Gamifying Language Resource Acquisition

By

CHRISTOPHER JAMES MADGE

School of Electronic Engineering And Computer Science
QUEEN MARY UNIVERSITY LONDON

Submitted in partial fulfillment of the requirements of the
Degree of Doctor of Philosophy

SEPTEMBER 2019

Abstract

Natural Language Processing, is an important collection of methods for processing the vast amounts of available natural language text we continually produce. These methods make use of supervised learning, an approach that learns from large amounts of annotated data. As humans, we're able to provide information about text that such systems can learn from. Historically, this was carried out by small groups of experts. However, this did not scale. This led to various crowdsourcing approaches being taken that used large pools of non-experts.

The traditional form of crowdsourcing was to pay users small amounts of money to complete tasks. As time progressed, gamification approaches such as GWAPs, showed various benefits over the micro-payment methods used before. These included a cost saving, worker training opportunities, increased worker engagement and potential to far exceed the scale of crowdsourcing. While these were successful in domains such as image labelling, they struggled in the domain of text annotation, which wasn't such a natural fit. Despite many challenges, there were also clearly many opportunities and benefits to applying this approach to text annotation. Many of these are demonstrated by Phrase Detectives. Based on lessons learned from Phrase Detectives and investigations into other GWAPs, in this work, we attempt to create full GWAPs for NLP, extracting the benefits of the methodology. This includes training, high quality output from non-experts and a truly game-like GWAP design that players are happy to play voluntarily.

ACKNOWLEDGEMENTS

I would first and foremost like to express my sincere thanks to my supervisors Massimo Poesio and Udo Kruschwitz, for their guidance, kind patience and always being there when I needed them.

Also, I would like to express my gratitude to my friends and colleagues, the DALI project team (Disagreements and Language Interpretation), which I consider myself very fortunate to be part of - with special thanks to:

- Jon Chamberlain - for his inspirational work on Phrase Detectives, contributions and advice
- Silviu Paun - contributions relating to probabilistic aggregation
- Juntao Yu - contributions relating to mention detection

A further thank you to IGGI for giving me the opportunity to pursue this Ph.D. and providing excellent Doctoral Training.

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

AUTHOR'S DECLARATION

I, *Christopher James Madge*, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

SIGNATURE: C. Madge
DATE: 2nd September 2019

Details of collaboration:

- The proposed metrics in Chapter 5 was a collaboration with *Jon Chamberlain*
- The two automated mention detectors used by TileAttack (Section 8.1) were developed by *Juntao Yu*
- The probabilistic aggregation used in Section 7.6.3 was developed by *Silviu Paun*

Publications:

- C. Madge, J. Chamberlain, U. Kruschwitz, and M. Poesio, "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*, (Amsterdam, The Netherlands), pp. 397–404, ACM Press, 2017
- C. Madge, U. Kruschwitz, J. Chamberlain, R. Bartle, and M. Poesio, "Testing game mechanics in games with a purpose for NLP applications," in *In Proceedings of Games4NLP: Using Games and Gamification for Natural Language Processing.*, (Valencia, Spain), p. 2, 2017

- C. Madge, J. Yu, J. Chamberlain, U. Kruschwitz, S. Paun, and M. Poesio, "Crowdsourcing and Aggregating Nested Markable Annotations," in *Proceedings of the 57th Conference of the Association for Computational Linguistics*, (Florence, Italy), pp. 797–807, Association for Computational Linguistics, July 2019
- C. Madge, J. Chamberlain, U. Kruschwitz, and M. Poesio, "Progression In A Language Annotation Game With A Purpose," in *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2019, Washington, USA, October 28-30, 2019.*, (Washington, USA), AAAI, 2019
- C. Madge, R. Bartle, J. Chamberlain, U. Kruschwitz, and M. Poesio, "Incremental Game Mechanics Applied to Text Annotation," in *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '19, (Barcelona, Spain), pp. 545–558, ACM, 2019
- C. Madge, R. Bartle, J. Chamberlain, U. Kruschwitz, and M. Poesio, "Making Text Annotation Fun with a Clicker Game," in *Proceedings of the 14th International Conference on the Foundations of Digital Games*, FDG '19, (San Luis Obispo, California), pp. 77:1–77:6, ACM, 2019
- C. Madge, R. Bartle, J. Chamberlain, U. Kruschwitz, and M. Poesio, "The Design Of A Clicker Game for Text Labelling," (London), IEEE, 2019

