Linguistic Acceptability and Usability Enhancement: A Case Study of GWAP Evaluation and Redesign

Wateen Aliady and Massimo Poesio
Queen Mary University of London, Imam Mohammad Ibn Saud University
United Kingdom, Saudi Arabia
{ w.a.a.aliady, m.poesio}@qmul.ac.uk

Abstract

Collecting high-quality annotations for Natural Language Processing (NLP) tasks poses challenges. Gamified annotation systems, like Games-with-a-Purpose (GWAP), have become popular tools for data annotation. For GWAPs to be effective, they must be user-friendly and produce high-quality annotations to ensure the collected data's usefulness. This paper investigates the effectiveness of a gamified approach through two specific studies on an existing GWAP designed for collecting NLP coreference judgments. The first study involved preliminary usability testing using the concurrent think-aloud method to gather open-ended feedback. This feedback was crucial in pinpointing design issues. Following this, we conducted semi-structured interviews with our participants, and the insights collected from these interviews were instrumental in crafting player personas, which informed design improvements aimed at enhancing user experience. The outcomes of our research have been generalized to benefit other GWAP implementations. The second study evaluated the linguistic acceptability and reliability of the data collected through our GWAP. Our findings indicate that our GWAP produced reliable corpora with 91.49% accuracy and 0.787 Cohen's kappa.

Keywords: games-with-a-purpose, natural language processing, coreference annotation, usability study, language acceptability

1. Introduction

Games-with-a-Purpose (GWAP) offers a promising approach to leveraging human computation for annotation tasks (Von Ahn and Dabbish, 2004; Von Ahn, 2006; Von Ahn and Dabbish, 2008; Von Ahn et al., 2006a,b; Madge et al., 2019a; Lafourcade et al., 2015; Chamberlain et al., 2008; Amspoker and Petruck, 2022; Morrison et al., 2023; Chaiko et al., 2022; Xu et al., 2022; Mount Cieri et al., 2020). They involve designing games to collect annotations from players, utilizing their gaming skills and language competence, with entertainment serving as the primary incentive (Poesio et al., 2013; Vannella et al., 2014; Jurgens and Navigli, 2014). These systems combine entertainment with task completion and hold significant potential across various fields, including data annotation and problem-solving.

Usability is a critical aspect of GWAPs, impacting their success and acceptance. Another crucial factor is the reliability of their annotations, especially in Natural Language Annotation (NLP) tasks, where linguistic acceptability is vital. Our work explores usability and linguistic acceptability in GWAPs, particularly in the context of a 3D game designed for Arabic NLP annotation. We aim to evaluate these components to improve understanding and assist researchers in this field.

This study has three primary objectives: (1) to conduct a preliminary usability study of the presented GWAP and identify areas for enhancement, (2) to redesign the GWAP based on usability findings, and (3) to evaluate the linguistic acceptability of the collected judgments. By addressing these objectives, we aim to contribute to creating more user-centered GWAPs. Particularly, we target the following research questions:

Q1: In the context of 3D games, mainly focusing on the interface/menu layer, what design elements, interaction techniques and user experience factors in the interface/menu layer improve the usability of the players as informed by participant feedback from our usability test?

Q2: Could our virtual world game, *Stroll-with-a-Scroll*, be used to collect linguistically acceptable coreference annotation? Coreference resolution is clustering the mentions in a text that refer to the same real-world entity.

The next section of this paper discusses related work. A brief description of the design of our GWAP follows this. Next, we show an in-depth description of the preliminary study that guided the redesign process. Finally, we present the linguistic acceptability of our game.

2. Background and Related Work

Games-with-a-Purpose (GWAPs) are typically designed to leverage players' skills and abilities, primarily for entertainment. They have found ap-

Games and NLP @LREC-COLING-2024, pages 85–96
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

plications in various domains such as biological data collection (Kleffner et al., 2017; Kawrykow et al., 2012), image processing in AI (Krause et al., 2010), assessment and comparison of Explainable AI (XAI) techniques (Morrison et al., 2023), music annotation (Kim et al., 2008), and dance movement annotation (Kougioumtzian et al., 2022). In dance movement annotation, for instance, a notator or movement analyst describes and documents dance movements by recording details of the body's actions using a coding system. Similarly, in Natural Language Processing (NLP), GWAPs are employed for tasks like text annotation (Venhuizen et al., 2013; Madge et al., 2019b; Fort et al., 2014; Kicikoglu et al., 2019; Bonetti and Tonelli, 2020; Dziedzic, 2016; Xu et al., 2022) or generating original content for annotation (Amspoker and Petruck, 2022).

The evaluation of the usability of these gamified systems holds significant importance as it contributes to reducing errors, training time, and learning effort while also enhancing productivity and satisfaction (Rajanen and Dorina, 2017). For instance, in (Tomé Klock et al., 2017), ten different gamified educational systems were assessed using ergonomic criteria guidelines that evaluate usability and user experience. Additionally, a systematic literature review in (Laine and Lindberg, 2020) provided generalized recommendations to improve motivation in gamified systems, such as offering feedback, using familiar vocabulary, ensuring actions align with goals, and maintaining consistency across elements. However, (Gouveia et al., 2023) demonstrated that usability significantly correlates with intrinsic motivation in a virtual reality gamified system designed for rehabilitation purposes.

In usability testing, qualitative research methods like interviews, surveys, and focus groups offer insights based on users' self-reports (Roberts et al., 2019). One effective method involves employing the think-aloud (TA) protocol, where participants articulate their thoughts and feelings while interacting with a product or system. This approach enables researchers to gain insights into users' cognitive processes, thereby identifying potential issues or challenges users may encounter during interaction.

TA protocols represent one of the most prevalent methods for identifying User Experience (UX) issues during usability testing (Fan et al., 2020; McDonald et al., 2013). There are two commonly used TA protocols in the industry: Concurrent Think-Aloud (CTA) and Retrospective Think-Aloud (RTA) (Fan et al., 2020; McDonald et al., 2013). In CTA, users vocalize their thoughts while performing a task, whereas in RTA, users complete the task and then articulate their thoughts by reviewing a recording. There has been an ongoing debate regarding which protocol is superior (Van den Haak and De Jong, 2003; Alshammari et al., 2015). We opted for CTA due to its popularity among UX practitioners (Fan et al., 2020; McDonald et al., 2013), as it allows UX evaluators to delve into participants' thought processes in real-time interaction with applications, which cannot be captured solely through retrospective self-reports.

