided to call for emergency.

Scenario A. The paramedic has arrived and is asking you some questions about Anna and you want your answers to be very accurate:

1. How long did the seizure last?
2. Was she conscious during the attack?
3. Did anything unusual happen before the accident?
4. How many attacks did she have lately?

Scenario B. Anna's mother is calling you because she could not reach her daughter. You told her about the attack and now she has more questions. Unfortunately, you are still in the office and you do not want people around you to overhear the details about Anna. As you cannot leave the common space and your colleagues suspect the topic of the conversation, you will need to answer indirectly:

1. How long did the seizure last?
2. Was she conscious during the attack?
3. Did anything unusual happen before the accident?
4. How many attacks did she have lately?

Scenario C. Matt from your team came by as he heard Anna had been taken to the hospital. He seems worried and is asking you questions

about Anna's condition. From what you know he and Anna are friends but Anna emphasised that she shared the information in secret, so you feel obliged to keep it. You understand his concerns but you are not going to reveal anything without permission:

1. How long did the seizure last?
2. Was she conscious during the attack?
3. Did anything unusual happen before the accident?
4. How many attacks did she have lately?

Scenario D. Rob, the annoying colleague from another department came by. He is known for his terrible gossiping habit and now is asking you questions about Anna's condition and information she told you in confidence. His behaviour irritates you and you do not want to talk with him about Anna. How will you react to his questions?:

1. How long did the seizure last?
2. Was she conscious during the attack?
3. Did anything unusual happen before the accident?
4. How many attacks did she have lately?

The paramedic team was able to rouse Anna and she seems all good but will be taken to the hospital for observation.

Story II. SECRET SANTA. You and your friends have decided to organise a Secret Santa event this year. Each member of your pack will receive a gift prepared jointly by the rest of the group members. After a short brainstorming session you proposed to give Joe the Craft Beer Brewing Kit and this idea was met with great enthusiasm. It costs 50 USD and this sum will be divided evenly between 5 people. You are responsible for collecting the money and purchasing the kit. You are going to make the purchase on Friday so your friends should give you their shares until then.

Scenario A. George (one of the conspiracy group) is not familiar with the arrangements and has just visited you for details. You can speak openly with George about the organisational details:

1. What are we giving to Joe?
2. How much is the contribution rate?

3. Who will make the purchase?
4. What is the deadline for collecting the money?

Scenario B. Jane was not present at the brainstorm meeting. She has called you and has some questions but Joe is in a car with you. You want to pass the information to Jane while hiding it from Joe. Try to provide indirect answers to the following questions:

1. What are we giving to Joe?
2. How much is the contribution rate?
3. Who will make the purchase?
4. What is the deadline for collecting the money?

Scenario C. Maggie, Joe's sister, is wishing to participate, too. You like her, but you do not trust her. She might share the secrets with Joe. She wants to know the details. Try to decline her in a polite manner:

1. What are we giving to Joe?
2. How much is the contribution rate?
3. Who will make the purchase?
4. What is the deadline for collecting the money?

Scenario D. Joe is extremely sneaky and is trying to draw some information on his gift from you. He has sent his younger brother to spy on you. You want to teach him a lesson of minding his own business and decline him in a rather rude way. How will you react to his questions?

1. What are we giving to Joe?
2. How much is the contribution rate?
3. Who will make the purchase?
4. What is the deadline for collecting the money?

B **STORY FOR THE GRAPHICAL / TEXTUAL
STUDY**

Story: TAWERNA / TAVERN. Jesteś właścicielem tawerny na obrzeżach miasta. Wczoraj zjawił się u ciebie tajemniczy podróżny, odziany

w szeroki, ciemny płaszcz z kapturem, który prawie całkiem zasłaniał jego twarz. Był bardzo wysoki i dało się zaobserwować niezwykle bladą twarz i wystające spod kaptura długie, rude włosy. W ręku trzymał niemały pakunek, przewiązany lnianą wstęgą. Położył go na koncie i zaproponował ci układ. W zamian za sowiłą zapłatę w postaci trzech dukatów, tajemnicza przesyłka zostaje pod twoim nadzorem. Ma zostać nietknięta, do czasu kiedy zjawi się człowiek, który będzie znał ustalone hasło.