Table of Contents

List of Tables

List of Figures

Figure **Page**

Part I

Introduction

CHAPTER 1

INTRODUCTION

1.1 Motivation

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. Originally, these "language resources" were typically hand-annotated by experts [39] over long periods of time. This process can be time consuming, expensive and tedious. Consequently, this requirement for annotated data remains an obstacle to progression for some NLP tasks. Willing experts are few and may be difficult to recruit. One proven method of reducing the time to gather annotations is crowdsourcing more judgements from a large pool of non-experts, then extracting wisdom from that crowd [40]. However, this still does not scale very well. An alternative method that has been explored is the application of gamification. Gamification has been described as "the use of game design elements in non-game contexts" [41]. When attempting to build large corpora, this approach can be cheaper [19], and provide better contributor engagement [42].

Going beyond the adoption of selected game elements (e.g. points or badges), Games-with-a-Purpose (GWAPs) attempt to attract large numbers of people by offering a complete game and harnessing human effort as a side effect of play [43]. This is an appealing proposition as games in general have proven to be a great way to attract people. A well designed game can attract people in the millions. Monument Valley for example, had over 26 million players in its second year [44]. GWAPs have been successful in many applications attracting large numbers of users to label datasets and solve real world problems [45]. Examples include *The ESP Game*, in which by playing, players contribute image labels [46], and *FoldIt*, in which players solve protein-structure prediction problems [47].

However, whilst gamification has been very effective in motivating text labelling (e.g. *Phrase Detectives* gathering anaphoric annotations [19]), 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 both greater challenges and greater opportunities [48]. The former typically adds a layer of game-like themes and carefully selected motivational game mechanics, while the later requires mapping the task completely into a game. 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 [27]. 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 [49]. The linguistic complexity of some text annotation tasks may not be immediately obvious or difficult to map into a game domain.

Identifying and overcoming the challenges in applying the GWAP approach to language resourcing may provide multiple benefits, including the much needed inexpensive large scale annotation experienced in other domains.

1.2 Research Questions

RQ1: Is it possible to develop enjoyable games for NLP that produce quality output? This is the core overarching question of this work. Multiple games have been proposed in the past for NLP tasks. However, they are often criticised for being either not very game-like and therefore not true GWAPs, or fail to ever gather useful quality annotations at scale. To answer this question, there are a number of other questions that must be answered.

RQ2: Can we develop truly entertaining games for NLP? Developing a "real game" for NLP is very challenging. Unlike image labelling, text annotation is not a natural fit into games. Before, we can ask if a game is truly entertaining, we first need to ask how we can evaluate its success. To this end, we propose a set of metrics selected and adapted from those used to evaluate Free-to-Play (a revenue model that gives the user access to the product for free but charges later for specific features) games. We also need to identify which challenges exist in this design space, preventing previous GWAPs for NLP from recruiting players comparable to GWAPs in other domains. Having identified these challenges, we identify a suitable matching game design and hypothesise that it is suited to fit this challenging design space. The impact of this design is measured using the aforementioned metrics.

