Proceedings of the LREC 2020 Workshop Games and Natural Language Processing, pages 79–84
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Aggregation Driven Progression for GWAPs

Doruk Kicikoglu,♣ Richard Bartle,♠ Silviu Paun,♣ Jon Chamberlain,♠ Massimo Poesio♣
o.d.kicikoglu@qmul.ac.uk, rabartle@essex.ac.uk, s.paun@qmul.ac.uk, jchamb@essex.ac.uk, m.poesio@qmul.ac.uk
♣Queen Mary Univ. Of London, United Kingdom
♠University Of Essex, United Kingdom

Abstract

As the use of Games-With-A-Purpose (GWAPs) broadens, their annotation schemes have increased in complexity. The types of annotations required within NLP are an example of labelling that can involve varying complexity of annotations. Assigning more complex tasks to more skilled players through a progression mechanism can achieve higher accuracy in the collected data while acting as a motivating factor that rewards the more skilled players. In this paper, we present the progression technique implemented in Wormingo , an NLP GWAP that currently includes two layers of task complexity. For the experiment, we have implemented four different progression scenarios on 192 players and compared the accuracy and engagement achieved with each scenario.

Keywords: GWAPs, player progression, Bayesian models, coreference annotation, citizen science

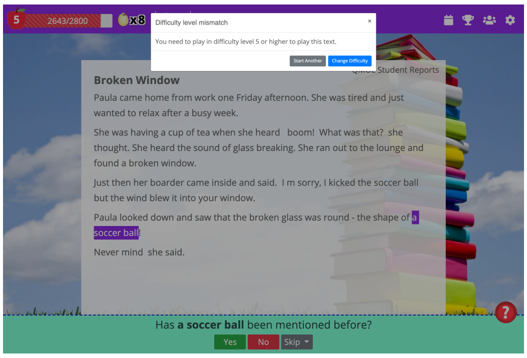
1. Introduction

The first GWAPs focused on simple tasks varying from text deciphering to image or sound labelling (von Ahn and Dabbish, 2004; Lafourcade et al., 2015; Barrington et al., 2009). Such GWAPs did not require their players to progress to more advanced tasks. However, modern GWAPs collecting more complex judgments, as in NLP, may require players to carry out annotations of varying complexity that may be harder to teach to entry-level players (Poesio et al., 2013). Such GWAPs may benefit from the practice, widely adopted within the gaming industry (Koster and Wright, 2004), of introducing a player to simpler tasks and proceeding to the more complicated ones once they have proven successful on the initial tasks. Such skill progression achieves higher motivation and engagement as the players are kept within flow (Csikszentmihalyi, 1991), meaning they face challenges corresponding to their improving competence. GWAPs can achieve a similar affect with this approach. In addition, this type of progression increases the quality of the data produced as players are assigned with more complicated tasks, only after they have reached a sufficient understanding of the annotation tasks within the system (Madge et al., 2019).

The fact that GWAP players vary in terms of competence makes it mandatory to assess the players by comparing to golden data, and proceed only when they reach a certain level of accuracy (Ipeirotis and Gabrilovich, 2015; Madge et al., 2019; Fort et al., 2014; Chamberlain et al., 2008). In addition to many GWAPs that utilize this method, Phrase Detectives and Zombilingo also implement progression techniques that assess the player accuracy based on the types of tasks that they are performing. These GWAPs include different types of tasks which vary in complexity. Players begin with simpler tasks, then move on to more complicated annotation tasks once they reached a certain level of success during the assessment period.

In addition to aligning the player progression along task complexity, another axis can be the difficulty of the labels; that can be defined as the difficulty of a label compared to the other labels within the same task i.e. some spans might be more ambiguous in Phrase Detectives, hence may be more difficult to resolve; creating more disagreement among the players. In a system where labels are identified and ranked by their difficulty, players can be assigned with more difficult tasks once they prove successful on the easier ones. Tile Attack and Quizz implement this technique, where players are assigned with labels matching their competence level (Ipeirotis and Gabrilovich, 2015; Madge et al., 2019).

Wormingo implements both of these approaches of progression. As players progress, they can advance to both more difficult documents (difficulty progression) and more complicated tasks (task progression). For difficulty progression, the documents in Wormingo are manually labelled into 5 levels of difficulty ranging from letter A to E. The documents in level A are considered as the easiest in terms of comprehension, while those in level E are the most difficult, that may include more sophisticated vocabulary or more complicated sentence structure. Wormingo uses a level-up mechanism which lets players reach higher levels (currently up to level 16) after collecting score points awarded for annotations. Players can play more difficult documents, only after reaching higher player levels (Figure 1).



