Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1740–1750
Marseille, 20-25 June 2022

Work Hard, Play Hard: Collecting Acceptability Annotations through a 3D Game

Federico Bonetti [1,2], Elisa Leonardelli[1], Daniela Trotta[3], Raffaele Guarasci[4], Sara Tonelli[1]
[1] Fondazione Bruno Kessler, Italy, [2] Università di Trento, Italy
[3] Università di Salerno, Italy, [4] ICAR–CNR, Naples, Italy
{fbonetti, eleonardelli, satonelli}@fbk.eu, dtrotta@unisa.it, raffaele.guarasci@icar.cnr.it

Abstract

Corpus-based studies on acceptability judgements have always stimulated the interest of researchers, both in theoretical and computational fields. Some approaches focused on spontaneous judgements collected through different types of tasks, others on data annotated through crowd-sourcing platforms, still others relied on expert annotated data available from the literature. The release of CoLA corpus, a large-scale corpus of sentences extracted from linguistic handbooks as examples of acceptable/non acceptable phenomena in English, has revived interest in the reliability of judgements of linguistic experts vs. non-experts. Several issues are still open. In this work, we contribute to this debate by presenting a 3D video game that was used to collect acceptability judgments on Italian sentences. We analyse the resulting annotations in terms of agreement among players and by comparing them with experts' acceptability judgments. We also discuss different game settings to assess their impact on participants' motivation and engagement. The final dataset containing 1,062 sentences, which were selected based on majority voting, is released for future research and comparisons.

Keywords: acceptability, gamification, annotation

1. Introduction

In recent years, studies on automatic assessment of acceptability have become very popular thanks to the release of the CoLA corpus (Warstadt et al., 2019), the first large-scale corpus of English acceptability, containing more than 10k sentences taken from linguistic literature, now included in the widely used GLUE benchmark (Wang et al., 2018). The corpus, whose Italian counterpart has been recently released (Trotta et al., 2021), has been developed based on the assumption that experts, i.e. linguists publishing handbooks and educational material on acceptability, have the required knowledge to define rules and examples to explain what is deemed acceptable or not in a language. In other terms, the example sentences reported in linguistic literature can be used to build a gold standard and train systems that perform acceptability judgments. Another strand of research related to acceptability, however, claims that informal collection of acceptability judgments can guarantee high annotation quality, even when annotation is crowd-sourced through Amazon Mechanical Turk (AMT). Specifically, (Sprouse, 2011) shows that crowd-sourced annotations are almost indistinguishable from data annotated in a controlled environment by university students. Other works along the same line have focused on the role of crowd-sourcing using AMT (Lau et al., 2014; Sprouse et al., 2013). However, to our knowledge, no previous work has explored the potential of gamification for acceptability annotation, which embeds the informal data collection criteria of AMT (i.e. non-expert annotation) while increasing annotators' engagement.

In this work we investigate the applicability of a 3D game to the annotation of linguistic acceptability in Italian sentences. Since our goal is to compare experts' and players' judgments, we annotate with the game a set of sentences from the ItaCoLA corpus (Trotta et al., 2021), originally extracted from linguistic handbooks. We also want to assess whether there are gameplay strategies that can positively affect players' enjoyment and motivation. Specifically, we address the following research questions:

Q1 : Is it possible to use a videogame to collect informal judgments on linguistic acceptability?

Q2 : How do the above annotations compare with experts' judgments extracted from linguistic literature?

Q3 : Which kind of strategies can be applied to obtain from players high engagement and enjoyment?

This paper describes both the adaptation of an existing game with a purpose (GWAP) to the task of acceptability annotation, and the evaluation of the data and the players' behaviour based on the contribution of around 90 participants. We also release 1,062 sentences from the ItaCoLA corpus with players' judgments (at least two acceptability annotations each),[1] to allow a direct comparison with the labels in the original dataset.

2. Related work

Acceptability judgements are still much-discussed in the literature, and many aspects are still controversial. The reliability of formal judgements extracted from the

[1] Available together with the official ItaCoLA release at `https://github.com/dhfbk/ItaCoLA-dataset`

literature and annotated by experts has been questioned (Langsford et al., 2019; Culbertson and Gross, 2009), as has the possibility of using annotators without any language training. Although the debate on the use of expert and non-expert annotators and on the effectiveness of different crowd-sourcing techniques is pervasive for all NLP annotation tasks (Snow et al., 2008; Wang et al., 2013), in acceptability judgements the problem is even more discussed. Indeed several studies comparing expert vs naïve judgements have been proposed (Dabrowska, 2010; Sprouse et al., 2013; Cho et al., 2021).

Apart from the theoretical debate around acceptability judgments, two main areas in NLP are related to this work: the creation of resources on acceptability and gamification techniques for linguistic annotation, which we summarise below.