You are the owner of a tavern in the suburbs. Yesterday a mysterious stranger came to you. He was wearing a wide, dark coat with a hood which covered almost all of his face. The stranger was very tall and he was extremely pale. You could observe long ginger hair under his hood. In his hand, he carried a significantly sized package with a linen ribbon. He put it on the counter and offered you a deal. He wanted to give the package to you for safekeeping and in turn he would pay you a fair price of 3 ducats. The package is to be left untouched until a man comes and tells you a password.

Scenario GUARD. Następnego dnia do twojej tawerny wkracza straż grodowa. Wygląda na to, że dziś nie przyszli na ciepły posiłek po służbie. Chcą zadać ci kilka pytań. Chociaż pytają o tajemniczego wędrowca, uznajesz że mądrze będzie odpowiadać im bezpośrednio i zgodnie z prawdą. Nie chcesz przecież popaść w konflikt z władzą. (Straż grodowa) Witaj karczmarzu! Dziś przybywamy w sprawie służbowej. Mamy kilka pytań.

On the next day, guards come to your tavern. It seems like they're not here to eat something warm after work. They want to ask you a few questions. Even though they're asking about the mysterious stranger, you decide that it will be wise to reply to them directly and truthfully. You don't want to get into a conflict with the guards. (Guards) Hello, innkeeper! Today we're here on business. We have a few questions to ask you.

Scenario BUSINESS PARTNER. Wieczorem zaczepia cię twój wspólnik, którego wczoraj nie było w gospodzie. Słyszał plotki od innych pracowników, dlatego postanawia wypytać cię o szczegóły. Odpowiedz mu wymijająco, ale uprzejmie – w końcu to twój wspólnik. (Wspólnik) Cześć chłopie! Dawno cię nie widziałem. Mam nadzieję, że wczorajszy obrót był wysoki. Muszę się przyznać, że słyszałem niepokojące plotki.

In the evening, your business partner comes up to you. He wasn't in the tavern yesterday. He's heard some gossip from other employees and he wants to know more details. Answer him in an evasive, but polite way, he's your business partner after all.

(Business partner) Hi, man! I haven't seen you in a while. I hope that yesterday's turnover was high. If I'm being honest, I've heard some unnerving rumours.

Scenario MINION. Tego samego wieczoru podchodzi do ciebie jeden ze sług zatrudnionych w karczmie. On również słyszał plotki. Sam dobrze wiesz, że te lubią rozchodzić się w zastraszającym tempie. Pacholek ma do ciebie parę pytań. Odpowiedz mu wymijająco – nie musisz być dla niego szczególnie uprzejmy. (Pacholek) Panie, wiem że ja tu tylko sprzątam, ale chciałbym cię o coś zapytać.

The same night one of the minions who work at your tavern comes to you. He's also heard the gossip. You know how fast they spread. The minion has a couple of questions for you. Answer him in an evasive way – you don't need to be polite. (Minion) Good sir, I know I'm a simple cleaner, but I would like to ask you about something.

Scenario WIFE. Kolejnego dnia z rana żona również bierze cię na wypytki. Ponieważ rozmowa toczy się przy kontuarze, przysłuchują się jej jak zawsze zaciekawieni goście gospody. Postaraj się udzielić żonie prawdziwych informacji, ale nie w bezpośredni sposób. (Żona) Witaj mężu. Mam nadzieję, że dobrze spałeś. Dopiero dziewiąta rano, a goście już pytają o zupę. Słyszałam od współnika niepokojące informacje – podobno odwiedziła nas straż grodowa. Mógłbyś mi rozjaśnić sprawę.

Next morning your wife wants to have a chat with you. The conversation is taking place at the counter, so as usual, curious tavern guests are listening out for information. Try to give your wife true information, but do it indirectly. (Wife) Hello, husband. I hope you slept well. It's only 9 in the morning and the guests are already asking for soup! Your business partner has given me some worrisome information – apparently the guards visited us. You could tell me what happened.