Difficulty level mismatch

You need to play in difficulty level 5 or higher to play this text.

Broken Window

Paula came home from work one Friday afternoon. She was tired and just wanted to relax after a busy week.

She was having a cup of tea when she heard boom! What was that? she thought. She heard the sound of glass breaking. She ran out to the lounge and found a broken window.

Just then her boarder came inside and said. I m sorry, I kicked the soccer ball but the wind blew it into your window.

Paula looked down and saw that the broken glass was round - the shape of a soccer ball!

Never mind she said.

Has **a soccer ball** been mentioned before?

Figure 1: Player attempts to access a document that is too difficult for their level

Level-up mechanisms are widely used within games (Zichermann and Cunningham, 2011). Although they are proven effective for rewarding commitment to the game, they do not necessarily indicate that the player is more competent. A player who performs poorly in terms of accuracy can simply hoard points by playing longer and still reach the next player level. Therefore, when assessing the players' competence for more advanced tasks, their annotation accuracy can be a better indicator rather than the points they managed to hoard.

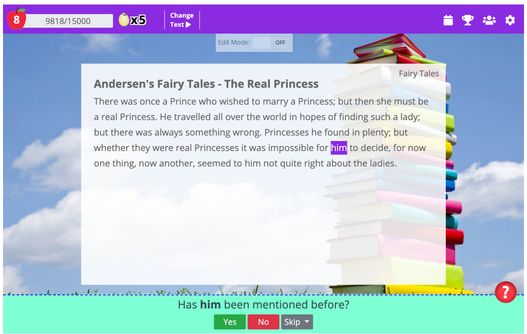
Comparing the players' annotations to the gold or aggregated data yields the player accuracy. However, cases in Phrase Detectives show that higher numbers of players can agree on a wrong annotation while fewer number of skilled players might contrarily have given the correct answer for a label (Paun et al., 2018). Relying solely on the number of annotators can be misleading in such cases. Therefore, Mention Pair Annotations model (MPA) builds a confidence-based model. MPA generates confidence scores for annotations, and players, via Bayesian models with the players' annotation accuracy taken into consideration. Players who have higher accuracy gain a higher confidence score from a range between 0 and 1. During data aggregation, the annotations of players with higher confidence scores are evaluated with higher weight. MPA also generates separate player confidence scores for each task, evaluating players' performance on individual tasks. This model overcomes the aforementioned problem and produces confidence ratings both for the aggregated data and the players. Wormingo uses the player confidence outcome when assessing their competence to progress to more complicated tasks.

2. Background

2.1. Annotation Tasks

Wormingo currently includes two types of annotation tasks, discourse-new and non-referring. The earlier versions of Wormingo already included the discourse-new task (Figure 2), which asks players if a label in the task has been mentioned before (Kicikoglu et al., 2019). In the current version of Wormingo , the non-referring task has been implemented as the second and more advanced task.

In the discourse-new task, the players annotate coreference chains. The game asks the players to annotate a label, such as the label "him" illustrated with purple colour in Figure 2. The player clicks "No" if this label was not mentioned in the text before, or "Yes" if it was mentioned. After clicking "Yes", clusters of phrases that we call "markables" are highlighted with colour yellow (Figure 3). The player chooses which of the markables that the label refers to in this interface.

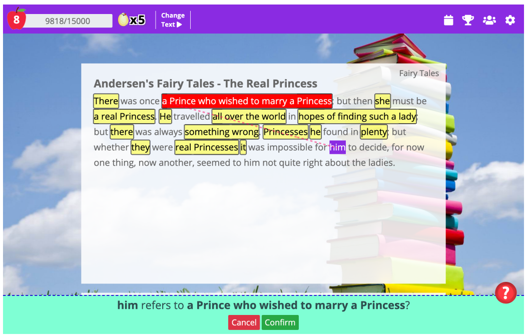


Andersen's Fairy Tales - The Real Princess

There was once a Prince who wished to marry a Princess; but then she must be a real Princess. He travelled all over the world in hopes of finding such a lady; but there was always something wrong. Princesses he found in plenty; but whether they were real Princesses it was impossible for him to decide, for now one thing, now another, seemed to him not quite right about the ladies.