2.1. Acceptability corpora

In recent years, the growing interest on automatic assessment of acceptability judgements driven by the release of the Corpus of Linguistic Acceptability CoLA (Warstadt et al., 2019) has shifted the topic of acceptability from a predominantly theoretical and psycholinguistic perspective to a more NLP-oriented one. CoLA is certainly not the first developed resource on acceptability, but it is rather the culmination of numerous previous works, each with different criteria, theoretical basis and method of data collection. The ongoing debate on acceptability judgements in the literature has its foundations on the still open and controversial theoretical issue on the status of syntax (Sprouse and Almeida, 2013; Lau et al., 2014) and on formal and informal data collection criteria (Culicover and Jackendoff, 2010; Gibson and Fedorenko, 2013). Among theoretically-driven datasets, (Sprouse et al., 2013) compare a random sample of 300 sentences extracted from the 'Linguistic Inquiry' informally annotated using AMT with ones collected using formal methods. Another dataset has been proposed by (Lau et al., 2014) extracting 600 sentences from the BNC (Consortium and others, 2007) with the deliberate addition of unacceptable sentences *ad-hoc* created using machine translation. The annotations were again made using AMT, since it is widely considered in the literature to be a reliable system for this type of task (Sprouse, 2011). More recently, (Marvin and Linzen, 2019) use a dataset of sentence pairs automatically built with templates in order to evaluate the behaviour of a neural model on specific syntactic phenomena. Concerning studies involving languages other than English, (Linzen and Oseki, 2018) collect data from different sources such as peer-reviewed papers, books and dissertations written in Hebrew and Japanese to evaluate informal acceptability judgments. Other studies using literature as a basis for extracting data have been conducted in Chinese (Chen et al., 2020) and French (Feldhausen and Buchczyk, 2020). A large corpus – containing around 9,600 sentences – was produced for the Swedish language (Volodina et al., 2021) exploiting language learners' data. Concerning Italian, two datasets have been released to date. The first one has been developed for the Evalita 2020 shared task on complexity and acceptability (Brunato et al., 2020) and includes acceptability scores on a 7-point Likert scale. The dataset is quite small (around 1,700 sentences) and it was built merging together different controlled corpora created for psycholinguistic purposes. Notice that this dataset is not a resource properly created for the acceptability task, but rather with the purpose to develop and evaluate methods to classify Italian sentences according to both Acceptability and Complexity. The other Italian corpus is ItaCoLA (Trotta et al., 2021), which has been developed following the same criteria as the English CoLA, and presents Boolean acceptability values. A subset of ItaCoLA, which we describe in Section 3, has been used for the annotation game presented in this paper.

2.2. Gamified Linguistic Annotation

Gamified linguistic annotation has gained traction in recent years as an alternative crowdsourcing technique. It follows from the success of some pioneering human computation games such as *ESP Game* (von Ahn and Dabbish, 2004), a multiplayer game for image annotation, and *Foldit!* (Cooper et al., 2010), a game that uses the intelligence of the crowd to predict protein structures.

Linguistic annotation is often time consuming and requires paid experts. Therefore, gamification represents an interesting, motivating and cheap alternative. It has been applied for many different tasks and to different degrees of gameplay complexity. Some of the most renowned examples of games with a purpose (GWAPs) in the field are the following: *Phrase Detectives* (Poesio et al., 2013) for anaphora resolution; *OnToGalaxy* (Krause et al., 2010) for semantic linking; *The Knowledge Towers* and *Infection* for validating and extending ontologies (Vannella et al., 2014); *Puzzle Racer* and *KaBoom!* (Jurgens and Navigli, 2014) for sense-image mapping and word sense disambiguation; *WordClicker* (Madge et al., 2019) for Part-of-Speech tagging; *Zombilingo* (Fort et al., 2014) for dependency syntax annotation, *Wordrobe* (Venhuizen et al., 2013) for word sense labeling, *Ambiguss* (Lafourcade and Brun, 2017) for word-sense disambiguation, *Wormingo* (Kicikoglu et al., 2019) for anaphoric annotation.

Many of the above works implement rewards as points, badges, leaderboards, cosmetic rewards, which are considered to be common incentives in gamification in general (Huotari and Hamari, 2012; Seaborn and Fels, 2015; Joubert, 2015). For example, *Zombilingo* uses an avatar that can be customized by making progress through the game. This approach is similar to mainstream commercial games such as *Fortnite* (Epic Games and People Can Fly, 2017). This mechanic is also present in our game, where we let players build their own character.

All of the above games use scores, which are represented by a 'School Quality' bar and player level number in our game. Contrary to how GWAPs are usually designed, our game lets players engage in tasks rather freely. It is up to them (although there is a time limit) to start annotating after exploring. In a way, this approach could be considered similar to the annotation-motivation paradigm proposed by (Kicikoglu et al., 2019), where annotation phases are alternated with more playful sessions. Another feature that is common to some of the above games, such as *Wormingo*, *Zombilingo* and *Phrase Detectives*, is a difficulty ranking system for the content to be annotated, which allows progression. This feature is however not yet present in our game.

Games like *OnToGalaxy*, *The Knowledge Towers* and *Puzzle Racer* present a full-fledged 2D environment, where players actually control an entity and have to shoot labels carried by spaceships when they are not related to a given concept. The game we present in this work tries to follow a similar approach concerning the game-like environment, but in 3D.

Regarding the impact of having limited resources, such as consumable items, while for instance (Naglé et al., 2021) have investigated the impact of collectibles on motivation in software training, there seems to be still little work on assessing the impact of limited vs. unlimited resources on task performance in games with a purpose. We address this aspect in Section 6.2.

3. Corpus Description

Our task is based on the re-annotation of a set of sentences taken from ItaCoLA, the Italian Corpus of Linguistic Acceptability (Trotta et al., 2021). The corpus was created with the purpose of representing a large number of linguistic phenomena while distinguishing between acceptable and not acceptable sentences. The methodology followed to create the corpus was similar as much as possible to the one proposed for the English CoLA in (Warstadt et al., 2019). In particular, ItaCoLA includes around 9,700 sentences from different manuals covering several linguistic phenomena.

Acceptability annotation relies on Boolean judgments as formulated by experts (i.e. the authors of the different data sources) in line with several previous works (Lawrence et al., 2000; Wagner et al., 2009; Linzen et al., 2016) to ensure robustness and simplify classification. Such sentences come from various types of linguistic publications covering four decades, which were manually transcribed and released in digital format.[2] Sources include theoretical linguistics textbooks (Graffi and Scalise, 2002; Simone and Masini, 2013) and works that focus on specific phenomena such as idiomatic expressions (Vietri, 2014), locative constructions (D'Agostino, 1983) and verb classification (Jezek, 2003). Few examples are listed in Table 1.

