Testing TileAttack with Three Key Audiences

Chris Madge, Massimo Poesio, Udo Kruschwitz, Jon Chamberlain
Queen Mary University London, University Of Essex
{c.j.madge, m.poesio}@qmul.ac.uk {udo, jchamb}@essex.ac.uk

Abstract

The Game-With-A-Purpose (GWAP) approach has shown some success and promise in language resource collection. However, player recruitment and accuracy can be challenging. In this work, *TileAttack*, a GWAP designed to gather annotations for text segmentation, is presented to the online linguistic community, an indie gaming community and the crowdsourcing community. We evaluate the results of this experiment both through traditional accuracy measures and adapted metrics from Free-to-Play games. With the addition of a tutorial, we find a high level of recall is achieved from crowdsourced non-expert workers.

Keywords: Gamification; Crowdsourcing; HCI

1. Introduction

Many Natural Language Processing (NLP) tasks require large amounts of annotated text to train statistical models, or as a gold standard to test the effectiveness of NLP systems. These are often hand-annotated contributions (Palmer et al., 2005) using annotation tools. These annotation tasks may be carried out using pre-built annotation tools such as MMAX2 (Müller and Strube, 2006), web-based crowdsourcing focused WebAnno (Yimam et al., 2013), or the wiki style web-based *GMB Explorer* (Basile et al., 2012). However, those tools are aimed at expert annotators and require some understanding on the part of the user. Willing and inexpensive experts can be difficult to recruit. This process can be time consuming, expensive and tedious. Consequently, this requirement for annotated data remains an obstacle to progression for some NLP tasks. One proven method of reducing the time to gather the annotations is crowdsourcing (Snow et al., 2008). However, this doesn't scale very well. When attempting to build large corpora gamification can be cheaper (Poesio et al., 2013), provide more accurate results and better contributor engagement (Lee et al., 2013).

In this work, we look at gathering **mentions**. These are candidate entities for co-refereerence that are usually detected in a co-reference pipeline in a step often referred to as Mention Detection. They are typically noun phrases, pronouns and named entities. Historically, the task of mention detection was rarely considered in isolation, but rather as a step in part of a pipeline for co-reference resolution (Peng et al., 2015). A rule-based approach, (e.g. pick all noun phrases (Haghighi and Klein, 2010)) was generally preferred with such systems usually aiming for high recall and compromise on precision, placing more confidence/importance on the co-reference resolution step (Kummerfeld et al., 2011) and being satisfied that incorrectly identified mentions will simply remain singletons which can be removed in post processing (Lee et al., 2011). However, this approach can result in a propagation of errors with singletons then being incorrectly identified as co-referent, particularly in the case of pleonastic entities (Lee et al., 2017). It has been pointed out by multiple researchers that this is a very important step for overall co-reference quality (Stoyanov et al., 2009; Hacioglu et al., 2005; Zhekova and Kübler, 2010). Recently, systems are now once again looking at machine learning approaches with the mention detection step being considered in isolation (Lee et al., 2017; Nguyen et al., 2016). This area is still identified as an area of challenge, particularly in under resourced languages (Soraluze et al., 2012) or domains, like biomedicine (Kim et al., 2011).

Games-with-a-Purpose (GWAPs) harness human effort as a side effect of playing a game (Von Ahn and Dabbish, 2008). GWAPs have been successful in many applications attracting large numbers of users to label datasets and solve real world problems (Lafourcade et al., 2015). Examples include *The ESP Game*, in which by playing, players contribute image labels (Von Ahn and Dabbish, 2004), and *FoldIt*, in which players solve protein-structure prediction problems (Cooper et al., 2010). In contrast, gamification has been described as "the use of game design elements in non-game contexts" (Deterding et al., 2011). Gamification has been very effective in motivating text labelling. For example, *Phrase Detectives* has been particularly effective in motivating participation in gathering anaphoric annotations (Poesio et al., 2013). However, there are limited examples of GWAPs for NLP. Creating a GWAP that produces annotations as a side effect, rather than applying gamification to motivate annotation, presents a greater challenge. The former requires mapping the task completely into a game, whilst the latter typically adds a layer of game-like themes and carefully selected motivational game mechanics. In exchange for this additional challenge, GWAPs have the potential for much higher player engagement.