Ensuring the linguistic reliability of the collected data is crucial. Consequently, some GWAPs have addressed this issue. For instance, *Phrase Detectives* (Poesio et al., 2013), a GWAP for English and Italian coreference annotation, initially employed majority voting to aggregate player feedback. They then assessed the acceptability of the collected judgments by comparing experts' annotations with the data derived from the majority vote of non-experts. The result indicated an 84% agreement across all cases, comparable to those observed when comparing an expert with an average annotator, typically trained students producing medium-quality annotations. Subsequently, *Phrase Detectives* improved aggregation by adopting Mention Pair Annotation (Paun et al., 2018), a dedicated probabilistic aggregation method for coreference. Here, players identify the nearest antecedent, and the best pairing is determined based on a probabilistic model (Paun et al., 2018). These pairs are then clustered to form a coreference chain, increasing the accuracy of the produced judgments to 92% (Poesio et al., 2019).

Games like the original von Ahn games and, for NLP, *Puzzle Racer*, have demonstrated the feasibility of entertaining GWAPs that generate high-quality annotations at a reduced cost (Jurgens and Navigli, 2014). Another example is *High School Superhero*, a GWAP developed for collecting acceptability judgments. It evaluated the resulting annotations in terms of agreement among players and compared them with experts' judgments (Bonetti et al., 2022). Additionally, *RigorMortis* measured acceptability in annotating multi-word expressions for French corpora (Fort et al., 2020).

3. Introduction to the Game: Stroll-with-a-Scroll

Stroll-with-a-Scroll represents the first virtual world GWAP designed for Arabic Natural Language Processing (NLP) tasks, featuring a treasure hunt theme set in an ancient Middle Eastern fictional town within a desert landscape. The game incorporates a narrative element at the outset of gameplay, inspired by the findings of a study on narrative importance (Krause et al., 2010). This narrative is presented through a cut scene, establishing the storyline and themes for players. As avatars dressed in traditional attire, players navigate the town, embarking on a quest to discover hidden chests scattered throughout the environment.

The game employs a navigation system displayed on the menu layer to aid players in locating chests, utilizing three colors (red, yellow, and green) to indicate proximity to the chest. Upon reaching a chest, players uncover a scroll containing text with torn sections. Given the age of these scrolls and the missing sections, players must solve puzzles to reconstruct the text. The puzzle mechanics, inspired by *Wormingo* (Kicikoglu et al., 2019), include selecting the correct word from the provided options and solving 'word search' puzzles within a grid of letters.

The coreference annotation task within the game follows the approach of *Phrase Detectives* (Chamberlain et al., 2008) and *Wormingo* (Kicikoglu et al., 2019), presenting players with annotation and validation questions. Annotation questions prompt players to identify whether a mention is new or old, with the option to select the antecedent if it is old or to skip the question. Validation questions, on the other hand, require players to evaluate other players' responses.

For post-processing, *Stroll-with-a-Scroll* adopts the methodology of *Phrase Detectives* (Chamberlain et al., 2008), utilizing Mention Pair Annotation (Paun et al., 2018) for probabilistic aggregation of coreference annotations. After collecting judgments from multiple players, this method selects the best pairing based on a probabilistic model, and then clusters pair to form coreference chains.

4. Preliminary Usability Study

A game's usability significantly impacts players' enjoyment and motivation, ultimately influencing participant numbers. Conducting early usability tests is crucial for enhancing user experience, streamlining navigation, promoting intuitive usage, and identifying design flaws. By observing user interactions with the proposed system, a usability study can pinpoint areas for improvement, thereby reducing dropouts. Moreover, usability has been demonstrated to affect users' engagement with Games-with-a-Purpose (GWAPs) (Bowser et al., 2013; Hamari and Keronen, 2017; Bui et al., 2020), highlighting its pivotal role in fostering user participation. Additionally, usability is a cornerstone of a successful virtual world (Lee and Chen, 2011), as it determines how effectively a virtual space facilitates specific tasks for particular users.

4.1. Participants and Procedure

In the preliminary study, we recruited 8 participants, consisting of 5 females and 3 males. The mean age of the participants was 28.5 years (SD=2.57).

A qualitative approach is typically preferred at this stage of development as it aids designers in identifying issues or bugs early on and making enhancements. Additionally, a qualitative approach can be utilized to comprehend player engagement in our game. Accordingly, numerous scholars have employed qualitative methodologies to explore engagement in virtual worlds (Chen and Kent, 2020; Bouta Cruz-Benito et al., 2015; Kohler et al., 2011). A think-aloud protocol (Lewis, 1982) is employed to gather data, allowing for open feedback collection. Given the early stage of game development, such open user feedback is crucial for testing usability and gaining initial insights into player engagement. We utilized the concurrent think-aloud (CTA) protocol, wherein users verbalize their thought processes while simultaneously working on a task, as it is more widely favored among UX practitioners (Fan et al., 2020; McDonald et al., 2013).

The study was conducted from October 24th to 30th, 2022, with each session lasting approximately 30-minute. Prior to commencing the preliminary test, participants were provided with informed consent outlining the study's objectives. Subsequently, they were introduced to the following tasks they were required to accomplish:

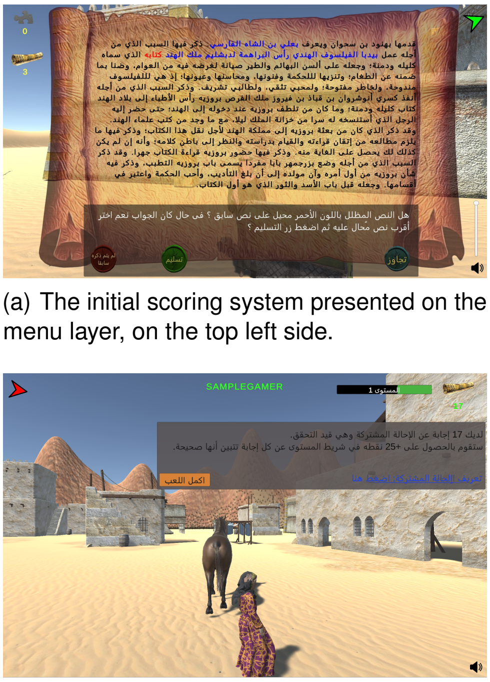
- First, sign up to join and start the game.
- Then, start the game and read the pre-game description.
- After you complete the pre-game part, navigate the scene to locate the chest.
- Finally, respond to the presented puzzles and the linguistic task, then navigate again to search for the following chest.