To perform our annotation with naïve users, we select a subset of the ItaCoLA corpus so to have 50% acceptable and 50% not acceptable sentences.

4. Game Description

4.1. High School Superhero for Acceptability Annotation

High School Superhero (henceforth HSS) is a 3D video game set in a small town that allows players to change or erase parts of sentences to annotate them. After a character creation screen, players can explore a town to perform the task in the dedicated spots. The game contains 2 different types of activities, so-called *mechanics*. In Task Mechanic 1 (Figure 1, left), players can listen to conversations happening among non-player characters and see a preview of what they are going to say. Players can then decide to change some tokens, or all of them, or leave the sentence unchanged. In this way the game collects pairs of acceptable and not acceptable examples when a sentence is left unchanged or is modified, respectively.

In Task Mechanic 2 (Figure 1, right), players erase graffiti tokens off the walls and floors of the 3D environment. In this mechanic, players can only erase tokens, which means that alternative sentences are not collected. Since players can erase an ambiguous portion of a word, as the mechanic is performed with a sponge, we consider a word annotated when 80% of its surface has been erased.

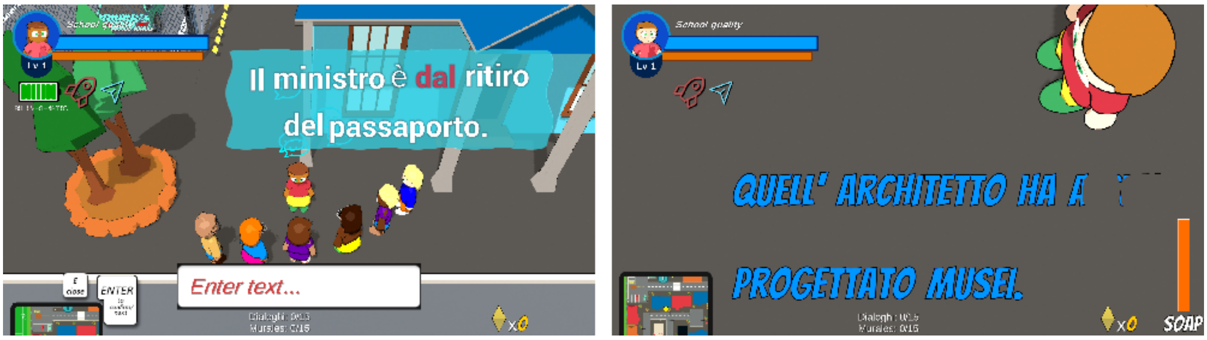
The game was first tested to collect judgments on abusive sentences (Bonetti and Tonelli, 2020). However, it was designed to accommodate different linguistic annotation tasks, therefore it has been easily adapted to the linguistic acceptability exercise. In particular, players were asked to erase sentences that are deemed unacceptable (Figure 1, right) or change the tokens in a given sentence that make a sentence unacceptable, if any (Figure 1, left).

After the character creation, where players can customize their avatar as they prefer, a brief narrative-oriented phase begins, where they also get the chance to read a tutorial and understand exactly how they are going to perform the tasks. Since this artifact is quite experimental, and presents itself as a game even though it tries to collect high quality data, administering an exhaustive tutorial concerning the task and the controls is crucial. Players were instructed to change sentences or erase them in such a way that they made sense; in the case of the graffiti, where it was only possible to erase tokens, they could erase the word(s) that made the sentence not acceptable. A couple of examples were given as part of a dialogue with an in-game character (the Professor). The following examples were given: '*Paolo ha detto che chiamerà la mela' (*Paolo said that he will call the apple*); '*Un aereo dalla decolla pista'

[2] https://github.com/dhfbk/ItaCoLA-dataset

Source	Label	Sentence
Vietri (2004)	0	*Quell'architetto ha alcuni progettato musei. (**That architect has some designed museums.*)
Graffi (1994)	1	Ho voglia di salutare Maria (*I want to greet Maria.*)
Elia et al. (1981)	0	*Il ministro è dal ritiro del passaporto. (**The minister is from passport withdrawal.*)
Simone and Masini (2013)	1	Questa donna mi ha colpito. (*This woman has impressed me.*)

Table 1: Example sentences from the ItaCoLA dataset. 1 = acceptable, 0 = not acceptable



Il ministro è dal ritiro del passaporto.

QUELL' ARCHITETTO HA A

PROGETTATO MUSEI.

Figure 1: Task Mechanic 1 (left): the player listens to a conversation and can decide to change the tokens that make the sentence unacceptable. In the example **Il ministro è dal ritiro del passaporto. (en: The minister is from passport withdrawal)*, the player is entering a new word to replace the selected one *dal (from)*. Task Mechanic 2 (right): the player sees a graffiti. The tokens that make the sentence unacceptable can be erased. In the example, the sentence **Quell'architetto ha alcuni progettato musei (en: That architect has some designed museums)* has been partially erased.

(*An airplane from the takes off runway*). In the former, *will call* or *apple* could be erased or changed. In the latter, *from the* and *off runway* could be erased or changed.