Has **him** been mentioned before?

Figure 2: Discourse New Annotation Interface



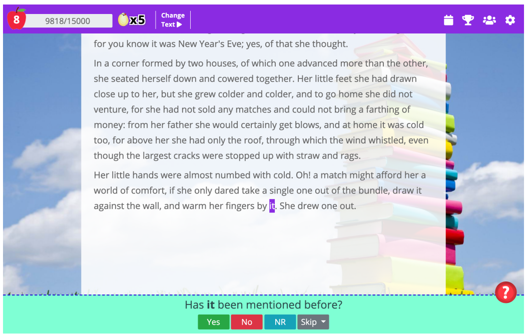
Andersen's Fairy Tales - The Real Princess

There was once a Prince who wished to marry a Princess; but then she must be a real Princess. He travelled all over the world in hopes of finding such a lady; but there was always something wrong. Princesses he found in plenty; but whether they were real Princesses it was impossible for him to decide, for now one thing, now another, seemed to him not quite right about the ladies.

him refers to **a Prince who wished to marry a Princess**?

Figure 3: Discourse New Interface - Marking coreference

In the non-referring task, labels such as "it" in the sentences "It is raining", "It is 3 o'clock" do not refer to a real object. Such occurrences should be labelled as non-referring (Chamberlain et al., 2009). However, this adds an extra layer to the discourse-new task implemented in the earlier versions of Wormingo , because in addition to the possibility of being a non-referring label, an occurrence of the word "it" can be a part of a coreference chain as well; such as in "I had a pizza, it was good!". Therefore, non-referring is considered as a more complicated task laid on top of the discourse-new task, as it includes the complexity of the discourse-new task with the non-referring option added on top. On the interface, non-referring task uses the same interface layout as the discourse-new task, but an additional "NR" button is added. Players who click this button annotate the given label as non-referring (Figure 4). Non-referring cases occur on expletive words "it" and "there", so only the labels with these string values were asked in the non-referring tasks.



for you know it was New Year's Eve; yes, of that she thought.

In a corner formed by two houses, of which one advanced more than the other, she seated herself down and cowered together. Her little feet she had drawn close up to her, but she grew colder and colder, and to go home she did not venture, for she had not sold any matches and could not bring a farthing of money: from her father she would certainly get blows, and at home it was cold too, for above her she had only the roof, through which the wind whistled, even though the largest cracks were stopped up with straw and rags.

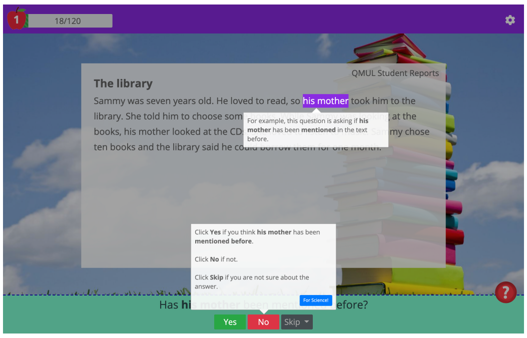
Her little hands were almost numbed with cold. Oh! a match might afford her a world of comfort, if she only dared take a single one out of the bundle, draw it against the wall, and warm her fingers by it. She drew one out.

Has **it** been mentioned before?

Figure 4: Non-referring Annotation Interface

2.2. Tutorials

Players are taught about the discourse-new task on their first annotation. This is done through freezing the interface and showing the player a message that explains the discourse new task. First an example whose correct answer is discourse-new (has not been mentioned before) is shown and the player can only continue by clicking the "No" button, which labels the annotation as discourse-new (Figure 5). On the following annotation, players are similarly shown a label that has been mentioned in the text before. Players can continue only by linking the label to one of its antecedents and clicking the "Confirm" button on this interface (Figure 6).



The library

Sammy was seven years old. He loved to read, so his mother took him to the library. She told him to choose som at the books, his mother looked at the CD my chose ten books and the library said he co

For example, this question is asking if **his mother** has been **mentioned** in the text before.