RQ3: Can we improve on the performance of an automated pipeline using a GWAP to correct its output? A common pattern for annotation is to correct the output of an existing pipeline. This serves to drastically reduce the amount of work required and inform non-expert annotation. Automated modules for NLP are often specifically configured to suit their purpose

of supporting judgements downstream. This may for example involve targeting a high recall rather than a high F_1. We also ask which configuration is best suited for correction. In addition, as we can expect annotators will be a mixed ability group of non-experts we will extract wisdom from the group by aggregating multiple judgements to find the most likely correct judgement for each case. We compare multiple aggregation methods to determine which performs best. We hypothesise that it is possible to improve upon a state-of-the-art pipeline by using a GWAP to correct its output.

RQ4: How can we perform annotation with non-expert players? Crowdsourcing was originally targeted tasks such as image labelling that were largely homogeneous in their nature, all of similar difficulty, requiring only common sense knowledge to complete. However, more recently it is being applied to increasingly difficult tasks. This can be a challenge when working with non-expert annotators. It has been said that one of the main benefits of using GWAPs over other crowdsourcing approaches, is the availability of training opportunities in games. We hypothesise that the introduction of game-like progression into a GWAP can improve annotation accuracy. We also test the idea that we can use a separate game to train players, and progress them on to GWAPs that target more complex annotation tasks.

1.3 Contributions

A design for engaging game-like text annotation Many design approaches have been proposed to address the challenge of integrating text annotation into a game. However, none have yet demonstrated the recruitment or retention required for large scale annotation. In Chapter 11, we investigate these challenges through the lens of game design and propose an approach. Through application of the approach we see rewards both in terms of player recruitment, play session length and player learning. This work is published in CHI PLAY '19 [5].

An approach for large scale nested sequence labelling In Chapter 11 we present a method for identifying markables for coreference annotation that combines high-performance automatic markable detectors with checking with a Game-With-A-Purpose (GWAP) and aggregation using a Bayesian annotation model. A key part of this contribution is its applicability to the case in which markables are nested. In evaluation this approach yields a result several percentage points higher than a state-of-the-art automated mention detector. This work is published in ACL '19 [3].

An approach to, and evaluation of, introducing game-like progression in a text-annotation task Within traditional games design, incorporating progressive difficulty is considered a fundamental principle. However, despite the clear benefits, progression is not such a prominent feature of Games-With-A-Purpose (GWAPs), nor one that is commonly evaluated. In Chapter

9 we present an approach to progression in GWAPs that generalizes to different annotation tasks with minimal, if any, dependency on gold annotated data. Using this method we show a statistically significant increase in accuracy over randomly showing items to annotators. This work is published in HCOMP '19 [4].

1.4 Publications

- C. Madge, J. Chamberlain, U. Kruschwitz, and M. Poesio, "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*, (Amsterdam, The Netherlands), pp. 397–404, ACM Press, 2017

- C. Madge, U. Kruschwitz, J. Chamberlain, R. Bartle, and M. Poesio, "Testing game mechanics in games with a purpose for NLP applications," in *In Proceedings of Games4NLP: Using Games and Gamification for Natural Language Processing.*, (Valencia, Spain), p. 2, 2017

- C. Madge, J. Yu, J. Chamberlain, U. Kruschwitz, S. Paun, and M. Poesio, "Crowdsourcing and Aggregating Nested Markable Annotations," in *Proceedings of the 57th Conference of the Association for Computational Linguistics*, (Florence, Italy), pp. 797–807, Association for Computational Linguistics, July 2019

- C. Madge, J. Chamberlain, U. Kruschwitz, and M. Poesio, "Progression In A Language Annotation Game With A Purpose," in *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2019, Washington, USA, October 28-30, 2019.*, (Washington, USA), AAAI, 2019

- C. Madge, R. Bartle, J. Chamberlain, U. Kruschwitz, and M. Poesio, "Incremental Game Mechanics Applied to Text Annotation," in *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '19, (Barcelona, Spain), pp. 545–558, ACM, 2019

- C. Madge, R. Bartle, J. Chamberlain, U. Kruschwitz, and M. Poesio, "Making Text Annotation Fun with a Clicker Game," in *Proceedings of the 14th International Conference on the Foundations of Digital Games*, FDG '19, (San Luis Obispo, California), pp. 77:1–77:6, ACM, 2019

