It's All Fun and Games until Someone Annotates: Video Games with a Purpose for Linguistic Annotation

David Jurgens and **Roberto Navigli**
Department of Computer Science
Sapienza University of Rome
{jurgens,navigli}@di.uniroma1.it

Abstract

Annotated data is prerequisite for many NLP applications. Acquiring large-scale annotated corpora is a major bottleneck, requiring significant time and resources. Recent work has proposed turning annotation into a game to increase its appeal and lower its cost; however, current games are largely text-based and closely resemble traditional annotation tasks. We propose a new linguistic annotation paradigm that produces annotations from playing graphical video games. The effectiveness of this design is demonstrated using two video games: one to create a mapping from WordNet senses to images, and a second game that performs Word Sense Disambiguation. Both games produce accurate results. The first game yields annotation quality equal to that of experts and a cost reduction of 73% over equivalent crowdsourcing; the second game provides a 16.3% improvement in accuracy over current state-of-the-art sense disambiguation games with WordNet.

1 Introduction

Nearly all of Natural Language Processing (NLP) depends on annotated examples, either for training systems or for evaluating their quality. Typically, annotations are created by linguistic experts or trained annotators. However, such effort is often very time- and cost-intensive, and as a result creating large-scale annotated datasets remains a long-standing bottleneck for many areas of NLP.

As an alternative to requiring expert-based annotations, many studies used untrained, online workers, commonly known as crowdsourcing. When successful, crowdsourcing enables gathering annotations at scale; however, its performance is still limited by (1) the difficulty of expressing the annotation task as a simply-understood task suitable for the layman, (2) the cost of collecting many annotations, and (3) the tediousness of the task, which can fail to attract workers. Therefore, several groups have proposed an alternate annotation method using games: an annotation task is converted into a game which, as a result of game play, produces annotations (Pe-Than et al., 2012; Chamberlain et al., 2013). Turning an annotation task into a Game with a Purpose (GWAP) has been shown to lead to better quality results and higher worker engagement (Lee et al., 2013), thanks to the annotators being stimulated by the playful component. Furthermore, because games may appeal to a different group of people than crowdsourcing, they provide a complementary channel for attracting new annotators.

Within NLP, gamified annotation tasks include anaphora resolution (Hladká et al., 2009; Poesio et al., 2013), paraphrasing (Chklovski and Gil, 2005), term associations (Artignan et al., 2009) and disambiguation (Seemakurty et al., 2010; Venhuizen et al., 2013). The games' interfaces typically incorporate common game elements such as scores, leaderboards, or difficulty levels. However, the game itself remains largely text-based, with a strong resemblance to a traditional annotation task, and little resemblance to games most people actively play.

In the current work, we propose a radical shift in NLP-focused GWAP design, building *graphical*, dynamic games that achieve the same result as traditional annotation. Rather than embellish an annota-

Transactions of the Association for Computational Linguistics, 2 (2014) 449–463. Action Editor: Mirella Lapata.
Submitted 10/2013; Revised 03/2014; Revised 08/2014; Published 10/2014.

tion task with game elements, we start from a video game that is playable alone and build the task into the game as a central component. By focusing on the game aspect, players are presented with a more familiar task, which leads to higher engagement. Furthermore, the video game interface can potentially attract more interest from the large percentage of the populace who play video games.

In two video games, we demonstrate how certain linguistic annotation tasks can be effectively represented as video games. The first video game, Puzzle Racer, produces a mapping between images and WordNet senses (Fellbaum, 1998), thereby creating a large-scale library of visual analogs of concepts. While resources such as ImageNet (Deng et al., 2009) provide a partial sense-image mapping, they are limited to only a few thousand concrete noun senses, whereas Puzzle Racer annotates all parts of speech and both concrete and abstract senses. Furthermore, Puzzle Racer's output enables new visual games for tasks using word senses such as Word Sense Disambiguation, frame detection, and selectional preference acquisition. The second game, Ka-boom!, performs Word Sense Disambiguation (WSD) to identify the meaning of a word in context by players interacting with pictures. Sense annotation is regarded to be one of the most challenging NLP annotation tasks (Fellbaum et al., 1998; Edmonds and Kilgarriff, 2002; Palmer et al., 2007; Artstein and Poesio, 2008), so we view it as a challenging application for testing the limits of visual NLP games.

Our work provides the following four contributions. First, we present a new game-centric design methodology for NLP games with a purpose. Second, we demonstrate with the first game that video games can produce linguistic annotations equal in quality to those of experts and at a cost reduction from gathering the same annotations via crowdsourcing; with the second game we show that video games provide a statistically significant performance improvement over a current state-of-the-art non-video game with a purpose for sense annotation. Third, we release both games as a platform for other researchers to use in building new games and for annotating new data. Fourth, we provide multiple resources produced by the games: (1) an image library mapped to noun, verb, and adjective WordNet senses, consisting of 19,073 images across 443 senses, (2) a set of associated word labels for most images, (3) sense annotations as a distribution over images and senses, and (4) mappings between word senses and related Web queries.

2 Related Work

Games with a Purpose Multiple works have proposed linguistic annotation-based games with a purpose for tasks such as anaphora resolution (Hladká et al., 2009; Poesio et al., 2013), paraphrasing (Chklovski and Gil, 2005), term associations (Artignan et al., 2009; Lafourcade and Joubert, 2010; Vannella et al., 2014), acquiring common sense knowledge (Kuo et al., 2009; Herdağdelen and Baroni, 2012), and WSD (Chklovski and Mihalcea, 2002; Seemakurty et al., 2010; Venhuizen et al., 2013). Notably, all of these linguistic games have players primarily interacting with text, in contrast to other highly successful games with a purpose such as Foldit (Cooper et al., 2010), in which players fold protein sequences, and the ESP game (von Ahn and Dabbish, 2004), where players label images with words.