Our analysis is grounded in Reflexive Thematic Analysis (Braun and Clarke, 2019, 2021) chosen for its suitability with a small sample size of participants– in this case, eight participants in this experiment, and due to its flexibility in organizing results into common themes. The data was initially transcribed and then coded to create an affinity diagram, from which themes emerged. The following themes were generated from the analysis:

4.1.1. User Interface Refinement, to Ensure Familiarity and Simplicity Theme

More Familiarity Theme: The study by (Abada and Onibere, 2009) demonstrated that prior computer experience plays a significant role in intuitively understanding and using new computer software. This principle extends to video games (Miller et al., 2019). Intuitive interfaces are crucial in game design, with schema theory explaining how individuals grasp gameplay mechanics without prior experience with a particular game. A concise definition of gameplay provided by Lindley and Sennersten (Lindley et al., 2008) describes it as "the structure and algorithm determining the management of attentional and other cognitive, perceptual, and motor resources required to realize the tasks involved in gameplay."

For example, Participant 1 commented, *"I am used to using the (W, S, D, A) buttons to move around in games."* Additionally, Participant 2 inquired, *"Does the Shift button speed up the character?"*, reflecting the common practice of using the Shift key to increase the speed in games. In our game, we utilize arrow keys for movement and the Shift button to enhance movement speed, aligning with standard video game conventions. However, providing additional explanations may be necessary, especially for novice users unfamiliar with these conventions. Two of our participants required instructions on manoeuvring the avatar, while three participants were unsure how to begin, necessitating an explanation of the functionality represented by the upper-right pointer, as depicted in Fig 1(a).



قدمها بهنود بن سحوان ويعرف بعلي بن الشاه الفارسي. ذكر فيها السبب الذي من
أجله عمل بيدبا الفيلسوف الهندي رأس البراهمة لدبشليم ملك الهند كتابه الذي سماه
كليله ودمنة؛ وجعله على ألسن البهائم والطير صيانة لغرضه فيه من العوام، وضنا بما
ضمنه عن الطغام؛ وتنزيها لللحكمة وفنونها، ومحاسنها وعيونها؛ إذ هي لللفيلسوف
مندوحة، ولخاطر مفتوحة؛ ولمحبي تثقي، ولطالبي تشريف. وذكر السبب الذي من أجله
أنفذ كسري أنوشروان بن قباذ بن فيروز ملك الفرص بروزيه رأس الأطباء إلى بلاد الهند
كتاب كليله ودمنة؛ وما كان من تلطف بروزيه عند دخوله إلى الهند؛ حتى حضر إليه
الرجل الذي أستنسخه له سرا من خزانة الملك ليلا، مع ما وجد من كتب علماء الهند.
وقد ذكر الذي كان من بعثة بروزيه إلى مملكة الهند لأجل نقل هذا الكتاب؛ وذكر فيها ما
يلزم مطالعه من إتقان قراءته والقيام بدراسته والنظر إلى باطن كلامه؛ وأنه إن لم يكن
كذلك لك يحصل على الغاية منه. وذكر فيها حضور بروزيه قراءة الكتاب جهرا. وقد ذكر
السبب الذي من أجله وضع بزرجمهر بابا مفردا يسمى باب بروزيه التطبب، وذكر فيه
شأن بروزيه من أول أمره وآن مولده إلى أن بلغ التأديب، وأحب الحكمة واعتبر في
أقسامها. وجعله قبل باب الأسد والثور الذي هو أول الكتاب.

هل النص المظلل باللون الأحمر محيل على نص سابق ؟ فى حال كان الجواب نعم اختر
أقرب نص مُحال عليه ثم اضغط زر التسليم ؟

(a) The initial scoring system presented on the menu layer, on the top left side.

لديك 17 إجابة عن الإحالة المشتركة وهي قيد التحقق.
ستقوم بالحصول على 25+ نقطه في شريط المستوى عن كل إجابة تتبين أنها صحيحة.

تعريف الحالة المشتركة: اضغط هنا

(b) The scoring system was updated, and descriptive info is added for the scoring process and coreference annotation.

Figure 1: Improved game scoring clarity: Usability test enhancements.

Another example of familiar design is allowing players to close instructional prompts permanently or view them at their discretion. For instance, Participant 1 expressed frustration with the repetitive annotation task pop-ups, commenting, *"It bothers me that the instructions keep showing. I am used to having the option of never showing that again."* Similarly, Participant 5 was displeased with the frequent closure of these pop-ups. However, despite these concerns, having instructional pop-ups is crucial, particularly for the coreference task. This is because players often tend to dismiss instructions without reading them, as highlighted in a study by (Fraser, 2015), where it was noted that students frequently close instructions without fully engaging with them. As a result, it was recommended to implement a pop-up before each task to ensure that players are adequately informed.

Fewer Chunks of Text Theme: Having fewer chunks of text is better for increasing reading comprehension and avoiding player frustration, as this was tested in the onboarding phase of *Lingotowns* (Althani et al., 2022). This design was followed by *PlayCoref* (Hladká et al., 2009) and *Wormingo* (Kicikoglu et al., 2019), English coreference annotation games. This technique is employed in *Wormingo* in the form of "chunks". In our preliminary study, P2 commented, *"There are too many linguistic questions for a single scroll. I am spending too much time on that, and it feels like a task rather than a game."* Also, P3 said *"The coreference task is just overwhelming; there are too many questions in a single chest."*. In addition, the rest of the players suggested making the task less overwhelming.