- C. Madge, R. Bartle, J. Chamberlain, U. Kruschwitz, and M. Poesio, "The Design Of A Clicker Game for Text Labelling," (London), IEEE, 2019

Part II

Background

CHAPTER 2

NATURAL LANGUAGE PROCESSING

Natural language interpretation is typically carried out using a series of steps. This has led to the development of NLP pipelines with a collection of modules such as the popular Stanford CoreNLP pipeline [50], a subset of the available steps consist of tokenisation module, sentence splitting module, a part-of-speech tagging module, a named entity recognition module etc. Each of these steps requires information from the previous step.

For example, a pipeline might look like:

Tokenization → Part-of-Speech Tagging → Mention Detection → Anaphora

2.1 Steps of a Pipeline

This section will focus on some of the core text annotation tasks and identification of tasks that would benefit from human annotation. We follow the pipeline going from tokenization, through to labelled anaphora. We discuss how annotations are used from previous tasks to support latter tasks in a pipeline. In addition, we have look at common errors that can occur in each stage and how they can cascade through the pipeline to impact later steps. This is particularly relevant for supervised learning. Of recent, motivated by this problem of cascading errors, some supervised learning systems have started to take an end-to-end approach [51]. However, it is clearly still highly desirable to achieve high quality annotations if possible in this independent pipeline steps.

2.1.1 Tokenization

Tokenization is a type of segmentation task that divides text into sentences and words. English is one of the easier languages for tokenizing sentences and words as they are typically delimited

by full stops and spaces respectively. However, this still remains challenging as those characters have multiple purposes, leading to ambiguity. Consider for example the sentence:

He said, "we're visiting Univ. Queen Mary". The train to London cost £5.40.

The simple rule of spaces and full stops would result in many errors. Examples of issues that arise with full stops include abbreviations (e.g. Univ.), acronyms (e.g. m.p.h.) and numeric expressions (e.g. £5.50). Common challenges that arise from tokenization include how to consider punctuation (e.g. "said," compared "Univ.", the first would be a separate token, the second would likely not) and clitic contraction (e.g. *we're* is effectively two words. There are many more, including character encoding.

Segmentation in other languages that do not make use of spaces to mark word boundaries, such as Chinese, is more challenging [52].

Previous context provides a lot of information to disambiguate characters leading to implementations such as Finite State Transducers [53, 54], that remains a popular method in modern pipelines [50].

2.1.2 Part-of-Speech Tagging

Parts-of-speech are word classes such as noun, verb pronoun preposition, adverb, conjunction, participle. These provide both semantic and syntactic information. For example, semantically we understand nouns to often be people, places or things (although this is not always the case). Syntactically nouns can often be preceded by determiners, e.g. *The cat*. Each of these classes can be open or closed, where closed means there is a small fixed set unlikely to change (e.g. articles: *a*, *an*, *the*), and open means the set is likely to change, (e.g. verbs: to *google*).

Using a traditional approach, tokenization is performed before Part-of-Speech tagging to separate the words and the neighbouring punctuation so classes can be applied.

The challenge with part-of-speech tags, like tokenization, is that each word can have multiple usages and therefore lexical categories. This prevents a basic word to category look up approach. For example:

The pain in Bob's $\textbf{back}_{\text{noun}}$ meant he couldn't $\textbf{back}_{\text{verb}}$ his car out the way, so he came $\textbf{back}_{\text{adverb}}$ into his house through the $\textbf{back}_{\text{adjective}}$ door.

Fortunately, unlike the example above, most words in English are unambiguous. However, unfortunately, the ambiguous words are some of the most common [52]. These challenges remain non-trivial and pipelines still struggle. For example, consider the sentence:

Recycle that can.