Most similar to our games are Wordrobe (Venhuizen et al., 2013) and Jinx (Seemakurty et al., 2010), which perform WSD, and The Knowledge Towers (Vannella et al., 2014), which associates images with senses. Wordrobe asks players to disambiguate nouns and verbs using multiple choice questions where options are sense definitions and disambiguation is limited to terms with at most five senses, a limitation that does not exist in our games. Jinx uses two players who both have to independently provide lexical substitutes of an ambiguous word and are then scored on the basis of their shared substitutes. While Jinx has a more game-like feel, producing annotations from the substitutes is non-trivial and requires looking for locality of the substitutes in the WordNet graph.

In contrast to Wordrobe and Jinx, we provide a game-centric design methodology for the seamless integration of the annotation task into a video game with dynamic, graphical elements.

The Knowledge Towers (TKT) is a video game for validating the associations between images and word senses in BabelNet (Navigli and Ponzetto, 2012) and associating each of the senses with new images acquired from a Web query of one of the sense's lemmas. To perform the annotation, players are shown a word and its definition and then asked to retrieve pictures matching that definition during game play.

In contrast, our Puzzle Racer game is purely visual and does not require players to read definitions, instead showing picture examples, increasing its video game-like quality. Furthermore, Puzzle Racer is tested on nouns, verbs, and adjectives whereas TKT is only applicable to annotate nouns since it relies on the BabelNet knowledge base to acquire its initial set of image-sense associations, which contains images only for nouns.

Image Libraries Associating images with conceptual entities is a long-standing goal in Computer Vision (Barnard et al., 2003) and two approaches have built large-scale image libraries based on the WordNet hypernym ontology. The data set of Torralba et al. (2008) contains over 80M images across all 75,062 non-abstract WordNet noun synsets. However, to support the size of the data set, images are down-scaled to 32x32 pixels; furthermore, their image-sense mapping error rates vary between 25-80% with more general concepts having higher error rates. The second significant image library comes from ImageNet (Deng et al., 2009), which contains 3.2M high-resolution images for 5,247 non-abstract WordNet noun synsets. Notably, both libraries focus only on concrete nouns. In contrast, the present work provides a methodology for generating image-sense associations for all parts of speech and for both abstract and concrete concepts. Within NLP resources, BabelNet (Navigli and Ponzetto, 2012) merges Wikipedia and WordNet sense inventories and contains mappings from WordNet senses to the pictures present on the corresponding Wikipedia page. However, since images come from an encyclopedia, the associations are inherently limited to only nouns and, due to inherently partial mapping, only 38.6% of the WordNet senses have images, with an average of 3.01 images for those senses. The present work also varies from previous approaches in that image-sense pairs are rated according to the strength of association between the image and sense, rather than having a binary ungraded association.

Crowdsourced WSD Many NLP areas have applied crowdsourcing (Wang et al., 2013); of these areas, the most related to this work is crowdsourcing word sense annotations. Despite initial success in performing WSD using crowdsourcing (Snow et al., 2008), many approaches noted the difficulty of performing WSD with untrained annotators, especially as the degree of polysemy increases or when word senses are related. Several approaches have attempted to make the task more suitable for untrained annotators by (1) using the crowd itself to define the sense inventory (Biemann and Nygaard, 2010), thereby ensuring the crowd understands the sense distinctions, (2) modifying the questions to explicitly model annotator uncertainty (Hong and Baker, 2011; Jurgens, 2013), or (3) using sophisticated methods to aggregate multiple annotations (Passonneau et al., 2012; Passonneau and Carpenter, 2013). In all cases, annotation was purely text based, in contrast to our work.

3 Game 1: Puzzle Racer

The first game was designed to fill an important need for enabling engaging NLP games: image representations of concepts, specifically WordNet senses. Our goals are two-fold: (1) to overcome the limits of current sense-image libraries, which have focused largely on concrete nouns and (2) to provide a general game platform for annotation tasks that need to associate lexical items with images. Following, we first describe the design, annotation process, and extensibility of the game, and then discuss how its input data is generated. A live demonstration of the game is available online.[1]

3.1 Design and Game Play

Puzzle Racer was designed to be as "video game-like" as possible, with no mentioning of linguistic terminology. Because the game is targeted for the layperson, we view this a fundamental design objective to make the game more engaging and long-lasting. Therefore, Puzzle Racer is modeled after popular games such as Temple Run and Subway Surfers, but with the twist of combining two game genres: racing and puzzle solving. Racing provides the core of game play, while the annotation is embedded as puzzle solving during and after the race.

[1] `http://www.knowledgeforge.org`

Following, we describe the game play and then detail how playing produces annotations.

Racing To race, players navigate a race car along a linear track filled with obstacles and enemy pieces. During play, players collect coins, which can be used to obtain non-annotation achievements and to increase their score. Enemies were added to introduce variety into the game and increase the strategy required to keep playing. Players begin the race with 2–4 health points, depending on the racer chosen, which are decreased when touching enemies. During game play, players may collect power-ups with familiar actions such as restoring lost health, doubling their speed, or acting as a magnet to collect coins. To bring a sense of familiarity, the game was designed using a combination of sprites, sound effects, and music from Super Mario World, Mario Kart 64, and custom assets created by us. Races initially last for 90 seconds, but may last longer if players collect specific power-ups that add time.

Puzzle Solving Prior to racing, players are shown three images, described as "puzzle clues," and instructions asking them to find the common theme in the three pictures (Fig. 1a). Then, during racing, players encounter obstacles, referred to as *puzzle gates*, that show a series of images. To stay alive, players must navigate their racer through the one picture in the series with the same theme as the puzzle clues. Players activate a gate after touching one of its images; a gate may only be activated once and racer movement over other pictures has no effect. Puzzle gates appear at random intervals during game play.