One of the goals of gamified solutions is to provide a positive and engaging user experience. Designing an interface for an application can present multiple challenges. This is particularly evident in application for text annotation. Text often has complex properties which can be difficult to visualise and present in an easy to use interface. The aforementioned tools take different approaches, for example, to embedded and overlapping annotations. There is no standardised and accepted interface for text annotation tools. Borrowing ideas from game interfaces can reduce the barriers to reach a wider audience of non-expert users. Designing for motivation carries additional complexity.

Games such as *Puzzle Racer* have demonstrated the feasibility of inexpensively creating an engaging GWAP that produces annotations. Furthermore, they report the annotations that are gathered are of a high quality and at a reduced cost compared with other methods (Jurgens and Navigli, 2014). However, such games have yet to achieve the player uptake or number of judgements comparable to GWAPs in other domains. GWAPs for annotation tasks often present additional unique challenges compared to those for image labelling and other similar tasks. For example, users can differentiate between image features easily, but not so easily with text features (Mason and Watts, 2010). The linguistic complexity of some text annotation tasks may not be immediately obvious or difficult to map into a game domain. Additionally, it may be challenging to find a representation that both entertains users and is easy to understand. *TileAttack* supports any text segmentation task with or without embeddings (e.g. noun-phrase embedding), that may be aligned, non-aligned or overlapping, making it broadly applicable to a variety of text annotation tasks including Named Entity Recognition, Information Extraction and Mention Detection.

In this work, we experiment with the GWAP *TileAttack*. [1] *TileAttack* is designed to gather mentions, a crucial step of the co-reference resolution pipeline which discovers potential referring expressions including noun-phrases and possessive pronouns (Lee et al., 2011). The following example shows the nested mentions enclosed in braces, (taken from the Phrase Detectives corpus (Chamberlain et al., 2016)) :

{A Wolf} had been gorging on {an animal {he} had killed}

In our previous work on testing game mechanics, we identified two additional important challenges with *TileAttack*: increasing player recruitment; and low annotation accuracy (Madge et al., 2017). This appears to be a challenge of effectively communicating the task to the players whilst retaining their interest. This is also a challenge in games. From studies in game design, the best approach is believed to be one that allows the player to play immediately, learning through a tutorial, without needing to read a manual (Sweetser and Wyeth, 2005). Naturally, traditional annotation tools, take a more utilitarian tool-like approach offering a manual and expecting a prior understanding of the task for which the tool will be used. *TileAttack* includes a game-like tutorial that plays similarly to an ordinary round but with more player feedback.

For Gamification and GWAPs to really achieve scale, they require communication of an arbitrarily complex task to a group of non-experts in a game setting. GWAPs are often tested against students from a department that have some interest or understanding in the task. In this experiment we ask if the current *TileAttack* is effective in the recruitment of non-experts and gathering accurate annotations with three distinct audiences: a linguistic community; a gaming community; and via crowdsourcing.

2. Related Work

The first Game With A Purpose was Von Ahn's *The ESP game*. This game was created to crowdsource image labels for web images, which may be used to train a supervised machine learning system. Human annotators play a game against a timer in which they were anonymously paired and rewarded scores for agreeing common labels to describe an image. In the interest of acquiring a comprehensive set of labels for each image, the game used a feature called *taboo words*. This resisted players contributing obvious image labels by displaying labels in a game as unavailable, once they had been contributed so many times. (Von Ahn and Dabbish, 2004)

The ESP game's design of rewarding based on agreement addresses the problem that an annotation task's latent correct labels are unknown by the system at the time the player is rewarded. Instead, given some input, it uses the agreement of multiple players output labels as a basis to determine whether points should be rewarded. This strategy has been described by Von Ahn as *output-agreement* (Von Ahn and Dabbish, 2008).

The GWAP concept was later applied in multiple fields to motivate player contribution including annotating text data for training NLP supervised learning systems. One notable example of a GWAP for text annotation is *Phrase Detectives*, in which players annotate and validate anaphora (Poesio et al., 2013). *Phrase Detectives* has gamification-like mechanics to motivate play such as points, leaderboards and levels, but also makes use of a game-like detective theme and tutorial section.