We're looking for VB DT NN. Stanford CoreNLP [50] gives VB DT MD. This pipeline incorrectly classifies *can* as a modal verb. NLTK gives NNP WDT MD. A similar error, but also incorrectly marks *Recycle* as a proper noun, rather than a verb.

Furthermore, there are ambiguities that occur at the semantic level. For example:

Bob was **entertaining**.

It is unclear if he is being described as entertaining (adjective), or he was performing the act of entertaining (verb). Resolution, in this case, would require broader context and understanding.

Part-of-speech tagging is used to support various other tasks, including mention detection (described in more detail in Section 2.1.3) and parsing. These errors can easily propagate through to other steps in the pipeline. This has lead to investigation of the optimal part-of-speech taggers for specific tasks, such as parsing [55, 56].

Accurate part-of-speech tagging is vital to most mention detection approaches. Taking into consideration the parts-of-speech found, we know that the following sequences of part-of-speech tags are valid noun-phrase: DT NN or NNP. Our pipelines would be unlikely to identify "that can" as a result of the automated classification above. Furthermore, in the case of NLTK, the false positive "Recycle" would show as a noun-phrase.

Algorithms and their implementations can be categorised into rule-based and probabilistic with minor exceptions that combine both (transformation based learning [57]). The rule based taggers use a series of rules or a grammar [58, 59]. Multiple probabilistic approaches have been used including Conditional Random Fields [60], Maximum Entropy Markov Models [61], HMMs [62] and neural networks (from [63] to [64]).

2.1.3 Mention Detection

Different coreference corpora adopt different definitions of a markable with respect to mention detection. [65, 66]. The definition of (candidate) mention used in this work is broadly speaking that adopted in corpora based on the MATE scheme [67], such as ONTONOTES [68] ARRAU [69] and Phrase Detectives 1.0 [70].

Mention detection is generally recognized as a very important step for overall coreference quality, if not the most important step [66, 71–73], so a number of good quality mention detectors exist, best known of which is the mention detector included in the Stanford CORE pipeline [74], which was used by many of the top-performing systems in the 2012 CONLL Shared Task [68]. In many of the most recent systems mention detection is carried out as a joint inference task with coreference resolution [75]–this is the case of the current top performing system on the CONLL 2012 dataset, [76]. But even such systems require mention-annotated corpora for training and testing of course. But this performance can still be improved.

But even the best automatic mention detectors do not achieve the accuracy required for high-quality corpus annotation, even when run in-domain, as shown by the fact that the difference in performance between running coreference resolvers on gold mentions and running them on system mentions can be of up to 20 percentage points; the results are of course even poorer when

running such systems out-of-domain, for domains like biomedicine [77] or for under-resourced languages [78]. So a manual checking step is still required to obtain high-quality results.

One difference between the mention detectors used for coreference resolvers and those used to preprocess data for coreference annotation is relevant for subsequent discussion. The former usually aim for high recall and compromise on precision, placing more confidence/importance on the coreference resolution step [79] and being satisfied that incorrectly identified mentions will simply remain singletons which can be removed in post processing [80]. The latter tend to go for high F.

Markable checking used to be very same individuals who carry out the coreference annotation. But increasingly, annotation is done using **crowdsourcing** [12, 13, 40], primarily for reasons of cost.

The annotation of mentions for coreference has similarities with the identification of the chunks for named entity resolution (NER), with the key difference that mentions can and often are nested, as in the following example, from the *Phrase Detectives* corpus [70]), where a mention of entity i is nested inside a mention of entity j.

$$[\text{A wolf}]_i \text{ had been gorging on } [\text{an animal } [\text{he}]_i \text{ had killed}]_j$$

2.1.4 Anaphora

We briefly discuss anaphora resolution here as this is the primary motivation for mention identification, and is itself used in many other complete tasks. These include Textual Entailment [81], Summarisation [82, 83], Term Extraction [83], Text Classification [83] and Sentiment Analysis [84].