The GWAPs presented in Section 2.2 rely on common gamification mechanics such as scores and cosmetic rewards. Something that is missing from these efforts is an integration between the game narrative and the annotation task. For example, in *OnToGalaxy*, a game where players shoot unrelated labels for semantic linking, if one were to take away the space-inspired fantasy, the task could be preserved by adopting another type of narrative without consequences. With HSS we seek to have the narrative and the mechanics as integrated as possible with the task. While shooting words requires some sort of gimmick, in that it makes sense only if the markable labels are carried by enemy spaceships, erasing or changing parts of sentences is compatible with an acceptability annotation task on its own. This type of design is sometimes referred to as *intrinsic integration* (Habgood and Ainsworth, 2011), albeit in the context of educational serious games.

In addition, HSS has been designed with the goal to explore the impact of *orthogonal mechanics* on the annotation task (Bonetti and Tonelli, 2021). Orthogonal mechanics are defined as those game-like mechanics that pose some kind of challenge or hurdle for users, such as aiming, jumping or limited resources (Tuite, 2014). Also in this case, we aim at investigating the impact of orthogonal mechanics when annotating acceptability judgments by collecting feedback through a questionnaire.

4.2. Questionnaire

At the end of the session, participants had to fill out two short questionnaires. The first one was aimed at collecting basic demographics. In particular, we wanted to assess whether gender differences or geographical information can lead to differences in the annotation outcome. The second questionnaire was about the players' experience with the task. In particular, the self report was collected by means of a well documented and established questionnaire, the Intrinsic Motivation Inventory (IMI), based on self-determination theory (Ryan and Deci, 2000). Five subscales were employed and translated to Italian: *Interest/Enjoyment*, *Pressure/Tension*, *Perceived choice*, *Perceived competence*, and *Effort/Importance*. The questionnaire is assessed with a 7-point Likert scale. The first one (*Interest/Enjoyment*) is a subscale that is thought to be a direct measurement of intrinsic motivation, and contains items such as "I enjoyed doing this activity very much" and "This activity was fun to do". *Pressure/Tension* is considered a negative predictor of intrinsic motivation and contains items such as "I did not feel nervous at all while doing this" and "I was anxious while working on this task". *Perceived choice* contains items such as "I believe I had some choice about doing this activity" and "I did this activity because I wanted to". *Perceived competence* contains items such as "I think I am pretty good at this activity" and "I am satisfied with my performance at this task". Finally, *Effort/Importance* contains items such as "I put a lot of effort into this" and "I tried very hard on this activity".

We left out the *Value/Usefulness* and the *Relatedness* (or *Belonging* (Ostrow and Heffernan, 2018)) subscales as we deemed them not relevant for the task at hand. The former contains items such as "I would be willing to do this again because it has some value to me" and "I think this is an important activity". These items seem to be more suitable for activities directed at improving the well being of the user or of a group of people. The latter contains items such as "I felt like I could really trust this person" and "It is likely that this person and I could become friends if we interacted a lot" and was therefore not suitable for a single player game.

5. Participants and procedure

5.1. Demographics

Participants (N=134) were recruited from the authors' research facilities and universities. The two questionnaires were administered when people had annotated at least 15 sentences from Mechanic 1 and 15 sentences from Mechanic 2. About 30% of the participants did not arrive to (or did not take) the final demographic questionnaire (40 people), and therefore no demographic information is available about them. Among those who completed the demographic questionnaire (N=94), 58.5% were females; 36.1% were males; 1% were non-binary and 4.2% did not specify. Regarding the age, 70.2% were aged 18-24; 23.4% were 25-34; 4.2% were 35-44 and finally 2.1% were aged 45-54. The regions of provenance were so divided: 57% were from Apulia; 12.7% from Veneto; 8.5% from Lombardy; 4.2% from Campania. The rest came from Tuscany, Basilicata and Trentino-South Tyrol (3.1% each), Lazio, Friuli and Calabria (2.1% each), and Liguria (1%).

5.2. Experimental design

Our annotation task has two goals: first, to assess the differences between naïve and expert annotators. To this purpose, players annotate with HSS sentences that have been previously judged as acceptable or not by linguistics scholars, so that a comparison between annotations can be carried out. Second, to understand which strategies work best to increase players' engagement. Participants were therefore randomly assigned to two different groups, playing two slightly different versions of the game: participants in one group had to replenish their resources in order to continue annotating, while in the other group resources were unlimited. In particular, Mechanic 1 (changing the dialogues) could be performed only when the energy in the battery was greater than zero (Figure 1, left, top-left corner), and in Mechanic 2 (erasing the graffiti) the erasing could be performed only when the soap bar value was greater than zero (Figure 1, right, bottom-right corner). In the other version, the one with unlimited resources, anybody could annotate without restrictions. The resource-limited version of the game is more similar to commercial games, since many meaningful gameplay actions are often subject to the availability of resources or power-ups (bullets or bombs, mana and stamina are some of the main examples). It is also worth noting that the version with limited resources encouraged exploration more, since resources could be bought in exchange for crystals that were found around the town.

Even in the condition where resources were limited, participants could annotate all the tokens they wanted if they were willing to go and gather the required resources. After seeing a graffiti it was possible to leave it for later if one thought the soap would not be sufficient; the same goes for the dialogues, since one could close a conversation without going any further with the annotation and restart from the same sentence.

Beside *limited and unlimited resource*, another independent variable considered in our experiments is *compulsoriness*. Indeed, some players were university students who had to carry out this activity as part of an academic course, while others were volunteers who did not receive any kind of compensation.

The experiment ended after 30 sentences were annotated (15 from Mechanic 1 and 15 from Mechanic 2) but participants were allowed to continue annotating. Every participant was thus assigned 30 mandatory sentences, which were necessary to reach the final questionnaires. The sentences were presented randomly with respect to the acceptability judgement given by expert linguists, so that participants had equal chance to annotate an acceptable or unacceptable sentence, regardless of it being in Mechanic 1 or 2. Annotating more than 30 sentences generated annotation overlaps between annotators. In this way it was possible to get a fair amount of annotations and a certain number of overlapping judgements.