QUESTIONS

1. Co niepokojącego ostatnio wydarzyło się w karczmie? / *What was the worrying thing that happened at the tavern?*

2. Był tu ostatnio podejrzany człowiek. Jak on wyglądał? / *A suspicious man came by recently. What did he look like?*
3. Co od niego dostałeś i jak to wyglądało? / *What did you get from him and how did it look?*
4. Co ci za to zaoferował? / *What did he offer you?*

STORIES FOR THE PLOT FORMULATION STUDY

C

Story I. KRADZIEŻ BIŻUTERII / JEWELRY THEFT. Z domu szanowanego hrabiego ukradziono cenną, rodową biżuterię. Jako zaufany kucharz, tej nocy przygotowywałeś dla pana domu kolację i przypadkowo wpadłeś na złodzieja, któremu jednak udało się uciec. Zdążyłeś mu się przyjrzeć, ale niestety nie widziałeś jego twarzy. Mimo to, wiesz, że: Złodziej był wysokim, szczupłym mężczyzną w ciemnej kurtce z kapturem. Biżuterię wyniósł w pudełku na buty. Złodziej uciekł z domu przez tylne wyjście. Kradzież nastąpiła około godziny 20.

From the respected count's house, valuable ancestral jewellery was stolen. As a trusted chef, you were preparing dinner for the household that night and accidentally stumbled upon the thief, who managed to escape. You had a chance to observe them, but unfortunately did not see his face. Nonetheless, you know that: The thief was a tall, slim man wearing a dark jacket with a hood. He carried the jewelry out in a shoebox. The thief fled the house through the back exit. The theft occurred around 8 p.m.

Scenario SZEFA POLICJI / POLICE DIRECTOR. Na miejscu zjawia się szef policji, który próbuje ustalić szczegóły kradzieży. Powinieneś udzielić prawdziwych i precyzyjnych odpowiedzi na jego pytania.

When the chief of police arrives at the scene to investigate the details of the theft, you should provide true and precise answers to his questions.

Scenario OFIARA KRADZIEŻY / THEFT VICTIM. Zostałeś poinstruowany, aby tymczasowo nie dzielić się szczegółami śledztwa z panem domu, ze względu na jego słabe zdrowie. Ponieważ jednak hrabia

próbujecie wypytać Cię o zajście, postaraj się dać mu do zrozumienia, że nie możesz udzielić odpowiedzi, jednak zrób to w sposób uprzejmy (w końcu to Twój pracodawca).

You've been instructed not to share details of the investigation with the count temporarily, due to his poor health. However, since the count is attempting to inquire about the incident, try to politely indicate that you cannot provide an answer, keeping in mind that he is your employer.

Scenario SŁUŻBA / MINIONS. Zaraz po udaniu się hrabiego do sypialni, podchodzi do Ciebie kilka osób ze służby. Ze względu na trwające śledztwo nie możesz udzielić im bezpośrednio informacji, jednak postaraj się odpowiedzieć zgodnie z prawdą.

Right after the count goes to his bedroom, a few members of the household staff approach you. Due to the ongoing investigation, you cannot directly provide them with information, but try to answer truthfully.

Scenario SĄSIAD / NEIGHBOUR. Następnego dnia spotykasz sąsiada hrabiego, którego sylwetka, według Ciebie, łądząco przypomina złodzieja biżuterii (o czym wspomniałeś także policjantom). Odpowiedz na jego pytania w taki sposób, żeby zrozumiał, że nie chcesz z nim rozmawiać (nie musisz być bardzo uprzejmy, jednak nie powinieneś kłamać ani zbyć go słowami „nie wiem”, ponieważ nie możesz pozwolić, aby domyślił się, że jest podejrzanym).

The next day, you encounter the count's neighbour, whose silhouette, in your opinion, strikingly resembles that of the jewellery thief (which you also mentioned to the police). Answer his questions in a way that makes him understand you don't want to engage in conversation (you don't have to be overly polite, but you shouldn't lie or brush him off with "I don't know," as you cannot allow him to suspect he's a suspect).

QUESTIONS

1. Jak wyglądał złodziej? / *What did the thief look like?*
2. W czym udało mu się wynieść biżuterię? / *How did he manage to carry the jewellery?*
3. W jaki sposób uciekł z domu? / *How did he escape from the house?*
4. O której godzinie zdarzyła się kradzież? / *At what time did the theft occur?*

Story II. ZARĘCZYNY / ENGAGEMENT. Twój dobry przyjaciel Piotr poprosił Cię o pomoc w organizacji imprezy-niespodzianki, na której chce oświadczyć się swojej dziewczynie Ewie. Twoim zadaniem jest potwierdzenie rezerwacji w restauracji oraz zadbanie o zaproszenie zaufanych gości. Przyjaciel zostawił Ci kilka wskazówek: Impreza ma zacząć się o godzinie 17 i potrwa do północy. Zaproszonych będzie 15 osób, w tym rodzice Piotra i Ewy. Na imprezie będzie podawane ulubione wino pary. W torcie przygotowanym na imprezę zostanie ukryty pierścionek zaręczynowy.