4.1.2. Reshaping and Adding Game Design Elements Theme

Reshape the Reward System Theme: Insufficient guidance in games can lead to player frustration. According to (Miller and Cooper, 2022), many issues encountered in citizen science games stemmed from designers failing to convey critical scientific concepts to players, resulting in frustration. Participant 1 expressed confusion regarding the game's dual scoring systems:

"Why do we have two scoring systems? I understand that the first scoring system is for puzzle points, but what does the other do?... It seems like the other one is used for answering the annotation questions, but I still don't understand why I receive points for each answer I submit. I even tried submitting a wrong answer and still received a point. Could this incentivize players to provide any answer to earn points?"

The challenge is that scoring for the annotation section is not immediate, as correct answers are not known immediately. Instead, all player-provided answers are recorded under the second scoring system, represented by a scroll icon (see Figure 1). Once validated, players receive additional points under this system without explanation. All participants highlighted the need to clarify why there are two scoring systems and how scoring is calculated.

Reshape the Feedback Theme: Feedback in Games-with-a-Purpose (GWAPs) is crucial as it impacts player retention, as players desire recognition for their contributions and reassurance that they are making a difference. This finding was corroborated by a citizen science game interviews, highlighting factors contributing to player immersion in the game world (Miller and Cooper, 2022). In the design of *Stroll-with-a-Scroll*, feedback is provided to players while solving puzzles, with a checkmark indicating a correct answer and a cross indicating an incorrect one. However, further improvements are necessary, as Participant 4 suggested: *"The feedback for the puzzles was too quick. It needs to be slowed down."* Additionally, Participant 3 commented on the puzzle scoring: *"There is varying difficulty between the two presented puzzles, the fill-in-the-blank and the word search puzzle, and therefore, there should be varying scoring based on difficulty and the time it takes to solve the puzzle."*

Add New Game Elements: Using leaderboards and assigning levels based on points is an effective motivator, with users often viewing these as targets to strive for (Lee et al., 2013; Von Ahn and Dabbish, 2004, 2008). Participant 5 emphasized the importance of leaderboards, stating, *"I think it is important to have a leaderboard as most games include that."* Additionally, Participant 1 suggested locking access to leaderboards for players until they reach a certain level, while Participant 7 underscored the significance of this feature. Moreover, a few players suggested incorporating puzzles within the virtual world, allowing players to explore the landscape while solving puzzles.

5. Player-Centred Design: Insights from Usability Testing for Game Development

Moving beyond basic understanding and truly grasping players' personalities when introducing gamification is crucial. This understanding aids in creating a user-friendly game that effectively motivates players. One method to achieve this is through player personas.

Player personas are not merely demographic profiles or stereotypes but crafted from authentic data from surveys, interviews, analytics, user testing, and other reliable sources. These personas capture players' motivations, frustrations, pain points, and aspirations, as well as provide insights into their gaming habits, preferences, and playing styles. As outlined in (Guzman-Mendoza et al., 2021), player personas are developed by studying and understanding player behavior. By observing how players navigate and interact with game mechanics, their patterns and interactions are analyzed to create meaningful personas.

5.1. Persona Profiling

To redesign our game, we developed player personas through in-depth interviews. After the usability test, participants were interviewed in person for about 20-minute in a semistructured format regarding their educational background, employment status, their play experiences and game skills, and their fluency in Modern Standerd Arabic (MSA). Examples of asked questions were "How long have you been playing games, and what types of games do you typically enjoy?", "Can you describe your level of proficiency in Modern Standard Arabic? and can you speak confidently in MSA during conversations or presentations?" and "Would you consider to participate in NLP based GWAP? what particular features or aspects of the game would encourage your participation? and what features might hinder your motivation to participate?". Also, participants were asked for their input, on how to enhance the current design to address any concerns they had raised. The interviews were audio-recorded and then transcribed for further analysis.

Our analysis revealed the goals,and challenges and preferred playing modes participants expressed during the interviews and game testing. This information is summarized in Table 1. Below are the key steps in our methodology for creating player personas:

- **Participants:** The sample comprised 8 participants: 5 students (2 undergraduates with gaming experience, 3 graduate students: one was an unemployed gamer and the other two were employees with gaming background ; one of these employees is a linguistic researcher), 1 unemployed individual had a high school degree with a gaming background, another employee individual had a bachelor's degree with no gaming experience but a strong linguistic background, and one employee held a master's degree with some gaming background and strong linguistic skills.
- **Instrument:** We devised a Face-to-face indepth interview methodology in a semistructured format regarding their educational background, employment status, their play experiences, game skills, and their fluency in MSA.
- **Procedure:** The interviews were audio-recorded to facilitate transcription of participants' comments and aid in analysis. We analyzed data to define the player persona based on Demographic Information, Professional Information and Playing Culture.
- **Data Analysis:** Our analysis involved identifying categories and codes based on constructs

	The Linguistic Enthusiastic	The Novice Player	The Gamer
Goal	They aim to improve their annotation skills and contribute to research projects in the field of NLP. Their aspiration is to collaborate with researchers and make contributions to the linguistic community.	They are interested in the notion of contributing to research through annotations but feel intimidated by the idea of using technological platforms.	They are looking for gaming experiences that suit their preferences, providing captivating gameplay mechanics without the need for complicated annotation tasks.
Pain Points	Balancing their workload and personal responsibilities while devoting time to annotations can be quite difficult for them.	They find it difficult to navigate and interact with interfaces related to gaming platforms.	They feel overwhelmed by the complexity of NLP annotation tasks.
Individual Achievement vs. Team Achievement	While they place importance on growth by enhancing their annotation skills and contributing meaningfully to research projects, they also recognize the value of teamwork in achieving research goals.	They value the opportunity to collaborate with others and benefit from their expertise while collectively working towards shared research objectives.	While they appreciate achievement in mastering gaming skills and conquering in-game challenges, some gamers also find joy in collaborating with other players, while others enjoy a competitive environment.

Table 1: Player-Persona Insights.

to extract information for designing the player persona.