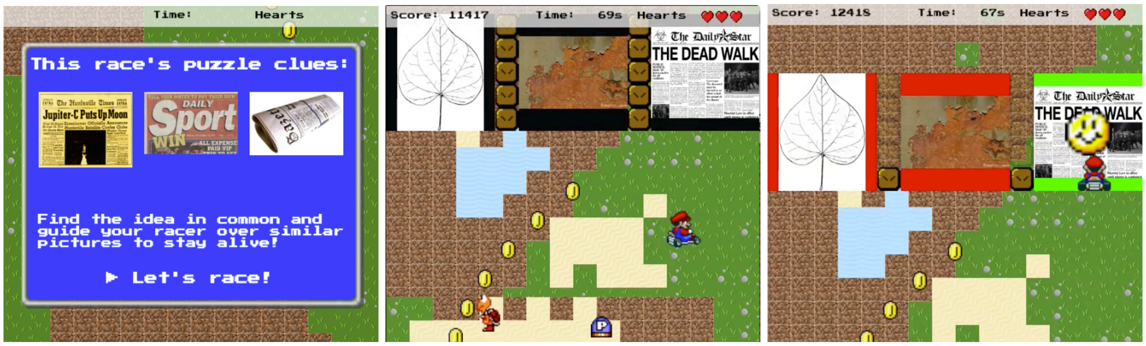
Two types of gate appear. In the first, the gate shows pictures where one picture is known to be related to the puzzle clues. We refer to these as golden gates. Racing over an unrelated image in a golden gate causes the player to lose one health point, which causes the race to end if their health reaches zero. The second type of gate, referred to as a mystery gate, shows three images that are *potentially* related to the clue. Moving over an image in a mystery gate has no effect on health. Prior to activating a gate, there is no visual difference between the two gates.

Figure 1b shows a racer approaching a puzzle gate. Upon first moving their racer on one of the gate's images, the player receives visual and auditory feedback based on the type of gate. In the case of a golden gate, the borders around all pictures change colors showing which picture should have been selected, a feedback icon appears on the chosen picture (shown in Figure 1c), and an appropriate sound effect plays. For mystery gates, borders become blue, indicating the gate has no effect.

Finally, when the race ends, players are asked to solve the race's puzzle by entering a single word that describes the race's puzzle theme. For example, in the race shown in Figure 1, an answer of "paper" would solve the puzzle. Correctly answering the puzzle doubles the points accumulated by the player during the race. The initial question motivates players to pay attention to picture content shown during the race; the longer the player stays alive, the more clues they can observe to help solve the puzzle.

Annotation Image-sense annotation takes place by means of the puzzle gates. Each race's puzzle theme is based on a specific WordNet sense. Initially, each sense is associated with a small set of gold standard images, G, and a much larger set of potentially-associated images, U, whose quality is unknown. At the start of a race, three gold standard images are randomly selected from G to serve as puzzle clues. The details of gold standard image selection are described later in Sec. 3.2. We note that not all gold standard images are shown initially as puzzle clues, helping mask potential differences between golden and mystery gates.

Mystery gates annotate the images in U. The images in a mystery gate are chosen by selecting the least-rated image in U and then pairing it with n-1 random images from U, where n is the number of pictures shown per gate. By always including the least-rated image, we guarantee that, given sufficient plays, all images for a sense will eventually be rated. When a player chooses an image in the mystery gate, that image receives n-1 positive votes for it being a good depiction of the sense; the remaining unselected images receive one negative vote. Thus, an image's *rating* is the cumulative sum of the positive and negative votes it receives. This rating scheme is zero-sum so image ratings cannot become inflated such that all images have a positive rating. However, we do note that if U includes many related images, due to the voting, some good images may have



This race's puzzle clues:

Find the idea in common and guide your racer over similar pictures to stay alive!

▶ Let's race!

(a) Puzzle clues (b) A puzzle gate prior to activation (c) An activated puzzle gate

Figure 1: Screenshots of the key elements of the Puzzle Racer game

negative ratings if even-better images become higher ranked.

Golden gates are used to measure how well a player understands the race's puzzle concept (i.e., the sense being annotated). The first three puzzle gates in a race are always golden gates. We denote the percentage of golden gates correctly answered thus far as α. After the three initial golden gates are shown, the type of new puzzle gates is metered by α: golden gates are generated with probability $0.3 + 0.7(1 - \alpha)$ and mystery gates are generated for the remainder. In essence, accurate players with high α are more likely to be shown mystery gates that annotate pictures from U, whereas completely inaccurate players are prevented from adding new annotations. This mechanism adjusts the number of new annotations a player can produce in real-time based on their current accuracy at recognizing the target concept, which is not currently possible in common crowdsourcing platforms. Last, we note that puzzle answering also provides labels for the race's images, data that might prove valuable for tasks such as image labeling (Mensink et al., 2013) and image caption generation (Feng and Lapata, 2013; Kulkarni et al., 2013).

Additional Game Elements Puzzle Racer incorporates a number of standard Game with a Purpose design elements (von Ahn and Dabbish, 2008), with two notable features: unlockable achievements and a leaderboard. Players initially start out with a single racer and power-up available. Players can then unlock new racers and power-ups through various game play actions of varying difficulty, e.g., correctly answering three puzzle questions in a row. This feature proved highly popular and provided an extrinsic motivation for continued playing. Second, players were ranked according to level, which was determined by the number of correct puzzle answers, correct golden gates, and their total score. The top-ranking players were shown at the end of every round and via a special screen in-game. A full, live-updated leaderboard was added halfway through the study and proved an important feature for new players to use in competing for the top ranks.

Extensibility At its core, Puzzle Racer provides three central annotation-related mechanics: (1) an initial set of instructions on how players are to interact with images, (2) multiple series of images shown during game play, and (3) an open-ended question at the end of the game. These mechanics can be easily extended to other types of annotation where players must choose between several concepts shown as options in the puzzle gates. For example, the instructions could show players a phrase such as "a bowl of *" and ask players to race over images of things that might fit the "*" argument in order to obtain selectional preference annotations of the phrase (à la Flati and Navigli (2013)); the lemmas or senses associated with the selected images can be aggregated to identify the types of arguments preferred by players for the game's provided phrase. Similarly, the instructions could be changed to provide a set of keywords or phrases (instead of images associated with a sense) and ask players to navigate over images of the words in order to perform image labeling.

Nouns	argument, arm, atmosphere, bank, difficulty,
	disc, interest, paper, party, shelter
Verbs	activate, add, climb, eat, encounter
	expect, rule, smell, suspend, win
Adjectives	different, important, simple

Table 1: Lemmas for Puzzle Racer and Ka-boom!