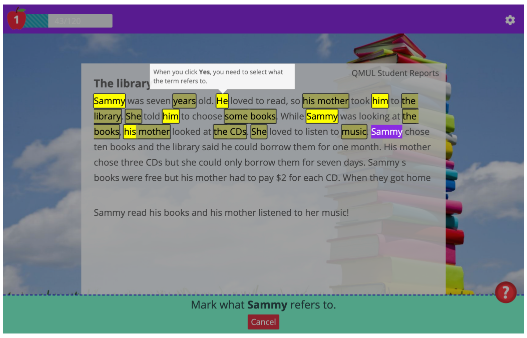
Click **Yes** if you think **his mother** has been **mentioned before**.

Click **No** if not.

Click **Skip** if you are not sure about the answer.

Has **hi** efore?

Figure 5: Tutorial for a discourse new label



The library

When you click **Yes**, you need to select what the term refers to.

Sammy was seven years old. He loved to read, so his mother took him to the library. She told him to choose some books. While Sammy was looking at the books, his mother looked at the CDs. She loved to listen to music. Sammy chose ten books and the library said he could borrow them for one month. His mother chose three CDs but she could only borrow them for seven days. Sammy s books were free but his mother had to pay $2 for each CD. When they got home

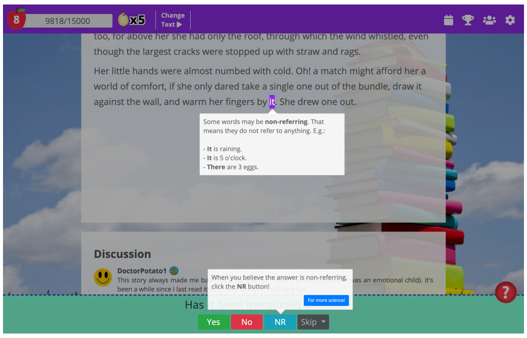
Sammy read his books and his mother listened to her music!

Mark what **Sammy** refers to.

Figure 6: Tutorial for marking coreference

The case-selection algorithm of Wormingo chooses the next documents and labels to represent to the players from a selection of available items, where incomplete labels that received at least one annotation are prioritized (Chen et al., 2010). Labels that have received less than 7 annotations are considered incomplete.

Once a player has been assessed to qualify to the non-referring task, the case-selection algorithm starts including expletive labels as well. Expletive labels gain higher priority scores; however the final case selection happens with a random selection where higher priority items gain higher probability -meaning an item with less probability still has a chance to appear as the next task depending on the generated random value. The player may also qualify to the non-referring task while playing a document that contains no expletive expressions at all. Thus, the player may not immediately encounter a non-referring task after qualifying to the non-referring tasks. Once they do encounter a non-referring task for a first time, the tutorial interface appears (Figure 7) and the players are explained about the non-referring task and introduced with the "NR" button that allows the players to annotate labels as non-referring.



too, for above her she had only the roof, through which the wind whistled, even though the largest cracks were stopped up with straw and rags.

Her little hands were almost numbed with cold. Oh! a match might afford her a world of comfort, if she only dared take a single one out of the bundle, draw it against the wall, and warm her fingers by it. She drew one out.

Some words may be **non-referring**. That means they do not refer to anything. E.g.:

- **It** is raining.
- **It** is 5 o'clock.
- **There** are 3 eggs.

Discussion

DoctorPotato1
This story always made me ba as an emotional child). It's been a while since I last read it

When you believe the answer is non-referring, click the **NR** button!

Has

Figure 7: Tutorial for non-referring tasks

2.3. Methodology

In the experiment, we divided the players into 4 groups. Each group needed to accomplish a different scenario to advance to the non-referring tasks. Group A needed to earn 350 score points, which corresponds to reaching level 3 and an average of 16.77 discourse-new annotations (players gain 25 points for each correct discourse-new annotation and 50 points for each correct non-referring annotation). The accuracy of this group was not considered when evaluating; hoarding enough points was sufficient for Group A to qualify to the non-referring task.

Groups B, C and D needed to pass the 350 point barrier like Group A. On top of this, they needed to achieve certain MPA confidence scores for their discourse-new annotations. Group B needed to reach 0.8 MPA confidence score in order to progress. Group C needed to reach 0.85 confidence score and Group D needed to reach 0.9. Comparing Group A to the other groups allowed observing the difference between assessing players based solely on their score, versus assessing players based on their accuracy. Comparing Group B, C and D allowed observing how the value of the qualification threshold affects the data produced.