6. Evolution of the Game: Stroll-with-a-Scroll(Version 2)

Based on the generated personas, we have redesigned the game into its second version. Firstly, addressing the 'More Familiarity' theme, we aimed to include novice players in the design process by adding clear directions on how to play. Drawing from a detailed framework of design strategies for enhancing learnability in video games (Poretski and Tang, 2022), we introduced just-in-time reminders (as shown in Fig 2), contextual prompts appearing in specific game situations that vanish once performed by the player. These prompts guide players on how to move around and what actions to take, eliminating the need to memorize instructions before gameplay. They appear only once before a new action is required. Additionally, we carefully considered the needs of gamers and experts who are always on the move during gameplay. Addressing participants' complaints about the lack of an option to 'never show again' or 'hide' task descriptive pop-ups and the need for a clear definition of coreference with examples, we introduced the coreference task description at the start and placed it within the scroll icon (as depicted in Fig 1(b)). Clicking on the scroll icon directs players to the task description, allowing them to view it upon request. To ensure players are aware of the description location, we added guidance at the start of gameplay, clarifying that they can refer to the task description by clicking on the scroll icon. Additionally, in response to player requests, we added (W, S, D, A) buttons for avatar movement control.

Secondly, to address the issue of overwhelming text highlighted in the 'Less chunks of text' theme, we adopted the chunk size approach used in *Wormingo* (Kicikoglu et al., 2019), as 'The Gamer' group of participants, the most intimidated by text size, did not report feeling overwhelmed during their experience with the game. When the player opens the scroll, the text is presented in chunks, one after the other. Each chunk contains a maximum of 50 words, ensuring complete sentences are displayed. Additionally, we implemented a gradual display of words, simulating an animated effect similar to *Wormingo*, to reduce cognitive load on players (Kicikoglu et al., 2019). We introduced virtual world puzzles to mitigate text overload raised mostly by 'The Gamer' group and to address the 'Add New Game Elements' theme. In these puzzles, players are tasked with searching the scene for lost letters in the scroll. The game presents three lost letters forming the word 'day'. As players search the scene, each missing letter is revealed with an Arabic coffee cup (as shown in Fig. 3). When a player finds a letter, it moves from the scene to be placed on top of the Arabic coffee cup, ultimately completing the word. Players have the option to hide and reveal the text by pressing the eye icon (as depicted in Fig. 3). Additionally, they can skip playing the game part by pressing the 'Skip' button, allowing those focused on annotation; 'The linguistic enthusiastic' group to continue without participating.

Thirdly, participants requested further explanations regarding the reward systems and their calculation processes, highlighting the need for a more intuitive presentation. Two rewarding systems were identified: instant rewards for solving puzzles and delayed rewards for solving annotations, as described in the 'Reshape the reward system' theme. To address this issue, we made adjustments to the menu depth items and the gameplay scene. We introduced a level bar, suggested by 'The Gamer' group, to mark progress providing instant points for solving puzzles while temporarily recording an-

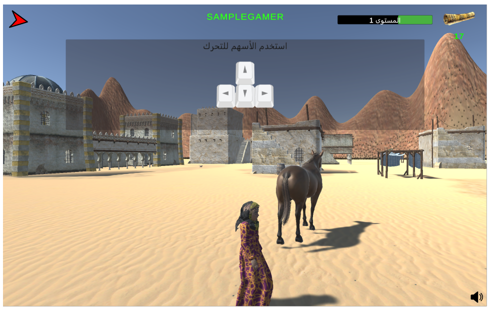


Figure 2: Directions on how to play, presented on the menu layer of the game.

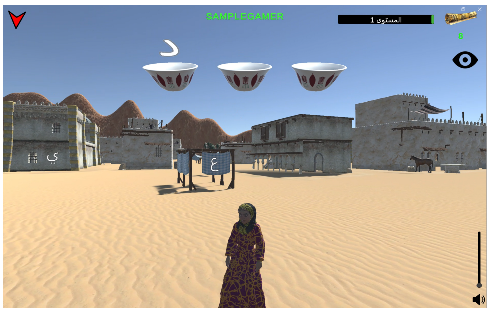


Figure 3: Lost letters puzzle: Scene search for the letters of the lost word.

notation answers within the scroll menu item until validated and presented within the progress bar (as shown in Fig. 1(b)). To clarify these rewarding systems we instructed players to click on the scoring systems at the start of gameplay: the level bar and then the scroll icon, where they received clarifications on the calculations and why instant points were not awarded for the annotation task. Additionally, the description of the coreference task was provided there for players to access as needed, eliminating pop-ups before each task. In response to the competitive nature of some players, a leaderboard was added to the home screen, addressing the 'Add New Game Elements' theme.

Furthermore, some players expressed a desire to enhance the enjoyment of game feedback, referring to it as "game juiciness", seeking elements that excite them. This was discussed in the 'Reshape the feedback Theme'. Game juiciness involves providing visual and audio feedback to induce a positive player experience (Rollings and Morris, 1999). In the initial version of the game, background music was included, with players able to control the sound level or mute it. Feedback was displayed as a check mark for correct answers and a cross mark for incorrect ones. However, players found this feedback too quick to absorb, prompting us to slow it down and add audio feedback for success and failure. We also implemented animated scoring similar to *Wormingo*, where correct answers are rewarded with an animated score, transitioning from the challenge to the corresponding reward system. Additionally, scoring now varies based on puzzle difficulty, with players receiving a more valuable animated treasure box for answering virtual world puzzles.

7. Redesign Validation

After making improvements, to our GWAP to make it more user friendly, it is important for us to carefully evaluate the effectiveness of these changes through thorough usability testing. In this section we will provide an explanation of how we validate the redesign and the methods we use to assess the systems usability.

We selected same series of tasks given in the first usability test. We had 3 participants, who were asked to perform the predefined tasks while thinking aloud. Following each task, participants were interviewed to gather feedback on their overall experience, usability challenges faced, and suggestions for improvement.

Based on the usability testing it seems that the redesign successfully enhanced the user friendliness of our GWAP, as participants successfully completed the task independently without raising any concerns about the issues that were identified in the first usability test.

8. Linguistic Acceptability Study

A debate persists regarding the use of expert versus non-expert annotators and the reliability of different crowdsourcing strategies in the realm of NLP annotation tasks. To address this, we tested our annotations' reliability to assess our GWAP's reliability. We aim to share these results with other researchers to encourage linguists to participate in annotating our GWAP and to disseminate them widely.