3.2 Image Data

Puzzle Racer requires sets of images G and U as input. Both were constructed using image queries via the Yahoo! Boss API as follows. Three annotators were asked to each produce three queries for each sense of a target word and, for each query, to select three images as gold standard images. Queries were checked for uniqueness and to ensure that at least a few of its image results were related to the sense. Each query was used to retrieve one result set of 35 images. Two additional annotators then validated the queries and gold standard images produced by the first three annotators. Validation ensured that each sense had at least three queries and $|G| \geq 6$ for all senses. After validation, the gold standard images were added to G and all non-gold images in the result set were added to U, discarding duplicates.

During game play, puzzle clues are sampled across queries, rather than sampling directly from G. Query-based sampling ensures that players are not biased towards images for a single visual representation of the sense.

While the construction of G and U is manual, we note that alternate methods of constructing the sets could be considered, including automatic approaches such as those used by ImageNet (Deng et al., 2009). However, as the focus of this game is on ranking images, a manual process was used to ensure high-quality images in G.

Importantly, we stress that the images in G alone are often insufficient due to two reasons. First, most senses – especially those denoting abstract concepts – can be depicted in many ways. Relying on a small set of images for a sense can omit common visualizations, which may limit downstream applications that require general representations. Second, many games rely on a sense of novelty, i.e., not seeing the same pictures repeatedly; however, limiting the images to those in G can create issues where too few images exist to keep player interest high. While additional manual annotation could be used to select more gold standard images, such a process is time-intensive; hence, one of our game's purposes is to eventually move high-quality images from U to G.

disc^4_n: a flat circular plate		
circular plate	dish plate	plate
paper^3_n: a daily or weekly publication on folded sheets		
news paper	daily newspaper	newspaper headline
simple^2_a: elementary, simple, uncomplicated		
simple problem	1+1=2	elementary equation
win^1_v: be the winner in a contest or competition		
olympic winner	lottery winner	world cup victory

Table 2: Examples of queries used to gather images

4 Puzzle Racer Annotation Analysis

Puzzle Racer is intended to produce a sense-image mapping comparable to what would be produced by crowdsourcing. Therefore, we performed a large-scale study involving over 100 players and 16,000 images. Two experiments were performed. First, we directly compared the quality of the game-based annotations with those of crowdsourcing. Second, we compared the difference in quality between expert-based gold standard images and the highest-ranked images rated by players.

4.1 Experimental Setup

To test the potential of our approach, we selected a range of 23 polysemous noun, verb, and adjective lemmas, shown in Table 1. Lemmas had 4-10 senses each, for a total of 132 senses. Many lemmas have both abstract and concrete senses and some are known to have highly-related senses (Erk and McCarthy, 2009). Hence, given their potential annotation difficulty, we view performance on these lemmas as a *lower* bound.

For all lemmas, during the image generation process (Sec. 3.2) annotators were able to produce queries for all but one sense, *expect*^2_v;[2] this produced 1356 gold images in G and 16,656 unrated images

[2] The sense *expect*^2_v has the definition, "consider obligatory; request and expect." Annotators were able to formulate many queries that could have potentially shown images of this definition, but the images results of such queries were consistently unrelated to the meaning.

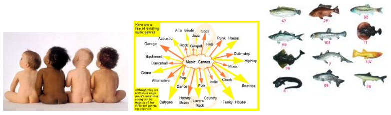
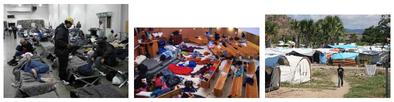
interest$_n^1$: a sense of concern with and curiosity about someone or something 	eat$_v^1$: take in solid food 	different$_a^1$: unlike in nature or quality or form or degree 
party$_n^2$: a group of people gathered together for pleasure 	expect$_v^6$: be pregnant with 	shelter$_n^5$: temporary housing for homeless or displaced persons 

Table 3: Examples of gold standard images

in U. Tables 2 and 3 show examples of the queries and gold standard images, respectively.

The game play study was conducted over two weeks using a pool of undergraduate students, who were allowed to recruit other students. After an email announcement, 126 players participated. Players were ranked according to their character's level and provided with an incentive that the four top-ranking players at the end of the study would be provided with gift cards ranging from $15-25USD, with a total compensation of $70USD.

4.2 Experiment 1: Crowdsourcing Comparison

The first experiment directly compares the image rankings produced by the game with those from an analogous crowdsourcing task. Tasks were created on the CrowdFlower platform using the identical set of examples and annotation questions encountered by players. In each task, workers were shown three example gold standard images (sampled from those configurations seen by players) and asked to identify the common theme among the three examples. Then, five annotation questions were shown in which workers were asked to choose which of three images was most related to the theme. Questions were created after the Puzzle Racer study finished in order to use the identical set of questions seen by players as mystery gates. Workers were paid $0.03USD per task.

To compare the quality of the Puzzle Racer image rankings with those from CrowdFlower, the three highest-rated images of each sense from both rankings were compared. Two annotators were shown a sense's definition and example uses, and then asked to compare the quality of three image pairs, selecting whether (a) the left image was a better depiction of the sense, (b) the right image was better, or (c) the images were approximately equal in quality. In the case of disagreements, a third annotator was asked to compare the images; the majority answer was used when present or, in the case of all three ratings, images were treated as equal, the latter of which occurred for only 17% of the questions. For all 396 questions, the method used to rank the image was hidden and the order in which images appeared was randomized.

Results During the study period, players completed 7199 races, generating 20,253 ratings across 16,479 images. Ratings were balanced across senses, with a minimum and maximum of 231 and 329 ratings per sense. Players accurately identified each race's theme, selecting the correct image in 83% of all golden puzzle gates shown. Table 4 shows example top-rated images from Puzzle Racer.

Experiment 1 measures differences in the quality of the three top-ranked images produced by Puzzle Racer and CrowdFlower for each sense. Puzzle Racer and CrowdFlower produced similar ratings, with at least one image appearing in the top-three positions of both ranks for 55% of the senses.