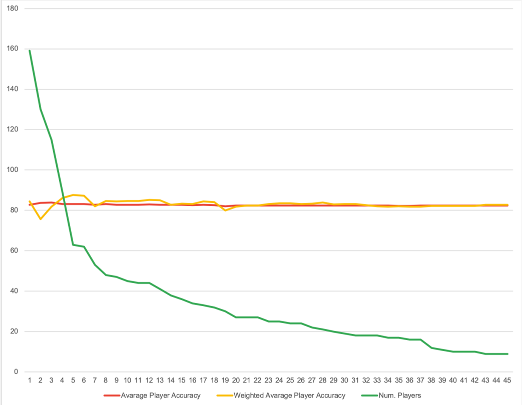


Figure 8: Average discourse new accuracy of players by number of annotations

Prior to the experiment, players were evaluated based on their discourse-new annotation accuracy over time. The yellow line in Figure 8 displays the average accuracy of players, varying by the number of annotations they have done. The red line is their weighted accuracy; calculated by comparing players' accuracy on each document to the average accuracy of all players on the respective document. The average weighted accuracy can vary on the first few annotations, but after players' 10th annotations, it reaches a plateau around 84% accuracy. Therefore, we took 10 annotations as the threshold -the number of discourse-new annotations a player must complete before being progressing to the non-referring tasks. Players who did annotations fewer than this threshold were not assigned to any of the observation group. The players who reached 350 points and did at least 10 annotations were assigned to an observation group.

3. Results

We analyze the data produced between 07 Feb 2020 and 17 Mar 2020. During this period, 192 Wormingo players did at least 1 annotation. The players came from the subreddits that we have posted on reddit.com and university e-mail groups with interest towards Computer Science and games.

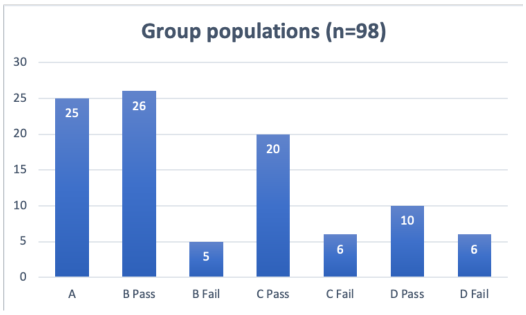


Figure 9: Number of players per group

Out of the 192 players, 98 completed the qualification requirements and were therefore assigned to a observation group. Figure 9 shows the number of players in each group. Groups B Pass, C Pass and D Pass are the groups of players who were originally in groups B, C and D respectively and have accomplished progression to the non-referring tasks. Similarly, groups B Fail, C Fail and D Fail contain the players who were in groups B, C and D respectively but failed to advance to the next task.

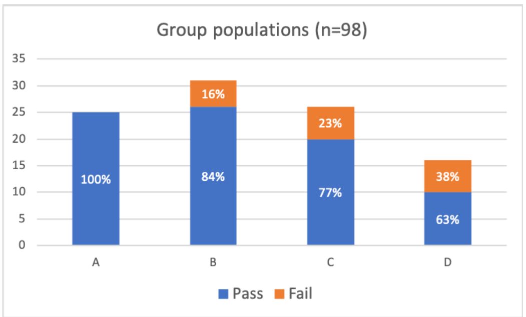


Figure 10: Pass/Fail percentages per group

Figure 10 shows each group's ratio of players who passed or failed progression to the non-referring task. The ratio of players increase as expected from Group B towards D; as the threshold for progression also increases towards this direction.