More recently, there have been increasingly game-like approaches taken (Vannella et al., 2014; Jurgens and Navigli, 2014). *Puzzle Racer* is a GWAP for image-sense annotation. Players tie images to word senses by racing through a series of gates, attempting to pass through gates that match a certain word sense (Jurgens and Navigli, 2014). Whilst a great example of a GWAP for NLP annotation, the game describes itself as "purely visual" and has a task itself that maps to images leaving the task not too far from being image labelling, rather than a typical NLP annotation task. *Puzzle Racer* recruited students incentivised by monetary prizes for top scoring players, and demonstrated a reduced cost over traditional crowdsourcing methods.

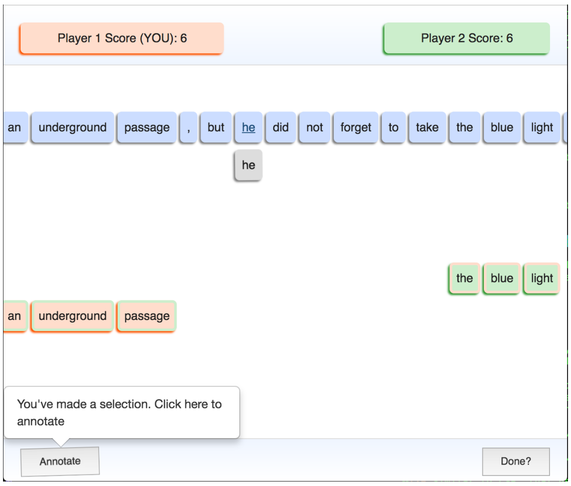
The *Wordrobe* suite of games (Bos et al., 2017) supports multiple games that perform similar annotations to that of *TileAttack* including tasks such as Named Entity Recognition and finding the referents of pronouns. However, unlike *TileAttack*, the *Wordrobe* games perform preprocessing to identify potential text segments, and then ask the player to identify which of those potential segments are correct. Whilst this fits nicely into a common game design that runs throughout the suite of games, it does constrain the players choices to potentially incorrect items. In comparison, *TileAttack* is only constrained to token boundaries.

3. TileAttack

TileAttack is a web-based two player blind game in which players are awarded points based on player agreement of the tokens they mark. The visual design of the game is inspired by *Scrabble*, with a tile like visualisation (shown in Figure 1).

In the game, players perform a text segmentation task which involves marking spans of tokens represented by tiles.

[1] https://tileattack.com



You've made a selection. Click here to annotate

Figure 1: In game screenshot from *TileAttack*

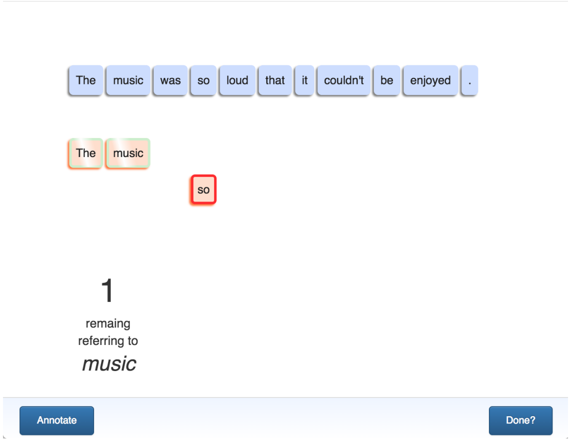


Figure 2: Tutorial screenshot from *TileAttack*

Our approach was to start with a game design that begins from as close as possible to an existing working recipe. We chose a design that is in many respects analogous to *The ESP Game*, but for text annotation. This provides the opportunity to test what lessons learned from games similar to *The ESP Game* still apply with text annotation games, and how, in the domain of text annotation, these lessons can be expanded upon. Like *The ESP Game*, we use the "output-agreement" format for the game, in which two players or agents are anonymously paired, and must produce the same output, for a given input (Von Ahn and Dabbish, 2008).

3.1. Gameplay

Following the documentation, but before the game, players are shown a two round tutorial (shown in Figure 2). For crowdsourced players, completion of this tutorial is mandatory. In the tutorial the player marks two sentences. They are informed of what entities are present in the sentence and how many mentions there are. They can incorrectly mark multiple items, which will be highlighted with a flashing red border, but will only be allowed to proceed once they have discovered all the correct items (shown by the glinting effect). They receive immediate and direct feedback to inform them of their progress.