Both annotators agreed in 72% of cases in selecting the best sense depiction, finding that in 88% of the agreed cases both images were approximately equal representations of the sense. In the remaining, the Puzzle Racer image was better in 4% and

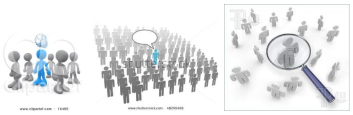
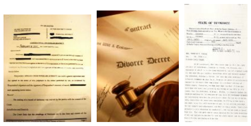
activate^4_v: aerate (sewage) so as to favor the growth of organisms that decompose organic matter 	argument^2_n: a contentious speech act; a dispute where there is strong disagreement 	atmosphere^4_n: the weather or climate at some place
climb^1_v: go upward with gradual or continuous progress 	important^1_a: of great significance or value 	rule^2_v: decide with authority 

Table 4: Examples of the three highest-rated images for six senses

CrowdFlower image better in 8%. When resolutions from a third annotator were included, a similar trend emerges: both images were equivalent in 79% of all cases, Puzzle Racer images were preferred in 7% and Crowflower images in 14%. These results show that, as a video game, Puzzle Racer produces very similar results to what would be expected under equivalent conditions with crowdsourcing.

4.3 Experiment 2: Image Quality

The second experiment evaluates the ability of the games to produce high-quality images by measuring the difference in quality between gold standard images and top-rated images in the game. CrowdFlower workers were shown a question with a sense's definition and example uses and then asked to choose which of two images was a better visual representation of the gloss. Questions were created for each of the three highest-rated images for each sense, pairing each with a randomly-selected gold standard image for that sense. Image order was randomized between questions. Five questions were shown per task and workers were paid $0.05USD per task.[3] The 2670 worker responses were aggregated by selecting each question's most frequent answer.

Results For senses within each part of speech, workers preferred the gold standard image to the top-rated image for nouns, verbs, and adjectives 57.4%, 53.1%, and 56.2% of the time, respectively. This preference is not significant at $p < 0.05$, indicating that the top-ranked images produced through Puzzle Racer game play are approximately equivalent in quality to images manually chosen by experts with full knowledge of the sense inventory.

4.4 Cost Comparison

Puzzle Racer annotations cost $70, or $0.0034USD per rating. In comparison, the analogous CrowdFlower annotations cost $256.60, or $0.0126USD per annotation. Because the game's costs are fixed, the cost per annotation is driven down as players compete. As a result, Puzzle Racer reduces the annotation cost to $\leq$27% of that required by crowdsourcing. We note that other factors could have contributed to the cost reduction over crowdsourcing beyond the video game itself. However, as we demonstrate in Vannella et al. (2014), players will play a video game with a purpose without compensation just as much as they do when compensated using a similar setup as was performed in this experiment. Hence, the video game itself is likely the largest motivating factor for the cost reduction.

Video game-based annotation does come with indirect costs due to game development. For example, Poesio et al. (2013) report spending £60,000 over a two-year period to develop their linguistic game with a purpose. In contrast, Puzzle Racer was created using open source software in just over a month and developed in the context of a Java programming class, removing any professional development costs.

[3] Workers were paid more for the second task to adjust for the time required to read each question's sense definition and example uses; thus, hourly compensation rates in the two experiments were approximately equivalent.

Furthermore, Puzzle Racer is easily extensible for other text-image annotation tasks, enabling the platform to be re-used with minimal effort.

The decreased cost does come with an increase in the time per annotation. All tasks on the CrowdFlower platform required only a few hours to complete, whereas the Puzzle Racer data was gathered over the two-week contest period. The difference in collection time reflects an important difference in the current resources: while crowdsourcing has established platforms with on-demand workers, no central platforms exist for games with a purpose with an analogous pool of game players. However, although the current games were released in a limited fashion, later game releases to larger venues such as Facebook may attract more players and significantly decrease both collection times and overall annotation cost.

5 Game 2: Ka-boom!

Building large-scale sense-annotated corpora is a long-standing objective (see (Pilehvar and Navigli, 2014)) and has sparked significant interest in developing effective crowdsourcing annotation and GWAP strategies (cf. Sec. 2). Therefore, we propose a second video game, *Ka-boom!*, that produces sense annotations from game play. A live demonstration of the game is available online.[4]

Design and Game Play Ka-boom! is an action game in the style of the popular Fruit Ninja game: pictures are tossed on screen from the boundaries of the screen, which the player must then selectively destroy in order to score points. The game's challenge stems from rapidly identifying which pictures should be destroyed or not destroyed as they appear.

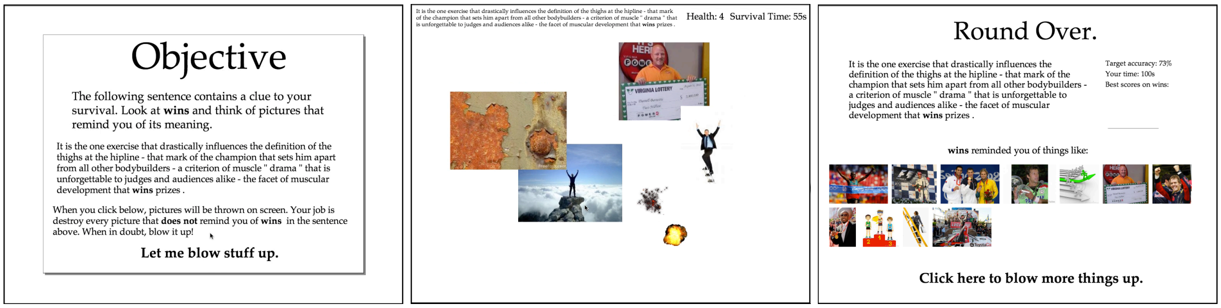
Prior to the start of a round, players are shown a sentence with a word in bold (Fig. 2a) and asked to envision pictures related to that word's meaning in the context. Players are then instructed to destroy pictures that do *not* remind them of the bolded word's meaning and let live pictures showing something reminiscent. Once finished reading the instructions, players begin a round of game play that shows (1) images for each sense of the word and (2) images for unrelated lemmas, referred to as distractor images.