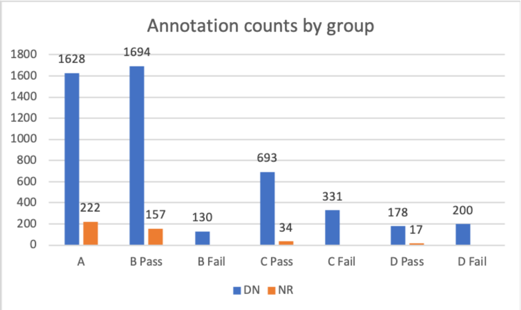


Figure 11: Total annotation counts per group

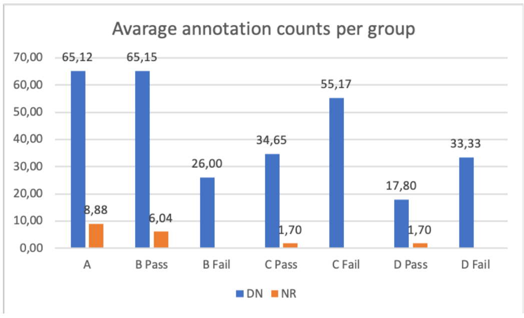


Figure 12: Average annotation counts per player

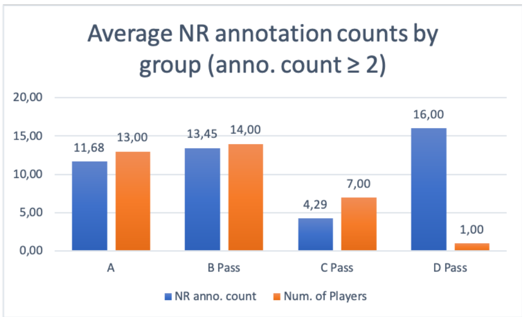


Figure 13: Average number of NR annotations and number of players who have done at least 2 NR annotations

Figures 11 and 12 display annotation counts per group and average annotations done by players within each group. Figure 12 includes players who have qualified to the NR task but have not done any non-referring annotations (since players may not immediately come across NR tasks after they qualify), hence the average annotation counts appear low. Figure 13 provides more meaningful average scores, as it displays values for players who have done at least 2 annotations. Groups A and B Pass contribute significantly higher number of annotations (DN and NR) in both total and average per player.

Figure 14 shows the groups' average accuracy and MPA confidence scores, wherein no significant difference in terms of NR accuracy is observed. However a significant difference is observed in D Pass group's NR MPA confi-

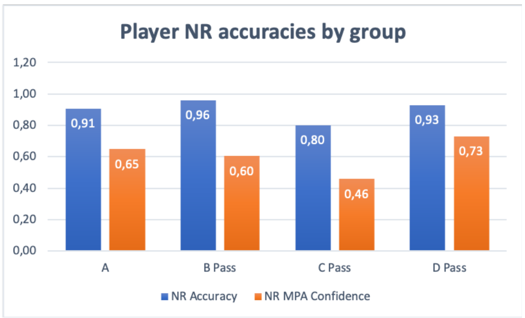


Figure 14: Average non-referring accuracy and MPA confidence scores per group

dence value (p=0.01). Although it might seem like a good strategy to set the qualification threshold to D Pass group's value, 0.9, this would potentially lead to generation of too small data, as D Pass group has only generated 17 NR annotations. B-Pass group however generated much more data (157 annotation) with an average of 0.60 confidence.

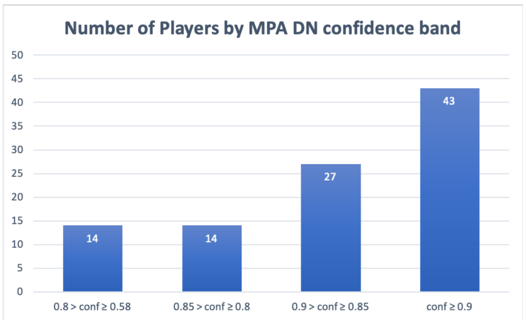


Figure 15: Number of players within each band of NR MPA confidence scores

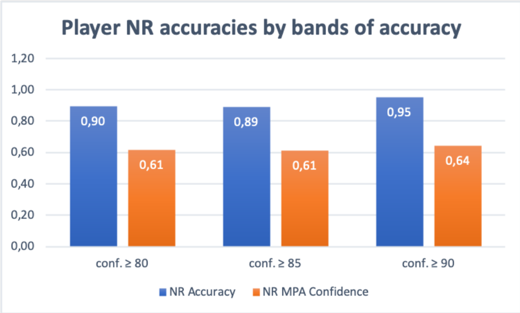


Figure 16: Non-referring accuracy and MPA confidence scores for each band of NR MPA confidence

Figure 15, 16 and 17 groups all players by their DN MPA confidence scores, instead of their observation groups. The bands "conf. $\geq$ 80", "conf. $\geq$ 85" and "conf. $\geq$ 90" are players whose DN confidence scores were higher than 0.8, 0.85 and 0.9 respectively and they are not exclusive of each other. We observe that a majority of players score higher than 0.85 DN MPA confidence in Figure 15. 43% of players score higher than 0.9 while 71% score higher than 0.85.