In each game round, the player is shown a single sentence to annotate. The players can choose to select a span from the sentence by simply selecting the start and end token of the item they wish to mark using the blue selection tokens. A preview of their selection is then shown immediately below. To confirm this annotation, they may either click the preview selection or click the *Annotate* button. The annotation is then shown in the player's colour. When the two players match on a selection, the tiles for the selection in agreement are shown with a glinting effect, in the colour of the player that first annotated the tiles and a border colour of the player that agreed. The players' scores are shown at the top of the screen.

Players receive a single point for marking any item. If a marked item is agreed between the two players, the second player to have marked the item receives the number of points that there are tokens in the selection, and the first player receives double that amount. The player with the greatest number of points at the end of the round wins.

When a player has finished, they click the *Done* button, upon which they will not be able to make any more moves, but will see their opponent's moves. Their opponent is also notified they have finished and invited to click *Done* once they have finished. Once both players have clicked *Done*, the round is finished and both players are shown a round summary screen. This screen shows the moves that both players agreed on, and whether they won or lost the round.

Clicking *Continue* then takes the player to a leaderboard showing wins, losses and the current top fifteen players. From this page they may click the *Next Game* button, to start another round.

4. Experiment

4.1. Task

In this experiment we will test *TileAttack* with three separate audiences discussed below. The results of the experiment will be compared on both accuracy, and as an evaluation of player recruitment, using a set of metrics adapted from Free-to-Play games (Xicota, 2014).

In this game, players mark "mentions". These entities would normally be collected by a mention detection system and are typically used as part of larger NLP pipelines, such as relation extraction systems or co-reference resolution systems (Lee et al., 2011). To determine how successfully players are annotating the corpus, they are given sentences from the gold standard Phrase Detectives corpus (Chamberlain et al., 2016) to annotate.

4.2. Recruitment

To test *TileAttack's* ability to attract players in a gaming audience, it has been integrated with the *Kongregate* platform. [2] *Kongregate* is a popular indie game platform with an audience exposure of approximately 40,000 players.

To test *TileAttack* with a group interested in the field of linguistics, *TileAttack* has been added to a new NLP games portal. The Linguistic Data Consortium - University of

[2] https://www.kongregate.com/

Pennsylvania (LDC) project, *LingoBoingo* [3]. The LDC advertised their new portal during that month via social media channels and a newsletter. This audience is most comparable with the previous experiment, that also focused on online communities interested in linguistics (Madge et al., 2017).

To test *TileAttack's* ability to gather annotations and the benefit of the new tutorial irrespective of the game qualities, *TileAttack* has been integrated with Amazon Mechanical Turk, a crowdsourcing platform that remunerates workers on behalf of requesters to carry out small tasks. These tasks are known as *Human Intelligence Tasks* (HITs). A requester can choose from one of several Amazon Mechanical Turk templates to upload data into, or creating a custom integration. They may also specify the number of unique workers to carry out each HIT, and requirements for those workers that include qualifications. These qualifications can be awarded by the requester and serve as a flag to positively or negatively filter workers.

In our implementation, we make use of the *ExternalQuestion API*. This results in *TileAttack* being displayed in a HTML IFrame in the MTurk requester interface as a custom question. Having successfully taken part we award workers with a qualification. This satisfies the requirement of each worker participating only once, by serving as a flag on their account that is checked to prevent future tasks being displayed to them.

4.3. Experimental Design

For both *Kongregate* and LDC players, their experience is exactly as described in TileAttack's usual gameplay.

TileAttack is integrated into Amazon Mechanical Turk. Workers are shown the game documentation, with game references removed. They are then taken to the tutorial. They must complete the tutorial before they are allowed to perform the annotation task itself. Having completed the tutorial they are then asked to annotate six sentences. The core game mechanics, including scores or any evidence of a second player, are removed. The game like interface remains. Having completed the tutorial and five sentences, the participants are then remunerated for their participation (0.50 USD). Each participant is only allowed to take part once.

5. Results

Of the participants that attempted the crowdsourcing task, approximately 15% continued to completion. We take all completed games in these results, including contributions from crowdsourcing participants that did not fully complete the crowdsourcing task.