6. Analysis

6.1. Naïve vs. Expert Annotations

A total of 4,686 annotations have been carried out by 134 participants. Every annotator evaluated an average of 35 sentences (SD=±22). A total of 2,465 unique sentences were annotated and each sentence received on average 1.9 annotations (SD=±1). Since in both Mechanics annotators were asked to modify the sentences that they considered unacceptable (either by erasing or correcting them), a sentence was considered 'acceptable' if the annotator left it unchanged. Conversely, any change to the sentence corresponded to an 'unacceptable' label. In total 2,390 labels were 'not acceptable' and 2,296 'acceptable'.

Formulating the task as a sentence modification activity makes it more time-consuming than just asking players for a boolean acceptable/not acceptable judgment. This in turn leads to a relatively low number of annotations per sentence (1.9 on average) compared to crowd-sourced tasks using Amazon Mechanical Turk, which usually collects 3 or 5 judgments per sentence. However, this allowed us obtain also manually modified versions of unacceptable sentences, which could be of interest, for instance, to implement or evaluate sentence correction systems.

Since several sentences were annotated by multiple players, we first compute inter-annotator agreement as a way to analyse if the task or the annotation setting present any issue. The sentences that received more than one annotation have a number of judgments between 2 and 6. We therefore compute Krippendorff's alpha (Hayes and Krippendorff, 2007), as this measure allows to handle different numbers of multiple raters. For calculation of K-α we considered all sentences with more than one annotation (1,359 sentences, 3,578 annotations), obtaining a K-α value of 0.672. Through a manual inspection of the disagreement cases, we noticed that few players seemed to annotate the sentences in an inconsistent way, randomly skipping sentences that are clearly not acceptable or erasing acceptable ones. On the other hand, annotators who seemed to have understood the assignment and took annotation seriously were consistent in their judgments, proving that the task is rather well-defined. This suggests also that effective ways to identify and discard *spammers* should be introduced, for example by checking annotation accuracy against gold standard sentences (with a filtering mechanism similar to Amazon Mechanical Turk) or by applying competence estimation techniques to annotators (Hovy et al., 2013).

In light of the relatively low value of K-α, we decided to include in our final dataset only the sentences with at least two annotations and for which a majority vote existed (e.g: sentences with two or four even annotations were discarded). This resulted in a smaller dataset of 1,062 sentences labelled as acceptable or not, which should contain only the most reliable judgments.

Against this smaller subset, which may reflect rather reliably the genuine annotation choices of naïve users, we compare the judgments provided by linguistics scholars. Indeed, all the sentences have been extracted from ItaCoLA, and therefore present also the acceptability annotations originally assigned by the authors of the source textbooks. The agreement between naïve annotators and linguists is 0.623 Cohen's kappa (Cohen, 1960), which corresponds to a moderate agreement. Table 2 displays the confusion matrix comparing the sentences annotated by experts and by naïve annotators (players). This analysis shows that the two types of annotators tend to disagree equally on the two classes. Contrary to the expectations, expert annotators do not seem to be generally stricter in their acceptability judgments. Indeed, experts consider not acceptable 545 sentences, while naïve users 561.

		Naïve		
		not Acc.	Acc.	Total
Expert	not Acc.	453	92	545
	Acc.	108	409	517
Total		561	501	1062

Table 2: Confusion matrix comparing expert and naïve annotations

If we consider the cases of disagreement, some general patterns can be observed, see examples in Table 3. Naïve users tend to consider marked syntactic structures, such as left and right dislocated sentences or hanging topics, as not acceptable, while they are typical examples of acceptable structures in linguistics textbooks. The same can be observed with nested relative clauses. On the contrary, naïve annotators tend to see as acceptable sentences with a slightly unusual wording, with expressions that are similar to the standard ones. For instance, the sentence 'Il treno si è un po' fermato' (*The train stopped a bit*) could be interpreted as *The train stopped for a while* in informal language, and has been therefore seen as acceptable by the game players. On the other hand, the two groups of annotators both judge as unacceptable clear cases of ungrammatical structures, for example missing subject–verb agreement. As far as it is beyond the scope of this work, note that disagreement between annotators in the judgments is biased by the binary forced-choice method. A comparison that takes into account multiple levels, i.e. naïve vs expert annotators, binary vs gradient scale for judgements - as suggested by some work (Sprouse et al., 2018; Lau et al., 2017; Lappin and Lau, 2018) could produce different results in terms of agreement.

6.2. Analysis of players' behaviour

The second analysis we want to carry out is aimed at assessing which strategies can be used to obtain high engagement and annotation quality from players, in order to answer Q3. We therefore focus on the set of players that completed both the demographic questionnaire and the one on Intrinsic Motivation Inventory (IMI) (see Section 4.2), i.e. 90 respondents in total. As explained above, the respondents include participants for which playing the game was a requirement for taking a class and those who were volunteers. We also distinguish between players having unlimited resources and those with limited battery energy and soap. We therefore compare the following groups: non compulsory/unlimited (N=17), compulsory/unlimited (N=26), non compulsory/limited (N=21) and compulsory/limited (N=26). The analysis was carried out by running two-way ANOVAs. Prior to running the model, we checked for heteroskedasticity with the Levene's Test to compare multiple sample variances, which revealed no significant differences of variance across

Expert	Naïve	Sentence
1	0	Questo libro, non lo avevo mai letto. (*This book, I had not read it before.*)
1	0	Accuratamente non mi pare che sia stato fatto, questo lavoro. (*Accurately I don't think it has been done, this job*)
1	0	La gente che va all'Università che ama la fisica otterrà il laboratorio. (*People who attend University who love Physics will get a lab.*)
1	0	Che libro dice che il professore ha raccomandato di leggere? (*Which book does he say that the professor recommended to read?*)
0	1	Il treno si è un po' fermato. (*The train stopped a bit.*)

Table 3: Example sentences with different judgments between expert and naïve annotators. 1 = acceptable, 0 = not acceptable

the subscale values. Results are reported in Table 4, showing the effect of each of the five IMI subscales on the different annotator groups. Significance across groups was found in *Interest/Enjoyment*, *Perceived Choice* and *Pressure/Tension*.