Players destroy pictures by clicking or touching them, depending on their device's input (Fig. 2b). Players are penalized for failing to destroy the distractor images. Rounds begin with a limit of at most three pictures on screen at once, which increases as the round progresses. The additional images provide two important benefits: (1) an increasing degree of challenge to keep the player's interest, (2) more image interactions to use in producing the annotation. Additionally, the increasing picture rate enables us to measure the interaction between game play speed and annotation quality in order to help tune the speeds of future games. The round ends when players fail to destroy five or more distractor images or 60 seconds elapses. Ending the game early after players fail to destroy distractor images provides Ka-boom! a mechanism for limiting the impact of inaccurate or adversarial players on annotation quality. After game play finishes, players are shown their score and all the lemma-related pictures they spared (Fig. 2c), proving a positive feedback loop where players can evaluate their choices.

Annotation Traditionally, sense annotation is performed by having an annotator examine a word in context and then chose the word's sense from a list of definitions. Ka-boom! replaces the sense definitions with image examples of that sense. A sense annotation is built from the senses associated with the images that the player spared. Images are presented to players based on a sequence of flights. Each flight contains one randomly-selected picture for each of a word's n senses and n distractor images. Images within a flight are randomly ordered. The structure of a flight's images ensures that, as the game progresses, players see the same number of images for each sense; otherwise, the player's annotation may become biased simply due to one sense's images appearing more often.

Once the game ends, the senses associated with the spared images are aggregated to produce a sense distribution. For simplicity, the sense with the highest probability is selected as the player's answer; in the case of ties, multiple senses are reported, though, we note that the game's annotation method could also produce a weighted distribution over senses (Erk et al., 2012), revealing different meanings that a player considered valid in the context.

[4] http://www.knowledgeforge.org/



Objective

The following sentence contains a clue to your survival. Look at **wins** and think of pictures that remind you of its meaning.

It is the one exercise that drastically influences the definition of the thighs at the hipline - that mark of the champion that sets him apart from all other bodybuilders - a criterion of muscle " drama " that is unforgettable to judges and audiences alike - the facet of muscular development that **wins** prizes .

When you click below, pictures will be thrown on screen. Your job is destroy every picture that **does not** remind you of **wins** in the sentence above. When in doubt, blow it up!

Let me blow stuff up.

It is the one exercise that drastically influences the definition of the thighs at the hipline - that mark of the champion that sets him apart from all other bodybuilders - a criterion of muscle " drama " that is unforgettable to judges and audiences alike - the facet of muscular development that **wins** prizes .

Health: 4 Survival Time: 55s

Round Over.

It is the one exercise that drastically influences the definition of the thighs at the hipline - that mark of the champion that sets him apart from all other bodybuilders - a criterion of muscle " drama " that is unforgettable to judges and audiences alike - the facet of muscular development that **wins** prizes .

Target accuracy: 73%
Your time: 100s
Best scores on wins:

wins reminded you of things like:

Click here to blow more things up.

(a) The context and target word

(b) Players destroying images

(c) The round-over summary

Figure 2: Screenshots of the three key elements of the Ka-boom! game

The highest probability of a sense from this distribution is then multiplied by the duration of the game to produce the player's score for the round. Players maximize their score when they consistently choose images associated with a single sense, which encourages precise game play.

The annotation design of having players destroy unrelated images was motivated by two factors. First, the mechanism of destroying unrelated images does not introduce noise into the annotation when a player mistakenly destroys an image; because only retained images count towards the sense annotation, players may be highly selective in which images they retain – even destroying some images that are associated with the correct sense – while still producing a correct annotation. Second, our internal testing showed the objective of destroying unrelated pictures keeps players more actively engaged. In the inverse type of play where players destroy only related pictures, players often had to wait for a single picture to destroy, causing them to lose interest.

Extensibility Ka-boom! contains two core mechanics: (1) instructions on which pictures should be destroyed and which should be spared, and (2) series of images shown to the player during game play. As with Puzzle Racer, the Ka-boom! mechanics can be modified to extend the game to new types of annotation. For example, instructions could display picture examples and ask players to destroy either similar or opposite-meaning ideas in order to annotate synonyms or antonyms. In another setting, images can be associated with semantic frames (e.g., from FrameNet (Baker et al., 1998)) and players must spare images showing the frame of the game's sentence in order to provide frame annotations.

6 Ka-boom! Annotation Analysis

Ka-boom! is intended to provide a complementary and more-enjoyable method for sense annotation using only pictures. To test its effectiveness, we perform a direct comparison with the state-of-the-art GWAP for sense annotation, Wordrobe (Venhuizen et al., 2013), which is not a video game.

6.1 Experimental Setup

Organizers of the Wordrobe project (Venhuizen et al., 2013) provided a data set of 111 recently-annotated contexts having between one and nine games played for each (mean 3.2 games). This data was distinct from the contexts used to evaluate Wordrobe in Venhuizen et al. (2013) in which case all contexts had six games played each. Contexts were for 74 noun and 16 verb lemmas with a total of 310 senses (mean 3.4 senses per word). Contexts were assigned the most-selected sense label from the Wordrobe games.

To gather the images for each lemma used with Ka-boom!, we repeated a similar image-gathering process as done for the gold standard images in Puzzle Racer. Annotators generated at least three queries for each sense, selecting three images for each query as gold standard examples of the sense. During annotation, four senses could not be associated with any queries that produced high-quality images. In total, 2594 images were gathered, with an average of 8.36 images per sense. The query data and unrated images are included in the data set, but were not used further in Ka-boom! experiments.

Game players were drawn from a small group of

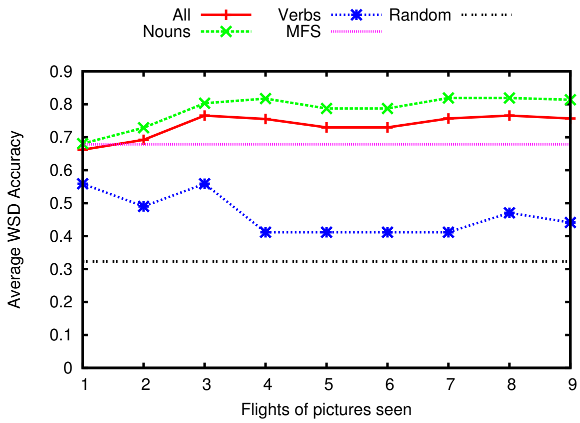


Figure 3: Players' average WSD accuracy within a single game relative to the number of flights seen

	Accuracy		
Method	All	Noun	Verb
Ka-boom!	**0.766**	**0.803**	0.559
Wordrobe	0.603	0.659	0.313
MFS	0.678	0.702	**0.588**
Random	0.322	0.325	0.312

Table 5: Sense disambiguation accuracies

fluent English speakers and were free to recruit other players. A total of 19 players participated. Unlike Puzzle Racer, players were not compensated. Each context was seen in at least six games.