5.1. Annotation Quality

The player's annotations are compared with that from the expert annotated Phrase Detectives corpus (Chamberlain et al., 2016). This corpus provides expert annotated data as corrections to an automated pipeline. The game does not attempt to apply the corrections from the corpus. This analysis of annotation quality uses a subset of the sentences that were expert approved without requiring corrections.

	LDC	Kongregate	MTurk
Precision	60.3	16.3	72.7
Recall	55.2	17.5	66.7
F-Measure	57.6	16.9	69.5

Table 1: User-based annotation accuracy from *TileAttack* used by 3 groups

	LDC	Kongregate	MTurk
Precision	60.1	29.6	38.9
Recall	61.7	60.7	89.3
F-Measure	60.9	39.8	54.2

Table 2: Item-based annotation accuracy from *TileAttack* used by 3 groups

	LDC	Kongregate	MTurk
Games	109	20	352
Items	56	5	9
Avg. Annotations	1.8	3.6	26.4
Participants	19	7	73

Table 3: User play data from *TileAttack* used by 3 groups

As we are interested in both the design of the system and its ability to gather accurate annotations, we take two measurements of accuracy. Table 1 is the average accuracy for each user, in each game. We use this to judge how successful the system was in communicating the task to a specific audience and enabling contribution. This is comparable to the previous experiment, albeit without a tutorial, in which *TileAttack* players achieved 56.6% precision and 59.4% recall (Madge et al., 2017).

Table 2 is the average accuracy over all items (taking a union of all annotations provided by all users in that group, for that item). This allows us to judge on the whole, how successful the system is at gathering annotations. It is also important to measure both due to the way tasks are distributed to players.

Table 3 shows the number of participants for each group, the games they played, how many items were annotated and the average annotations per item. A higher number of annotations per item is very likely to raise recall. This occurs when there is a wide spread in the number of games played by the users. If a few users play many games, the system will present those users with games they have not seen before, so many individual annotations per item will be received for that group. This does impact the results shown in Table 2, but not those in Table 1. The average annotations per item are far higher for the MTurk players, as the system ensured everyone played six games, so items were more evenly annotated.

The crowdsourced players (MTurk), on average achieved a high average precision and recall. Their contribution over the items had a much higher recall, but also a much lower precision. These players were forced to take the tutorial and motivated financially. This demonstrates the system does appear to be effective in gathering annotations.

TileAttack did not appear to be successful in terms of accuracy on the *Kongregate* platform. Over a period of one

[3] https://lingoboingo.org/

month on the Kongregate platform, only 7 players chose to play *TileAttack*. They rated the game at 1.3/5 stars.

LDC players achieved precision and recall comparable to that of online linguistic groups in the previous experiment (Madge et al., 2017).

5.2. Analysis using Free-To-Play Metrics

	LDC	Kongregate	MTurk
LTJ (mention)	8	2	40
LTJ (sentence)	1	2	8
AJpP (mention)	8	2.5	16
AJpP (sentence)	1	2	2
ALP (secs)	115	180	193
MAU	19	7	73
Retention (1 day)	0	0	0

Table 4: Free-to-Play metrics for *TileAttack* used by 3 groups

Table 4 shows adapted free to play metrics for TileAttack. These metrics are defined as follows: *Lifetime Judgements (LTJ)* is the average number of items annotated per player over their lifetime of play. *Average Judgements per Player (AJpP)* is the average number of items marked per player, per gaming session. *Average Lifetime Play* is the average session length in time. *Monthly Active Users (MAU)* is the number of users in a month, the active part refers specifically to those that finished a game. *Retention and churn* is the players that were kept and lost respectively, over some time period.

6. Conclusion

TileAttack presents a fast and usable interface for sequence labelling with embedding. The system, including the design of features such as the tutorial, appear to be effective in communicating the nature of the desired annotation to non-experts. When players are financially incentivised, *TileAttack* does now achieve a high level of recall. Obviously, the strengths of a crowdsourcing approach is based on robust aggregation methods that extract the wisdom of the crowd and filter out outliers. However, here we aim to obtain high-quality annotations in the first place independent of various aggregation methods that may be added later.