Source	df	Mean Sq	F	p
Interest/Enjoyment				
Resources	1	9.56	4.18	**.044**
Compuls.	1	1.17	.51	.47
Interaction	1	11.4	4.98	**.028**
Residuals	86	2.3		
Perceived choice				
Resources	1	15.28	108	**<.01**
Compuls.	1	125.93	83	**<.001**
Interaction	1	10.38	6.8	**.01**
Residuals	86	1.5		
Perceived competence				
Resources	1	.077	.038	.84
Compuls.	1	.6	.28	.59
Interaction	1	.68	.33	.56
Residuals	86	2		
Effort/Importance				
Resources	1	1.52	.076	.38
Compuls.	1	5	2.51	.11
Interaction	1	.0058	.002	.96
Residuals	86	1.9		
Pressure/Tension				
Resources	1	2.54	1.48	.38
Compuls.	1	26.67	15.56	**<.001**
Interaction	1	3.33	1.94	.16
Residuals	86	1.7		

Table 4: Summary of the two-way ANOVAs run on the IMI results. Five subscales were used. The significant outcomes (main effects and interactions) regard Interest/Enjoyment (main effect of Resources and interaction effect), Perceived choice (main effect of both Resources and Compulsoriness, and interaction effect) and Pressure/Tension (main effect of Compulsory).

People with limited resources seem to have reported slightly higher values in the *Interest/Enjoyment* subscale (main effect $F(1,86)=4.18$, $p<.05$). There is also an interaction effect ($F(1,86)=4.98$, $p<.05$) which indicates that participants whose task was compulsory tended to benefit, motivationally speaking, from the limited resources (unlimited: $M=3.32$, $SD=\pm1.45$; limited: $M=4.57$, $SD=\pm1.8$). This was confirmed by a post-hoc Tukey's test ($p<.05$). This could be due to additional objectives (such as replenishing the resources and exploring) being added to the simple objective of annotating unacceptable sentences. Regarding *Perceived Choice*, as one may expect, people whose task was compulsory reported significantly lower values ($F(1,86)=83$, $p<.001$). An interaction between Resources and Compulsoriness reveals that limited resources tended to contribute to the feeling of choice for those in the Compulsory group with limited resources (unlimited: $M=2.64$, $SD=\pm1.24$; limited: $M=3.92$, $SD=\pm1.28$, $F(1,86)=6.8$, $p<.05$). The Tukey's post-hoc test on this last difference revealed significance at $p<.01$. Since the residuals of the ANOVA deviated from normality according to the Shapiro-Wilk test, we performed a Kruskal-Wallis test with two two-sample Wilcoxon (rank sum) tests, which confirmed the significance of the two main effects.[3]

Finally, there is a main effect of Compulsoriness on *Pressure/Tension*. Participants who were in the compulsory groups reported significantly higher levels of pressure ($F(1,86)=15.5$, $p<.001$). It is also worth noting that people with limited resources reported on average lower levels of pressure in the Compulsory condition, albeit without significance.[4] A detail of the results obtained for the three subscales with significant differences is plotted in Figure 2.

Our results suggest that adding resources to be bought in exchange for collectibles in order to carry out the

[3] Non-parametric test results for the Choice subscale: Kruskal-Wallis: $p<.001$; Wilcoxon test on main effect for Compulsoriness: $p<.001$; Wilcoxon test on main effect for Resources: $p<.05$

[4] Again, since the residuals deviated from normality, we performed a non-parametric test on the Pressure/Tension subscale, which confirmed the result with $p<.001$. for the main effect of Compulsoriness.

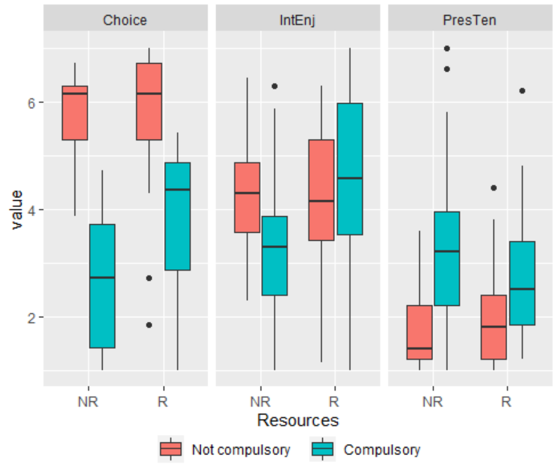


Figure 2: A boxplot of the three IMI subscales that yielded significant differences among groups: Perceived Choice, Interest/Enjoyment, Pressure/Tension.

annotation task can be beneficial to players' engagement.[5] This finding is in line with the results presented in (Naglé et al., 2021). Although the above finding does not apply to people who performed the task voluntarily, we did not find significant opposite trends in this respect either, and therefore we conclude that limited resources did not seem to do any harm to motivation. On the other hand, people who had to do the task compulsorily seemed to benefit from the limited resources significantly. This is probably due to in-game collectibles providing at least an additional playful objective to the compulsory task. Based on these findings, it follows that collectibles, limited resources and exploration increase the engagement value of the GWAP or at least do not yield negative effects.