Your good friend Piotr has asked you for help in organising a surprise party, where he plans to propose to his girlfriend Ewa. Your task is to confirm the restaurant reservation and ensure the invitation of trusted guests. Your friend left you some guidelines: The party is to start at 5 p.m. and last until midnight. 15 people will be invited, including Piotr and Ewa's parents. The couple's favorite wine will be served at the party. An engagement ring will be hidden in the cake prepared for the party.

Scenario SZEFE RESTAURACJI / RESTAURANT MANAGER. Na umówionym spotkaniu omawiasz szczegóły przyjęcia z szefem restauracji. Odpowiedz na jego pytania jak najdokładniej, aby wiedział, jak się przygotować na imprezę.

At the scheduled meeting, you discuss the details of the reception with the restaurant manager. Answer his questions as accurately as possible so he knows how to prepare for the event.

Scenario TELEFON OD MACIEJA / PHONE FROM MACIEJ. Po rozmowie z szefem restauracji jedziesz spotkać się z Ewą. Podczas Waszego spotkania dzwoni do Ciebie brat Piotra, Maciej, który wie o imprezie i chce dopytać Cię o szczegóły. Odpowiedz na jego pytania w zrozumiałym sposobie, ale tak, aby Ewa nie domyśliła się, o czym rozmawiacie.

After the conversation with the restaurant manager, you go to meet Ewa. During your meeting, Piotr's brother, Maciej, who knows about the party, calls you and wants to inquire about the details. Answer his questions in an understandable way, but ensure Ewa doesn't suspect what you're discussing.

Scenario EWA / EWA. Po zakończonej rozmowie z Maciejem, Ewa próbuje wypytać Cię o to, o czym rozmawialiście. Wie ona tylko, że im-

preza odbędzie się w najbliższą sobotę, jednak cała reszta powinna pozostać niespodzianką. Postaraj się odpowiedzieć na jej pytania tak, aby dać do zrozumienia, że nie możesz jej nic zdradzić, ale bądź uprzejmy, aby jej nie zdenerwować.

After the conversation with Maciej, Ewa tries to ask you about what you talked about. She only knows that the party will take place next Saturday, but everything else should remain a surprise. Try to answer her questions in a way that implies you can't reveal anything, but be polite so as not to upset her.

Scenario ZNAJOMA / FRIEND. Gdy wracasz do domu po spotkaniu, spotykasz na ulicy znajomą, za którą nie przepadają Ewa i Piotr. Nie jest ona zaproszona na imprezę, jednak usłyszała o niej od swojego kolegi. Odpowiedz na jej pytania tak, aby zrozumiała, że nie chcesz z nią rozmawiać (nie musisz być bardzo uprzejmy, jednak nie powinieneś kłamać ani zbyć jej słowami „nie wiem”).

When you return home after the meeting, you meet an acquaintance on the street, whom Ewa and Piotr don't particularly like. She's not invited to the party, but she heard about it from her friend. Answer her questions in a way that makes her understand you don't want to talk to her (you don't have to be very polite, but you shouldn't lie or brush her off with "I don't know").

QUESTIONS

1. W jakich godzinach odbędzie się impreza? / *What time will the party take place?*
2. Ile osób jest na nią zaproszonych? / *How many people are invited?*
3. Jaki alkohol zostanie podany na imprezie? / *What alcohol will be served at the party?*
4. Jakie są zaplanowane niespodzianki? / *What surprises are planned?*