In our continued progress with the *TileAttack* game, we have demonstrated, with the recent addition of a tutorial, we can reach a fair level of accuracy using non-expert annotators. If the crowdsourced participants were permitted to continue contributing, we may reasonably expect that the accuracy of their contribution may increase further with their experience.

Whilst *TileAttack* did not perform very well on *Kongregate*, this was by far the most challenging setting so far. Set alongside indie games, *TileAttack* still fails to attract the volumes of players necessary to annotate a large corpora. Now the interface and instructions appear to be satisfactory, more work must be done for *TileAttack* to work in a game setting. This will involve further testing of game design concepts and mechanics to improve both *TileAttack's* ability to attract and retain players.

7. Bibliographical References

Basile, V., Bos, J., Evang, K., and Venhuizen, N. (2012). A platform for collaborative semantic annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 92–96. Association for Computational Linguistics.

Bos, J., Basile, V., Evang, K., Venhuizen, N., and Bjerva, J. (2017). The groningen meaning bank. In Nancy Ide et al., editors, *Handbook of Linguistic Annotation*, volume 2, pages 463–496. Springer.

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2016). Phrase detectives corpus 1.0 crowdsourced anaphoric coreference. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z., et al. (2010). Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760.

Deterding, S., Dixon, D., Khaled, R., and Nacke, L. (2011). From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, pages 9–15. ACM.

Hacioglu, K., Douglas, B., and Chen, Y. (2005). Detection of entity mentions occurring in english and chinese text. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 379–386. Association for Computational Linguistics.

Haghighi, A. and Klein, D. (2010). Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393. Association for Computational Linguistics.

Jurgens, D. and Navigli, R. (2014). It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *TACL*, 2:449–464.

Kim, Y., Riloff, E., and Gilbert, N. (2011). The taming of reconcile as a biomedical coreference resolver. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 89–93. Association for Computational Linguistics.

Kummerfeld, J. K., Bansal, M., Burkett, D., and Klein, D. (2011). Mention detection: heuristics for the ontonotes annotations. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 102–106. Association for Computational Linguistics.

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

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning:*

Shared Task, pages 28–34. Association for Computational Linguistics.

Lee, T. Y., Dugan, C., Geyer, W., Ratchford, T., Rasmussen, J. C., Shami, N. S., and Lupushor, S. (2013). Experiments on motivational feedback for crowdsourced workers. In *ICWSM*.

Lee, H., Surdeanu, M., and Jurafsky, D. (2017). A scaffolding approach to coreference resolution integrating statistical and rule-based models. *Natural Language Engineering*, pages 1–30.

Madge, C., Chamberlain, J., Kruschwitz, U., and Poesio, M. (2017). Experiment-driven development of a gwap for marking segments in text. In *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '17 Extended Abstracts, pages 397–404, New York, NY, USA. ACM.

Mason, W. and Watts, D. J. (2010). Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter*, 11(2):100–108.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with mmax2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, 3:197–214.

Nguyen, T. H., Sil, A., Dinu, G., and Florian, R. (2016). Toward mention detection robustness with recurrent neural networks. *arXiv preprint arXiv:1602.07749*.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Peng, H., Chang, K.-W., and Roth, D. (2015). A joint framework for coreference resolution and mention head detection. In *CoNLL*, volume 51, page 12.

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

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 conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.

Soraluze, A., Arregi, O., Arregi, X., Ceberio, K., and De Ilarraza, A. D. (2012). Mention detection: First steps in the development of a basque coreference resolution system. In *KONVENS*, pages 128–136.

Stoyanov, V., Gilbert, N., Cardie, C., and Riloff, E. (2009). Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 656–664. Association for Computational Linguistics.

Sweetser, P. and Wyeth, P. (2005). Gameflow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)*, 3(3):3–3.

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 *ACL (1)*, pages 1294–1304.

Von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *SIGCHI*, pages 319–326. ACM.

Von Ahn, L. and Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, 51(8):58–67.

Xicota, D. (2014). Free to play and its Key Performance Indicators.

Yimam, S. M., Gurevych, I., de Castilho, R. E., and Biemann, C. (2013). Webanno: A flexible, web-based and visually supported system for distributed annotations. In *ACL (Conference System Demonstrations)*, pages 1–6.

Zhekova, D. and Kübler, S. (2010). Ubiu: A language-independent system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 96–99. Association for Computational Linguistics.