7. Conclusions

This paper details how a 3D GWAP has been adapted to collect acceptability judgments from players. We describe both the game and the annotation process, showing what are the main differences between naïve and expert annotators, and which strategies are most effective to improve user engagement and attitude. We also release the annotated corpus in Italian, so to enable further comparisons with ItaCoLA.

A first analysis of inter-annotator agreement among players shows that a lack of quality control strategies integrated in the game can be detrimental to annotation quality. In the next game version, intermediate checks based on gold standard sentences or on the analysis of annotators' reliability should be introduced. Also checking whether there are patterns of disagreement among annotators may be useful to select only reliable raters. In our case, retaining only the sentences with a majority vote mitigated the problem but reduced significantly the size of the final corpus.

Comparing naïve and expert annotators provided interesting insights: while the two groups tend to agree on the most obvious cases of acceptability related to grammaticality, some syntactic structures are considered less acceptable in the game setting. We refer for example to marked structures such as dislocated sentences, hanging topics and nested relative clauses. Indeed, the typical made-up examples present in linguistics handbooks to illustrate the above constructions may seem wrong simply because they are not very frequent in standard language.

Concerning players' behaviour, we observe that introducing limited resources in the game (which encouraged the collection of gems needed to buy them) increases enjoyment and perceived autonomy when the activity is compulsory, probably because these strategies make the GWAP more similar to commercial video games and set more gameful objectives for players. On the other hand, recruiting participants by making the task mandatory for a class increases players' tension, making them feel under pressure, which is the contrary of what we would like to achieve using a GWAP for linguistic annotation.

In the future we aim at collecting more judgments, particularly for the sentences that were discarded because they were annotated only once. We will also extend our participants' analysis considering demographic information such as self-declared gender, age and geographical provenance. Furthermore, we plan to perform classification experiments by comparing the performance of an acceptability classifier when trained on ItaCoLA and on our novel dataset. Concerning annotator's judgements, it might be an interesting future development to compare binary judgments with other ones collected using a continuous scale. In general, the reliability of judgements collection is still under debate and its limitations are well known in the literature, i.e. there is not yet agreement on a rigorous formal method for collecting and evaluate acceptability ratings. However, the possibility to create and test a model of gradient acceptability could be a challenging target for the future.

8. Acknowledgements

We thank all participants that played with the game to annotate the dataset. Part of this work has been funded by the KID_ACTIONS REC-AG project (n. 101005518) on "Kick-off preventIng and responDing to children and AdolesCenT cyberbullyIng through innovative mOnitoring and educatioNal technologieS".

9. Bibliographical References

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

[5] Although collectible crystals were present in all versions of the game, they were *relevant* only in the condition with limited resources.

Bonetti, F. and Tonelli, S. (2021). Measuring orthogonal mechanics in linguistic annotation games. *Proceedings of the ACM on Human-Computer Interaction*, 5(CHI PLAY):1–16, October.

Brunato, D., Chesi, C., Dell'Orletta, F., Montemagni, S., Venturi, G., and Zamparelli, R. (2020). Accompl-it @ EVALITA2020: Overview of the Acceptability & Complexity Evaluation Task for Italian. In Valerio Basile, et al., editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Chen, Z., Xu, Y., and Xie, Z. (2020). Assessing introspective linguistic judgments quantitatively: The case of the syntax of chinese. *Journal of East Asian Linguistics*, 29(3):311–336.

Cho, J., Li, Y., and Shields, R. (2021). Gradient acceptability between naïve and expert linguistic intuitions.

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

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z., and players, F. (2010). Predicting Protein Structures with a Multiplayer Online Game. *Nature*, 466(7307):756–760, August.

Culbertson, J. and Gross, S. (2009). Are linguists better subjects? *The British journal for the philosophy of science*, 60(4):721–736.

Culicover, P. W. and Jackendoff, R. (2010). Quantitative methods alone are not enough: Response to gibson and fedorenko. *Trends in Cognitive Sciences*, 6(14):234–235.

Dabrowska, E. (2010). Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*, 27(1):1–23.

D'Agostino, E. (1983). *Lessico e sintassi delle costruzioni locative: materiali per la didattica dell'italiano*. Liguori.

Elia, A., Martinelli, M., and d'Agostino, E. (1981). *Lessico e strutture sintattiche: introduzione alla sintassi del verbo italiano*. Liguori Napoli.

Epic Games and People Can Fly. (2017). Fortnite. Game [Windows]. Epic Games, Cary (NC), USA.

Feldhausen, I. and Buchczyk, S. (2020). Testing the reliability of acceptability judgments for subjunctive obviation in French. In *Going romance 2020.*

Fort, K., Guillaume, B., and Chastant, H. (2014). Creating zombilingo, a game with a purpose for dependency syntax annotation. In *Proceedings of the First International Workshop on Gamification for Information Retrieval - GamifIR '14*, pages 2–6, Amsterdam, The Netherlands. ACM Press.

Gibson, E. and Fedorenko, E. (2013). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1-2):88–124.

Graffi, G. and Scalise, S. (2002). *Le lingue e il linguaggio. Introduzione alla linguistica*. Il Mulino, Bologna, Italy.

Graffi, G. (1994). *Le strutture del linguaggio. Sintassi*. Il Mulino, Bologna, Italy.

Habgood, M. P. J. and Ainsworth, S. E. (2011). Motivating children to learn effectively: Exploring the value of intrinsic integration in educational games. *Journal of the Learning Sciences*, 20(2):169–206, April.