8.1. The data

Our objective is to compare players' judgments with those of experts, so our players annotated a gold standard document extracted from the OntoNotes 5.0 datasets. OntoNotes is widely utilized for coreference resolution (R. et al., 2014; Björkelund and Kuhn, 2014; Martschat and Strube, 2015; Clark and Manning, 2015, 2016a,b; Lee et al., 2017, 2018) and has been a key resource since the CoNLL 2011 and 2012 shared tasks (Pradhan et al., 2011). It encompasses documents in three languages: Arabic (300K tokens), Chinese (950K tokens), and English (1.6M tokens), spanning various genres, with news being the predominant genre. Our study used a single 'Art News'

CoNLL document containing MSA text annotated with coreference.

8.2. Participants and Procedure

We aim to evaluate whether our virtual world game, *Stroll-with-a-Scroll*, can effectively collect linguistically acceptable coreference annotations. To achieve this, we conducted an experiment in August 2023 to compare the annotations provided by naive participants (our participants) with those of expert annotators.

We recruited some of our participants (N=77) through Prolific, a platform for online participant recruitment. We used the demographic filters provided by the platform, to selectively enroll participants whose first language was Arabic. This measure was implemented to mitigate potential confounding variables that might impact the accuracy metrics within our research investigation. Participating individuals were paid £7 (£12 per hour) upon successful completion of a 35-minute study entitled "Study about a Game-with-a-Purpose." Additionally, we enlisted volunteers (N=29) who received invitation emails, and whose first language is Arabic. These emails were sent to academic faculty in Saudi universities, requesting them to share the game with their students. Due to technical constraints, the experimental protocol could only be executed on desktop or laptop web browsers such as Chrome and Firefox; consequently, participation via mobile devices was not feasible due to these limitations.

Of the 106 participants who completed the demographic questionnaire, 44.34% were female, and 55.66% were male. Regarding age distribution, 28.30% were aged 18 to 24, 46.23% were aged 25 to 34, 15.09% were aged 35 to 44, 8.49% were aged 45 to 54, and 1.88% were aged 55 and above. The participants comprised 23.58% Saudis, 13.20% Lebanese, and 13.20% Syrians. The remainder represented various nationalities, including Algerian, Iraqi, Jordanian, Moroccan, Palestinian, Somali, Sudanese, and Tunisian.

In our annotation task, players are presented with a text window highlighting a specific word or phrase in red. Their initial task is to determine whether the highlighted word or phrase is newly introduced to the conversation or if it refers to something previously mentioned. If it refers to a previous mention, players must locate it by selecting one of the highlighted ones in blue. Once the player has made their selection, they can submit their annotation by clicking on the Submit button. During the validation mode, players confirm other players' answers, which is activated only when players submit different answers. The experiment concluded after 47 markables were annotated, and all answers were aggregated using MPA (Mention Pair Annotation) (Paun et al., 2018) and stored in an inline XML file.

8.3. Analysis and Results

The agreement between naive annotators and linguists is 91.49% overall accuracy, calculated by comparing the markable in the generated XML file with the gold standard file. This result is excellent. Additionally, our Cohen's kappa coefficient, a more robust measure accounting for the possibility of chance agreement, is 0.787 (Cohen, 1960), indicating substantial agreement. Our participants failed to answer 4 out of the 47 presented mentions.

In terms of precision, recall, and F-Measure, players' annotations were compared to the gold standard (the OntoNotes annotation). The data suggested that Precision, Recall, and F-Measure collectively evaluate annotation accuracy and completeness, with a balanced score of 0.84615 indicating both precision and recall around 85% for players' annotations. Furthermore, we compared individuals who were paid and those who volunteered for our experiment, specifically examining their accuracy levels when completing annotation tasks. Our analysis revealed that both paid and volunteer participants achieved similar accuracy scores. This consistency across participant groups demonstrates our GWAP's strength and dependability.

9. Discussion

In this article, we have discussed two factors that contributing to the success of GWAPs: user-friendliness and the reliability of the generated data. First, to ensure user-friendliness, we conducted a usability study that helped us create personas and guide our design process accordingly. From this study, we generalized our findings to inform the design of other NLP annotation games, as we aim to answer our first research question: "In the context of 3D games, particularly focusing on the interface/menu layer, what design elements, interaction techniques, and user experience factors improve the usability and productivity of the player?"

Firstly, introducing breaks between annotations enhances overall enjoyment, but it's essential to make these breaks optional for participants who prioritize contributing over gameplay. Secondly, instructions and tutorials should be concise, quick to understand, and easily accessible during gameplay to ensure a seamless experience without interruptions. Thirdly, clear explanations of the calculation process for annotation tasks reduce frustration and enhance understanding among participants. Fourthly, incorporating visually and audibly satisfying feedback mechanisms for player actions improves engagement. Fifthly, incorporating both competitive and collaborative elements is recommended to accommodate diverse preferences and play styles. Finally, simplifying NLP tasks by breaking them into smaller, manageable tasks enhances user involvement and potentially creates more reliable data.

Usability is an iterative process, and we actively seek more participants to conduct further tests. These tests may involve creating or updating personas with more detailed information. By improving our understanding of user preferences, our goal is to make our GWAP more user-friendly and effective.

We tested the reliability of our generated corpora to answer our second research question, “Could our GWAP be used to collect linguistically acceptable coreference annotation?” We achieved excellent results by assigning more weight to reliable players when aggregating annotation answers (Paun et al., 2018) instead of simply annotating with the value submitted the most. Out of all markables presented, our participants failed to annotate only 4, which is less than 10% of the total.

10. Concluding Remarks

This paper presents two evaluations of a 3D virtual world game designed for NLP annotation. First, we conducted a preliminary study to improve user experience and identify design flaws. The usability test involved observing how users interacted with the system and identifying areas for enhancement or correction to minimize dropouts. Tasks were assigned to participants, who provided feedback using the think-aloud protocol. We redesigned the game tailored to player personas based on qualitative research findings. Secondly, we evaluated the reliability and acceptability of the game for collecting annotations by comparing aggregated player feedback to the OntoNotes 5.0 gold standard corpus. Our analysis indicates that annotations produced through the game are of acceptable quality.