REFERENCES

- Anne H. ANDERSON, Miles BADER, Ellen Gurman BARD, Elizabeth H. BOYLE, Gwyneth M. DOHERTY, Simon C. GARROD, Stephen D. ISARD, Jacqueline C. KOWTKO, Jan M. MCALLISTER, Jim MILLER, Catherine F. SOTILLO, Henry S. THOMPSON, and Regina WEINERT (1991), The HCRC Map Task Corpus, *Language and Speech*, 34(4):351–366.
- Lou BURNARD, editor (2007), *Reference guide for the British National Corpus (XML Edition)*, Oxford University Computing Services on behalf of the BNC Consortium, <http://www.natcorp.ox.ac.uk/XMLedition/URG/>, access 20.03.2017.
- Jon CHAMBERLAIN, Massimo POESIO, and Udo KRUSCHWITZ (2008), PhraseDetectives: A web-based collaborative annotation game, in *Proceedings of the International Conference on Semantic Systems (I-Semantics' 08)*, pp. 42–49.
- Seth COOPER, Adrien TREUILLE, Janos BARBERO, Andrew LEAVER-FAY, Kathleen TUIITE, Firas KHATIB, Alex Cho SNYDER, Michael BEENEN, David SALESIN, David BAKER, Zoran POPOVIĆ, and > 57,000 Foldit PLAYERS (2010), The challenge of designing scientific discovery games, in *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, pp. 40–47.
- Cristian DANESCU-NICULESCU-MIZIL and Lillian LEE (2011), Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs, in *Proceedings of the 2nd workshop on cognitive modeling and computational linguistics*, pp. 76–87, Association for Computational Linguistics.
- Lorna DSILVA, Shubhi MITTAL, Brian KOEPNICK, Jeff FLATTEN, Seth COOPER, and Scott HOROWITZ (2019), Creating custom foldit puzzles for teaching biochemistry, *Biochemistry and Molecular Biology Education*, 47(2):133–139.
- Dagmara DZIEDZIC (2016), Use of the free to play model in games with a purpose: the RoboCorp game case study, *Bio-Algorithms and Med-Systems*, 12(4):187–197.
- Jonathan GINZBURG, Zulipiye YUSUPUJIANG, Chuyuan LI, Kexin REN, Aleksandra KUCHARSKA, and Paweł LUPKOWSKI (2022), Characterizing the response space of questions: data and theory, *Dialogue & Discourse*, 13(2):79–132.
- Jonathan GINZBURG, Zulipiye YUSUPUJIANG, Chuyuan LI, Kexin REN, and Paweł LUPKOWSKI (2019), Characterizing the response space of questions: a corpus study for English and Polish, in *Proceedings of the 20th annual SIGdial meeting on discourse and dialogue*, pp. 320–330.

Oliwia IGNASZAK and Paweł ŁUPKOWSKI (2017), Inferential Erotetic Logic in modelling of cooperative problem solving involving questions in the QuestGen game, *Organon F*, 24(2):214–244.

Charlene JENNETT, Anna L. COX, Paul CAIRNS, Samira DHOPAREE, Andrew EPPS, Tim TIJS, and Alison WALTON (2008), Measuring and defining the experience of immersion in games, *International Journal of Human-Computer Studies*, 66(9):641–661.

Chris J. LINTOTT, Kevin SCHAWINSKI, Anže SLOSAR, Kate LAND, Steven BAMFORD, Daniel THOMAS, M. Jordan RADDICK, Robert C. NICHOL, Alex SZALAY, Dan ANDREESCU, Phil MURRAY, and Jan VANDENBERG (2008), Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey, *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189.

Paweł ŁUPKOWSKI and Dagmara DZIEDZIC (2016), Building players' engagement – a case study of games with a purpose in science, *Homo Ludens*, 1(9):127–145.

Paweł ŁUPKOWSKI, Mariusz URBAŃSKI, Andrzej WIŚNIEWSKI, Wojciech BŁĄDEK, Agata JUSKA, Anna KOSTRZEWA, Dominika PANKOW, Katarzyna PALUSZKIEWICZ, Oliwia IGNASZAK, Joanna URBAŃSKA, Natalia ŻYLUK, Andrzej GAJDA, and Bartosz MARCINIAK (2017), Erotetic Reasoning Corpus. A data set for research on natural question processing, *Journal of Language Modelling*, 5(3):607–631.

Paweł ŁUPKOWSKI and Patrycja WIETRZYCKA (2015), Gamification for question processing research – the QuestGen game, *Homo Ludens*, 7(1):161–171.