Hayes, A. F. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. (2013). Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Huotari, K. and Hamari, J. (2012). Defining gamification: a service marketing perspective. In *Proceeding of the 16th International Academic MindTrek Conference on - MindTrek '12*, page 17, Tampere, Finland. ACM Press.

Jezek, E. (2003). *Classi di verbi tra semantica e sintassi*. Edizioni ETS, Pisa, Italy.

Joubert, Alain; Lafourcade, M. L. B. N. (2015). *Games with a Purpose (GWAPS)*. Focus series (London England); Cognitive science and knowledge management series. Wiley-ISTE, 1 edition.

Jurgens, D. and Navigli, R. (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, December.

Kicikoglu, D., Bartle, R., Chamberlain, J., and Poesio, M. (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, San Luis Obispo California USA, August. ACM.

Krause, M., Takhtamysheva, A., Wittstock, M., and Malaka, R. (2010). Frontiers of a paradigm: Exploring human computation with digital games. In *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '10*, pages 22–25, Washington DC. ACM Press.

Lafourcade, M. and Brun, N. L. (2017). Ambiguss, a game for building a sense annotated corpus for French. In *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*.

Langsford, S., Hendrickson, A. T., Perfors, A., Kennedy, L., and Navarro, D. (2019). A systematic comparison and reliability analysis of formal measures of sentence acceptability.

Lappin, S. and Lau, J. H. (2018). Gradient probabilistic models vs categorical grammars: A reply to sprouse et al.(2018). *Science of Language*.

Lau, J. H., Clark, A., and Lappin, S. (2014). Measuring gradience in speakers' grammaticality judgements. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Lau, J. H., Clark, A., and Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.

Lawrence, S., Giles, C. L., and Fong, S. (2000). Natural language grammatical inference with recurrent neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 12(1):126–140.

Linzen, T. and Oseki, Y. (2018). The reliability of acceptability judgments across languages. *Glossa: a journal of general linguistics*, 3(1).

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

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

Marvin, R. and Linzen, T. (2019). Targeted syntactic evaluation of language models. *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 373–374.

Naglé, T., Bateman, S., and Birk, M. V. (2021). Pathfinder: The behavioural and motivational effects of collectibles in gamified software training. *Proceedings of the ACM on Human-Computer Interaction*, 5(CHI PLAY):1–23, October.

Ostrow, K. S. and Heffernan, N. T. (2018). Testing the validity and reliability of intrinsic motivation inventory subscales within assistments. In Carolyn Penstein Rosé, et al., editors, *Artificial Intelligence in Education*, volume 10947, pages 381–394. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2013). Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation. *ACM Transactions on Interactive Intelligent Systems*, 3(1):1–44, April.

Ryan, R. M. and Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, page 11.

Seaborn, K. and Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, 74:14–31, February.

Simone, R. and Masini, F. (2013). *Nuovi fondamenti di linguistica*. McGraw Hill.

Snow, R., O'connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Sprouse, J. and Almeida, D. (2013). The empirical status of data in syntax: A reply to gibson and fedorenko. *Language and Cognitive Processes*, 28(3):222–228.

Sprouse, J., Schütze, C. T., and Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, 134:219–248.

Sprouse, J., Yankama, B., Indurkhya, S., Fong, S., and Berwick, R. C. (2018). Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *The Linguistic Review*, 35(3):575–599.

Sprouse, J. (2011). A validation of amazon mechanical turk for the collection of acceptability judgments in linguistic theory. *Behavior research methods*, 43(1):155–167.

Tuite, K. (2014). GWAPs: Games with a problem. In *Proceedings of the 9th International Conference on the Foundations of Digital Games*.

Vannella, D., Jurgens, D., Scarfini, D., Toscani, D., and Navigli, R. (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, Baltimore, Maryland. Association for Computational Linguistics.

Venhuizen, N. J., Evang, K., Basile, V., and Bos, J. (2013). Gamification for word sense labeling. In *Proceedings of the International Conference on Computational Semantics (IWCS)*, pages 397–403.

Vietri, S. (2004). *Lessico-grammatica dell'italiano. Metodi, descrizioni e applicazioni*. UTET Università.

Vietri, S. (2014). *Idiomatic constructions in Italian: a lexicon-grammar approach*, volume 31. John Benjamins Publishing Company.

von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*, pages 319–326, Vienna, Austria. ACM Press.

Wagner, J., Foster, J., van Genabith, J., et al. (2009). Judging grammaticality: Experiments in sentence classification. *Calico Journal*, 26(3):474–490.

Wang, A., Hoang, C. D., and Kan, M.-Y. (2013). Perspectives on crowdsourcing annotations for natural language processing. *Lang. Resour. Eval.*, 47(1):9–31, mar.

Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural network acceptability judgments. *Transac-*

tions of the Association for Computational Linguistics, 7:625–641, March.

10. Language Resource References

Consortium, B. et al. (2007). British national corpus. *Oxford Text Archive Core Collection*.

Lau, J. H., Clark, A., and Lappin, S. (2014). Measuring gradience in speakers' grammaticality judgements. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Trotta, D., Guarasci, R., Leonardelli, E., and Tonelli, S. (2021). Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2929–2940, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Volodina, E., Mohammed, Y. A., and Klezl, J. (2021). DaLAJ – a dataset for linguistic acceptability judgments for Swedish. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 28–37, Online, May. LiU Electronic Press.

Wang, Alex and Singh, Amanpreet and Michael, Julian and Hill, Felix and Levy, Omer and Bowman, Samuel. (2018). *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*. Association for Computational Linguistics.