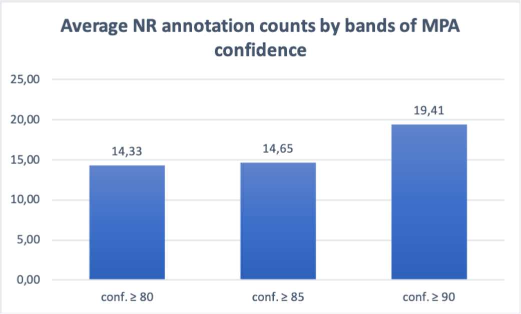


Figure 17: Average Non-referring annotation counts for each band of NR MPA confidence

We do not observe significant difference in terms of non-referring task competence between bands "conf. $\geq$ 80" and "conf. $\geq$ 85" bands (Figure 16). A slight increase is observed in the "conf. $\geq$ 90" band, however we do not have yet sufficient evidence to conclude that the threshold should be set to 0.9. Players in "conf. $\geq$ 90" band do produce more NR annotations per player (Figure 16), however setting the threshold at this level would rule out 57% of players who perform sufficiently well in terms of accuracy at the lower levels (Paun et al., 2018). We hope that future studies with more players, more data, and more levels of complexity can could provide more definitive results.

4. Discussion

In this paper, we have tested 4 different scenarios of skills progression in Wormingo. The fact that the players have voluntarily come to the game rather than for a paid reward, assures more relevance of this data to the general GWAP audience. However, the few number of participants that arrived within the limited time hinders the accuracy of our measurements, leaving room for future research on the area, possibly with more advanced tasks added.

Players who score high on discourse new tasks also achieve high accuracy on non-referring tasks. This fact is encouraging, as it supports the claim that allowing only competent players to do more complicated tasks produces cleaner data. However, this comes with a cost. Setting a threshold too high will hinder the players who have the potential to score adequately on the more complicated tasks. Setting it too low pollutes the produced data. The results show that players can perform higher accuracy on more advanced tasks, if they have were sufficiently trained on the preceding tasks. An optimal threshold that will neither rule out skilled annotators nor pollute the data can be calculated based upon the players' performance on the initial tasks.

5. Acknowledgements

This research was supported in part by the DALI project, ERC Grant 695662, in part by the EPSRC CDT in Intelligent Games and Game Intelligence (IGGI), EP/L015846/1.

6. References

Barrington, L., O'Malley, D., Turnbull, D., and Lanckriet, G. (2009). User-centered design of a social game to tag music. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '09, pages 7–10, New York, NY, USA. ACM.

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008). Phrase detectives: A web-based collaborative annotation game. 01.

Chamberlain, J., Kruschwitz, U., and Poesio, M. (2009). Constructing an anaphorically annotated corpus with non-experts: Assessing the quality of collaborative annotations. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, People's Web '09, page 57–62, USA. Association for Computational Linguistics.

Chen, L.-J., Wang, B.-C., and Chen, K.-T. (2010). The design of puzzle selection strategies for gwap systems. *Concurrency and Computation: Practice and Experience*, 22(7):890–908.

Csikszentmihalyi, M. (1991). *Flow: The Psychology of Optimal Experience*. Harper Perennial, New York, NY, March.

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, page 2–6, New York, NY, USA. Association for Computing Machinery.

Ipeirotis, P. G. and Gabrilovich, E. (2015). Quizz: Targeted crowdsourcing with a billion (potential) users. *CoRR*, abs/1506.01062.

Kicikoglu, D., Bartle, R., Chamberlain, J., and Poesio, M. (2019). Wormingo: a 'true gamification' approach to anaphoric annotation. In *FDG '19*.

Koster, R. and Wright, W. (2004). *A Theory of Fun for Game Design*. Paraglyph Press.

Lafourcade, M., Joubert, A., and Brun, N. L. (2015). *Games with a Purpose (GWAPS) (Focus Series in Cognitive Science and Knowledge Management)*. Wiley-ISTE.

Madge, C., Yu, J., Kruschwitz, U., Paun, S., and Poesio, M. (2019). Progression in a language annotation game with a purpose. In *FDG '19*.

Paun, S., Chamberlain, J., Kruschwitz, U., Yu, J., and Poesio, M. (2018). A probabilistic annotation model for crowdsourcing coreference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1926–1937, Brussels, Belgium, October-November. Association for Computational Linguistics.

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

von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, page 319–326, New York, NY, USA. Association for Computing Machinery.

Zichermann, G. and Cunningham, C. (2011). *Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps*. O'Reilly Media, Inc., 1st edition.