11. Acknowledgements

Wateen Aliady is supported by the Saudi Arabian Cultural Bureau in the UK and Imam Mohammad Ibn Saud Islamic University (IMSIU) in Saudi Arabia. Massimo Poesio was in part supported by the ARCIDUCA project, funded by European Research Council (EPSRC).

12. Bibliographical References

G. O. Abada and E. A. Onibere. 2009. The effect of rehearsed computer use on icon recognition. *International Journal of Computers and Applications*, 31:9–15.

T. Alshammari, O. Alhadreti, and P. Mayhew. 2015. When to ask participants to think aloud: A comparative study of concurrent and retrospective think-aloud methods. *International Journal of Human Computer Interaction*, 6(3):48–64.

F. Althani, C. Madge, and M. Poesio. 2022. Less text, more visuals: Evaluating the onboarding phase in a gwap for nlp. *Proceedings of the 9th Workshop on Games and Natural Language Processing within the 13th Language Resources and Evaluation Conference*, pages 17–27.

E. Amspoker and M. Petruck. 2022. A gamified approach to frame semantic role labeling. In *Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances)*, page 37–42.

A. Björkelund and A. Kuhn. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of the 52nd ACL*, volume 1, page 47–57.

F. Bonetti, E. Leonardelli, D. Trotta, G. Raffaele, and S. Tonelli. 2022. Work hard, play hard: Collecting acceptability annotations through a 3d game. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1740–1750. European Language Resources Association.

F. Bonetti and S. Tonelli. 2020. A 3d role-playing game for abusive language annotation. In *Workshop on Games and Natural Language Processing*, pages 39–43.

J. Bouta Cruz-Benito, R. Theron, F. J. Garcia-Penalvo, and E. P. Lucas. 2015. Discovering usage behaviors and engagement in an educational virtual world. *Computers in Human Behavior*, 47:18–25.

A. Bowser, D. Hansen, Y. He, C. Boston, M. Reid, L. Gunnell, and J. Preece. 2013. Using gamification to inspire new citizen science volunteers. In *Proceedings of the First International Conference on Gameful Design, Research, and Applications, Gamification'13, NewYork,NY,USA, October*, page 18–25. Association for Computing Machinery.

V. Braun and V. Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11:589–597.

V. Braun and V. Clarke. 2021. One size fits all? what counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, 18:328–352.

P. Bui, G. Rodríguez-Aflecht, B. Brezovszky, M. M. Hannula-Sormunen, S. Laato, and E. Lehtinen. 2020. Understanding students' game experiences throughout the developmental process of the number navigation game. *Educational Technology Research and Development*, 68:2395–2421.

N. Chaiko, S. Sepanta, and R. Zamparelli. 2022. The "actors challenge" project: Collecting data on intonation profiles via a web game. In *Proceedings of the 9th Workshop on Games and Natural Language Processing within the 13th Language Resources and Evaluation Conference*, page 49–53. European Language Resources Association.

J. Chamberlain, M. Poesio, and U. Kruschwitz. 2008. Phrase detectives: A web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics' 08)*, pages 42–49.

J. C. Chen and S. Kent. 2020. Task engagement, learner motivation and avatar identities of struggling english language learners in the 3d virtual world. *System*, 88:102168.

K. Clark and C. D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the ACL*.

K. Clark and C. D. Manning. 2016a. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of EMNLP*.

K. Clark and C. D. Manning. 2016b. Improving coreference resolution by learning entity level distributed representations. In *Proceedings of the ACL*.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

D. Dziedzic. 2016. Use of the free to play model in games with a purpose: the robocorp game case study. *Bio-Algorithms and Med-Systems*, 12:187–197.

M. Fan, S. Shi, and K. N. Truong. 2020. Practices and challenges of using think-aloud protocols in industry: An international survey. *Journal of Usability Studies*, 15(2).

K. Fort, B. Guillaume, and H. Chastant. 2014. Creating zombilingo, a game with a purpose for dependency syntax annotation. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, pages 2–6.

K. Fort, B. Guillaume, Y. Pilatte, M. Constant, and N. Lefèbvre. 2020. Rigor mortis: Annotating mwes with a gamified platform. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, page 4395–4401. European Language Resources Association.

N. Fraser. 2015. Ten things we've learned from blockly. *IEEE Blocks and Beyond Workshop (Blocks and Beyond)*.

E.R. Gouveia, M. Nascimento, C. França, P. Campos, A. Ihle, K. Przednowek, A. Marques, N. Nunes, and B.R. Gouveia. 2023. Correlates of presence in a virtual reality gamification environment for rehabilitation after musculoskeletal injury. *PRESENCE: Virtual and Augmented Reality*.

I.E. Guzman-Mendoza, M. Mirna, C.-R. Héctor, and M. Jezreel. 2021. Designing a player-persona for gamification learning experiences. *CEUR Workshop Proceedings*.

J. Hamari and L. Keronen. 2017. Why do people play games? a meta-analysis. *International Journal of Information Management*, 37(3):125–141.

B. Hladká, J. Mírovský, and P. Schlesinger. 2009. Play the language: Play coreference. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 209–212.

D. Jurgens and R. Navigli. 2014. It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *Transactions of the Association for Computational Linguistics*, 2:449–464.

A. Kawrykow, G. Roumanis, A. Kam, D. Kwak, C. Leung, C. Wu, E. Zarour, L. Sarmenta, M. Blanchette, and J. Waldispuhl. 2012. Phylo: A citizen science approach for improving multiple sequence alignment. *PLOS ONE*.

D. Kicikoglu, R. Bartle, J. Chamberlain, and M. Poesio. 2019. Wormingo: a 'true gamification' approach to anaphoric annotation. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*, pages 1–7.

Y.E. Kim, E.M. Schmidt, and L. Emelle. 2008. Moodswings: A collaborative game for music mood label collection. In *International Society for Music Information Retrieval Conference*.

R. Kleffner, J. Flatten, A. Leaver-Fay, D. Baker, J. B. Siegel, F. Khatib, and S. Cooper. 2017. Foldit standalone: a video game-derived protein structure manipulation interface using rosetta. *Bioinformatics*, 33:2765–2767.