WSD performance is measured using the traditional precision and recall definitions and the F1 measure of the two (Navigli, 2009); because all items are annotated, precision and recall are equivalent and we report performance as accuracy. Performance is measured relative to two baselines: (1) a baseline that picks the sense of the lemma that is most frequent in SemCor (Miller et al., 1993), denoted as MFS, and (2) a baseline equivalent to performance if players had randomly clicked on images, denoted as Random.[5]

6.2 Results

Two analyses were performed. Because Ka-boom! continuously revises the annotation during gameplay based on which pictures players spare, the first analysis assesses how the accuracy changes with respect to the length of one Ka-boom! game. The second analysis measures the accuracy with respect to the number of games played per context.

In the first analysis, each context's annotation was evaluated using the most-probable sense after each flight of gameplay. Figure 3 shows results after six games were played. Players were highly accurate at playing, surpassing the MFS baseline after seeing two flights of pictures (i.e., two pictures for each sense). Accuracy remained approximately equivalent after three rounds for noun lemmas, while verb lemmas showed a small drop-off in performance. We believe that the increased rate at which images occur on screen likely caused lower performance, where players were unable to react quickly enough. Many noun lemmas had easily-recognizable associated images, so higher-speed game play may still be accurate. In contrast, verbs were more general (e.g., "decide," "concern," and "include"), which required more abstract thinking in order to recognize an associated picture; as the game speed increased, players were not able to identify these associated pictures as easily, causing slightly decreased performance.

Table 5 shows the players' disambiguation accuracy after three flights in comparison to the players' accuracy with Wordrobe and the two baselines. Ka-boom! provides an increased performance over Wordrobe that is statistically significant at $p < 0.01$.[6] Ka-boom! also provides a performance increase over the MFS baseline, though it is statistically significant only at $p = 0.14$. The time required to gather annotations after three flights varied based on the number of senses, but was under a minute in all cases, which puts the rate of annotation on par with that of expert-based annotation (Krishnamurthy and Nicholls, 2000).

In the second analysis, disambiguation accuracy was measured based on the number of games played for a context.[7] Because the provided Wordrobe data set has 3.2 games played per context on average, results are reported only for the subset of contexts played in at least four Wordrobe games in order to obtain consistent performance estimates. Ka-

[5] This baseline is similar to random sense selection but takes into account differences in the number of pictures per sense.

[6] We note that, although Venhuizen et al. (2013) report a higher accuracy for Wordrobe in their original experiments (85.7 F1), that performance was measured on a different data set and used six games per context.

[7] In all cases, players played at most one game per context.

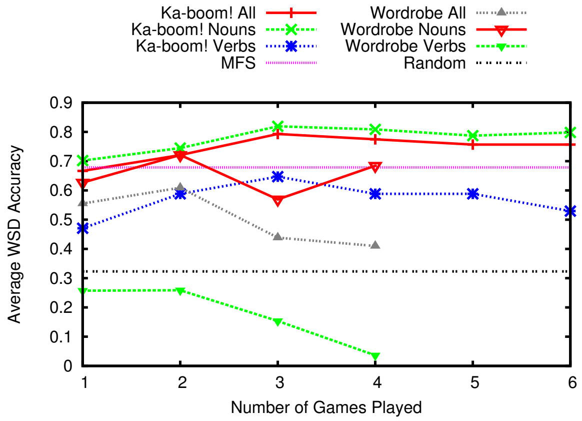


Figure 4: Average WSD accuracy as a function of the number of games played for a context

boom! annotations are recorded after three flights were seen in each game. Figure 4 shows the performance relative to the number of annotators for both Ka-boom! and Wordrobe, i.e., the number of games played for that context by different players.

For nouns, Ka-boom! is able to exceed the MFS baseline after only two games are played. For both nouns and verbs, multiple rounds of Ka-boom! game play improve performance. In contrast to Ka-Boom!, Wordrobe accuracy declines as the number of players increases; when multiple players disagree on the sense, no clear majority emerges in Wordrobe, lowering the accuracy of the resulting annotation. In contrast, a single player in Ka-boom! produces multiple sense judgments for a context in a single game via interacting with each flight of images. These interactions provide a robust distributional annotation over senses that can be easily aggregated with other players' judgments to produce a higher-quality single sense annotation. This analysis suggests that Ka-boom! can produce accurate annotations with just a few games per context, removing the need for many redundant annotations and improving the overall annotation throughput.

7 Conclusion and Future Work

In this work we have presented a new model of linguistic Games with a Purpose focused on annotation using *video games*. Our contributions show that designing linguistic annotation tasks as video games can produce high-quality annotations. In the first game, Puzzle Racer, we demonstrated that game play can produce a high-quality library of images associated with WordNet senses, equivalent to those produced by expert annotators. Moreover, Puzzle Racer reduces the cost of producing an equivalent resource via crowdsourcing by at least 73% while providing similar-quality image ratings. In the second game, Ka-boom!, we demonstrated that a video game could be used to perform accurate word sense annotation with a large improvement over the MFS baseline and a statistically significant improvement over current game-based WSD.

While not all linguistic annotations tasks are easily representable as video games, our two games provide an important starting point for building new types of NLP games with a purpose based on video games mechanics. Software for both games will be open-sourced, providing a new resource for future game development and extensions of our work. Furthermore, the multiple data sets produced by this work are available at `http://lcl.uniroma1.it/videogames`, providing (1) a sense-image mapping from hundreds of senses to tens of thousands of images, (2) word labels for most images in our dataset, (3) Web queries associated with all senses, and (4) image-based word sense annotations.