Adam PRZEPIÓRKOWSKI, Mirosław BAŃKO, Rafał L. GÓRSKI, Barbara LEWANDOWSKA-TOMASZCZYK, Marek ŁAZIŃSKI, and Piotr PĘZIK (2011), National Corpus of Polish, in *Proceedings of the 5th language & technology conference: Human language technologies as a challenge for computer science and linguistics*, pp. 259–263, Fundacja Uniwersytetu im. Adama Mickiewicza Poznań.

Piotr PĘZIK (2014), Spokes search engine for Polish conversational data, <http://hdl.handle.net/11321/47>, CLARIN-PL digital repository.

Carolyn P. ROSÉ, Barbara Di EUGENIO, and Johanna D. MOORE (1999), A dialogue-based tutoring system for basic electricity and electronics, in Susanne P. LAJOIE and Martial VIVET, editors, *Artificial intelligence in education*, pp. 759–761, IOS, Amsterdam.

Paweł STROJNY and Agnieszka STROJNY (2014), Kwestionariusz immersji – polska adaptacja i empiryczna weryfikacja narzędzia, *Homo Ludens*, 1(6):171–186.

Mariusz URBAŃSKI, Katarzyna PALUSZKIEWICZ, and Joanna URBAŃSKA (2016a), Erotetic problem solving: From real data to formal models. An analysis

of solutions to erotetic reasoning test task, in F. PAGLIERI, L. BONETTI, and S. FELLETTI, editors, *The Psychology of Argument: Cognitive Approaches to Argumentation and Persuasion*, pp. 33–46, College Publications.

Mariusz URBAŃSKI, Natalia ŻYLUK, Katarzyna PALUSZKIEWICZ, and Joanna URBAŃSKA (2016b), A formal model of erotetic reasoning in solving somewhat ill-defined problems, in D. MOHAMMED and M. LEWIŃSKI, editors, *Argumentation and Reasoned Action. Proceedings of the 1st European Conference on Argumentation*. London: College Publications, pp. 973–983, College Publications.

Noortje VENHUIZEN, Valerio BASILE, Kilian EVANG, and Johan BOS (2013), Gamification for word sense labeling, in *Proceedings of the 10th International Conference on Computational Semantics (IWCS'13) – Short Papers*, pp. 397–403.

Anthony J. VIERA and Joanne M. GARRETT (2005), Understanding interobserver agreement: the kappa statistic, *Family Medicine*, 37(5):360–363.

Aleksandra WASIELEWSKA and Paweł ŁUPKOWSKI (2022), IMUW the questionnaire measuring the engagement of attention in a task execution, <https://osf.io/6dt8f/>.

Andrzej WIŚNIEWSKI (2013), *Questions, inferences and scenarios*, College Publications, London.

Zulipiye YUSUPUJIANG and Jonathan GINZBURG (2020), Designing a GWAP for collecting naturally produced dialogues for low resourced languages, in *Workshop on Games and Natural Language Processing*, pp. 44–48.

Zulipiye YUSUPUJIANG and Jonathan GINZBURG (2021), Data collection design for dialogue systems for low-resource languages, *Conversational Dialogue Systems for the Next Decade*, pp. 387–392.

Zulipiye YUSUPUJIANG and Jonathan GINZBURG (2022), UgChDial: A Uyghur chat-based dialogue corpus for response space classification, in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 3140–3149.

Paweł Łupkowski

ORCID 0000-0002-5335-2988

Pawel.Lupkowski@amu.edu.pl

Ewelina Chmurska

Adrianna Płatosz

Aleksandra Kwiecień

Barbara Adamska

Magdalena Szkalej

Faculty of Psychology and Cognitive
Science

Adam Mickiewicz University

Szamarzewskiego 89a, 60-568 Poznań

Jonathan Ginzburg

ORCID 0000-0001-5737-0991

yonatan.ginzburg@u-paris.fr

Université Paris Cité, CNRS,
Laboratoire de Linguistique Formelle
5 Rue Thomas Mann,
75205, Paris

Paweł Łupkowski, Jonathan Ginzburg, Ewelina Chmurska, Adrianna Płatosz, Aleksandra Kwiecień, Barbara Adamska, and Magdalena Szkalej (2024), *QRGS – Question Responses Generation via crowdsourcing*, *Journal of Language Modelling*, 12(1):213–270

DOI <https://dx.doi.org/10.15398/jlm.v12i1.372>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

CC BY <http://creativecommons.org/licenses/by/4.0/>