T. Kohler, J. Fueller, K. Matzler, D. Stieger, and J. Füller. 2011. Co-creation in virtual worlds: The design of the user experience. *MIS Quarterly*, page 773–788.

L. Kougioumtzian, K. El Raheb, A. Katifori, and M. Roussou. 2022. Blazing fire or breezy wind? a story-driven playful experience for annotating dance movement. *Frontiers in Computer Science*, 4:957274.

M. Krause, A. Takhtamysheva, M. Wittstock, and R. Malaka. 2010. Frontiers of a paradigm: exploring human computation with digital games. In *Proceedings of the acm sigkdd workshop on human computation*, pages 22–25.

M. Lafourcade, A. Joubert, and N. Le Brun. 2015. *Games with a Purpose (GWAPS)*. John Wiley and Sons.

T. Laine and R. Lindberg. 2020. Designing engaging games for education: A systematic literature review on game motivators and design principles. *IEEE Transactions on Learning Technologies*, 13(4):804–821.

K. Lee, L. He, M. Lewis, and L. Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of EMNLP*.

K. Lee, L. He, M. Lewis, and L. Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of ACL*.

T. Y. Lee, C. Dugan, W. Geyer, T. Ratchford, J. Rasmussen, N. S. Shami, and S. Lupushor. 2013. Experiments on motivational feedback for crowdsourced workers. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 341–350.

Y. Lee and A. N. K. Chen. 2011. Usability design and psychological ownership of a virtual world. *Journal of Management Information Systems*, 28:269–308.

C. Lewis. 1982. Using the “thinking-aloud” method in cognitive interface design. *IBM TJWatson Research Center Yorktown Heights*.

C. A. Lindley, C. Sennersten, et al. 2008. Game play schemas: from player analysis to adaptive game mechanics. *International Journal of Computer Games Technology*, 2008.

C. Madge, R. Bartle, J. Chamberlain, U. Kruschwitz, and M. Poesio. 2019a. The design of a clicker game for text labelling. In *2019 IEEE Conference on Games (CoG)*, pages 1–4. IEEE.

C. Madge, R. Bartle, J. Chamberlain, U. Kruschwitz, and M. Poesio. 2019b. Incremental game mechanics applied to text annotation. In *in Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pages 545–558.

S. Martschat and M. Strube. 2015. Latent structures for coreference resolution. *Transactions of the Association for Computational Linguistics*, 3:405–418.

S. McDonald, T. Zhao, and H. M. Edwards. 2013. Dual verbal elicitation: the complementary use of concurrent and retrospective reporting within a usability test. *International Journal of Human-Computer Interaction*, 29(10):647–660.

J. A. Miller and S. Cooper. 2022. Barriers to expertise in citizen science games. In *CHI Conference on Human Factors in Computing Systems*, pages 1–25.

J. A. Miller, U. Narayan, M. Hantsbarger, S. Cooper, and M. S. El-Nasr. 2019. Expertise and engagement: re-designing citizen science games with players’ minds in mind. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*, pages 1–11.

K. Morrison, M. Jain, J. Hammer, and A. Perer. 2023. Eye into ai: Evaluating the interpretability of explainable ai techniques through a game with a purpose. In *Proceedings of the ACM on Human-Computer Interaction 7(CSCW2)*, pages 1–22.

C. Mount Cieri, J. Fiumara, and J. Wright. 2020. Using games to augment corpora for language recognition and confusability. *Interspeech*, pages 1887–1891.

S. Paun, J. Chamberlain, U. Kruschwitz, J. Yu, and M. Poesio. 2018. A probabilistic annotation model for crowdsourcing coreference. In *http://aclweb. org/anthology/D18-1000*, pages 1926–1937.

M. Poesio, J. Chamberlain, U. Kruschwitz, L. Robaldo, and L. Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, 3(1):Article 3.

M. Poesio, J. Chamberlain, S. Paun, J. Yu, A. Uma, and U. Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2019–06–02 – 2019–06–07, Minneapolis, Minnesota.

L. Poretski and A. Tang. 2022. Press a to jump: Design strategies for video game learnability. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*, pages 155, 1–26.

S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Zhang Y. 2011. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea.

Eraldo R., C´ıcero N. dos Santos Fernandes, and Ruy L. Milidi´u. 2014. Latent trees for coreference resolution. *Computational Linguistics*, 40(4):801–835.

M. Rajanen and R. Dorina. 2017. Usability benefits in gamification. In *GamiFIN Conference*.

A. R. Roberts, B. De Schutter, K. Franks, and M. E. Radina. 2019. Older adults' experiences with audiovisual virtual reality: Perceived usefulness and other factors influencing technology acceptance. *Clinical gerontologist*, 42(1):27–33.

A. Rollings and D. Morris. 1999. *Game Architecture and Design*. Paraglyph Press.

A. C. Tomé Klock, E. J. de Borba, I. Gasparini, D. Lichtnow, Pimenta M. S., and G. Rodriguez. 2017. Evaluation of usability and user experience regarding the gamification of educational systems. In *Twelfth Latin American Conference on Learning Technologies (LACLO)*, pages 1–8.

M. J. Van den Haak and M. D. T. De Jong. 2003. Exploring two methods of usability testing: concurrent versus retrospective think-aloud protocols. In *IEEE International Professional Communication Conference, 2003. IPCC 2003. Proceedings.*, pages 3–pp. IEEE.

D. Vannella, D. Jurgens, D. Scarfini, D. Toscani, and R. Navigli. 2014. Validating and extending semantic knowledge bases using video games with a purpose. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1294–1304.

N. Venhuizen, K. Evang, V. Basile, and J. Bos. 2013. Gamification for word sense labeling. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*.

L. Von Ahn. 2006. Games with a purpose. *Computer*, 39:92–94.

L. Von Ahn and L. Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326.

L. Von Ahn and L. Dabbish. 2008. Designing games with a purpose. *Communications of the ACM*, 51:58–67.

L. Von Ahn, M. Kedia, and M. Blum. 2006a. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 75–78.

L. Von Ahn, R. Liu, and M. Blum. 2006b. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 55–64.

L. Xu, E. Dhonnchadha, and M. Ward. 2022. Faoi gheasa: An adaptive game for irish language learning. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, page 133–138. Association for Computational Linguistics.