Based on our results, three directions for future work are planned. The two games presented here focus on concepts that can be represented visually and thus lend themselves to annotations for lexical semantics. However, the fact that the games are graphical does not prevent them from showing textual items (see Vannella et al. (2014)) and more apt video games could be developed for text-based annotations such as PP-attachment or pronoun resolution. Therefore, in our first future work, we plan to develop new types of video games for textual items as well as extend the current games for new semantic tasks such as selectional preferences and frame annotation. Second, we plan to scale up both games to a broader audience such as Facebook, creating a larger sense-image library and a standard platform for releasing video games with a purpose. Third, we plan to build multilingual games using the images from Puzzle Racer, which provide a language-independent concept representation, and could therefore be used to enable the annotation and validation of automatically-created knowledge resources (Hovy et al., 2013).

Acknowledgments

The authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234.

We thank the many game players whose collective passion for video games made this work possible.

References

Guillaume Artignan, Mountaz Hascoët, and Mathieu Lafourcade. 2009. Multiscale visual analysis of lexical networks. In *Proceedings of the International Conference on Information Visualisation*, pages 685–690.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics,* Montréal, Québec, Canada, 10–14 August 1998, Montreal, Canada.

Kobus Barnard, Pinar Duygulu, David Forsyth, Nando De Freitas, David M. Blei, and Michael I. Jordan. 2003. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135.

Chris Biemann and Valerie Nygaard. 2010. Crowdsourcing WordNet. In *The 5th International Conference of the Global WordNet Association (GWC-2010)*.

Jon Chamberlain, Karën Fort, Udo Kruschwitz, Mathieu Lafourcade, and Massimo Poesio. 2013. Using games to create language resources: Successes and limitations of the approach. In Iryna Gurevych and Jungi Kim, editors, *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, pages 3–44. Springer.

Tim Chklovski and Yolanda Gil. 2005. Improving the design of intelligent acquisition interfaces for collecting world knowledge from web contributors. In *Proceedings of the International Conference on Knowledge Capture*, pages 35–42. ACM.

Tim Chklovski and Rada Mihalcea. 2002. Building a Sense Tagged Corpus with Open Mind Word Expert. In *Proceedings of ACL 2002 Workshop on WSD: Recent Successes and Future Directions*, Philadelphia, PA, USA.

Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit players. 2010. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255.

Philip Edmonds and Adam Kilgarriff. 2002. Introduction to the special issue on evaluating word sense disambiguation systems. *Natural Language Engineering*, 8(4):279–291.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 440–449, Singapore.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2012. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

Christiane Fellbaum, Joachim Grabowski, and Shari Landes. 1998. Performance and confidence in a semantic annotation task. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database*, pages 217–237. MIT Press.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Yansong Feng and Mirella Lapata. 2013. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):797–812.

Tiziano Flati and Roberto Navigli. 2013. SPred: Large-scale Harvesting of Semantic Predicates. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1222–1232, Sofia, Bulgaria.

Amaç Herdağdelen and Marco Baroni. 2012. Bootstrapping a game with a purpose for common sense collection. *ACM Transactions on Intelligent Systems and Technology*, 3(4):1–24.

Barbora Hladká, Jiří Mírovskỳ, and Pavel Schlesinger. 2009. Play the language: Play coreference. In *Proceedings of the Joint Conference of the Association for Computational Linguistics and International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, pages 209–212. Association for Computational Linguistics.

Jisup Hong and Collin F. Baker. 2011. How Good is the Crowd at "real" WSD? In *Proceedings of the Fifth Linguistic Annotation Workshop (LAW V)*, pages 30–37. ACL.

Eduard H. Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)*, pages 556–562.

Ramesh Krishnamurthy and Diane Nicholls. 2000. Peeling an onion: The lexicographer's experience of manual sense-tagging. *Computers and the Humanities*, 34(1-2):85–97.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903.

Yen-ling Kuo, Jong-Chuan Lee, Kai-yang Chiang, Rex Wang, Edward Shen, Cheng-wei Chan, and Jane Yung-jen Hsu. 2009. Community-based game design: experiments on social games for commonsense data collection. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 15–22.

Mathieu Lafourcade and Alain Joubert. 2010. Computing trees of named word usages from a crowdsourced lexical network. In *Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT)*, pages 439–446, Wisla, Poland.

Tak Yeon Lee, Casey Dugan, Werner Geyer, Tristan Ratchford, Jamie Rasmussen, N. Sadat Shami, and Stela Lupushor. 2013. Experiments on motivational feedback for crowdsourced workers. In *Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 341–350.

Thomas Mensink, Jakob J. Verbeek, and Gabriela Csurka. 2013. Tree-structured crf models for interactive image labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):476–489.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. 1993. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pages 303–308, Plainsboro, N.J.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli. 2009. Word Sense Disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):1–69.

Martha Palmer, Hoa Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(2):137–163.

Rebecca J. Passonneau and Bob Carpenter. 2013. The benefits of a model of annotation. In *7th Linguistic Annotation Workshop and Interoperability with Discourse*, August 8–9.

Rebecca J. Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):209–252.

Ei Pa Pa Pe-Than, DH-L Goh, and Chei Sian Lee. 2012. A survey and typology of human computation games. In *Information Technology: New Generations (ITNG), 2012 Ninth International Conference on*, pages 720–725. IEEE.

Mohammad Taher Pilehvar and Roberto Navigli. 2014. A Large-scale Pseudoword-based Evaluation Framework for State-of-the-Art Word Sense Disambiguation. *Computational Linguistics*, 40(4).

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*, 3(1):3:1–3:44, April.

Nitin Seemakurty, Jonathan Chu, Luis Von Ahn, and Anthony Tomasic. 2010. Word sense disambiguation via human computation. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 60–63. ACM.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 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,* Waikiki, Honolulu, Hawaii, 25-27 October, pages 254–263.

Antonio Torralba, Robert Fergus, and William T Freeman. 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970.

Daniele Vannella, David Jurgens, Daniele Scarfini, Domenico Toscani, and Roberto Navigli. 2014. Validating and extending semantic knowledge bases using video games with a purpose. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1294–1304.

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

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 319–326.

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

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, 47(1):9–31.