Paraphrasing Jurafsky and Martin, Reference Resolution is identifying which pronouns refer to their noun phrases [85]. When two references refer to the same entity, it is known as coreference resolution. Information gathered relating to coreference resolution is very useful in information extraction and summarization tasks. In the context of information extraction it is important when identifying entities, to achieve a good coverage, that any pronouns referring to those entities are also recovered.

Coreference resolution is typically described in one of two broad categories.

Anaphora Alice got to work late. She missed the bus.

Cataphora He won the game, Bob had been practising for weeks.

In the context of this work, it is useful to identify what features, or information must be extracted from a document before these relationships can be identified. A typical pipeline consists of some parsing; mention detection ; coreference resolution and some post processing [80].

Features typically include distance between the mentions or number of mentions, various syntactic features (including part of speech tags), semantic features (including named entity

type), character level features (including exact or partial matching) and lexical features (head word of the mention) [86].

The field of co-reference resolution uses a great deal of esoteric context sensitive terminology, and idioms which benefit from definition for the purpose of understanding other materials relating to this topic. The above subsection gives a simple example, but the following subsection will step through some of the more common cases using the correct terminology.

A *discourse* is a structured text of related sentences that are coherent. This is typical of most text. To situate this, there are sub classes of *discourse* such as a monologue (a single speaker), or a dialogue (two speakers conversing).

A phrase that refers to something (e.g. "he", "she", "the author of the book", "the 40 year old", "the CEO of the company"), is known as the *referring phrase*. The thing that is referred to, is known as the *referent*. It may be worth noting that the *referring phrase* is typically a *noun phrase*, but not all *noun phrases* are *referring phrases*. Furthermore, the identification of *noun phrases* should not confuse their *adjectival forms* such as the "the British tourists", in which British is used as an adjective. *Pleonastic* noun phrases should also be ignored. These typically start a sentence. For example, *It* in "*It* is sunny today", or "*It* is possible that". Referring phrases can also be *verbal*, that is to say, they refer to *verb phrases*. For example, "Please don't answer your phone on the plane. Doing so, may disturb the equipment.", in which "*doing so*" refers to "*answer your phone on the plane*".

If there are two or more *referring expressions* that target the same *referent*, those expressions are said to *co-refer*.

Anaphora, is a type of reference relationship, where the *referent* precedes the *referring expression* (example given above). The *referent* may then be known as the *antecedent*. If there are multiple antecedents being referred to by the referring phrase, this *antecedent* may be called a *split antecedent* (e.g. *Alice* and *Bob* were late. *They* missed the bus.). This use in linguistics, is not to be confused with the rhetoric use of the word anaphora, which uses the same word in repetition, at the beginning of successive phrases for effect.

Anaphora can be split into three categories. *Pronominal anaphora* is anaphora where the *referring expression* is a pronoun such as he/she (personal), his/hers (reflexive), him/her (demonstrative), which (relative) and is a subtask of *coreference resolution*. It is important to note that not all pronouns are referring expressions. *Nominal anaphora* is the linguistic concept, is when the *referrent* is *non-pronominal*.

Bridging, is when there is a weaker tie between the *referring expression* and *referent*. For example, perhaps some component of the *referring expression* is mentioned by the *referent*. For example, "I saw a *couple* walking down the street, the *man* was wearing a hat". The man is part of the couple, but isn't a direct reference to the couple as a whole as for example, *they* would be. [87]

2.2 Traditional Annotation

In this section we will look at traditional methods of annotation with a particular focus on nested sequence labelling. This includes the processes and design and interfaces of the tools used (summarised in Table 2.1).

Annotation Tool	**Year of release**	**Platform**
GATE [88]	1996	Desktop (Java)
MMAX [89]	2001	Desktop (Java)
MMAX2 [9]	2006	Desktop (Java)
BRAT [10]	2012	Web-based
WebAnno [11]	2013	Web-based
GATE Teamware [90]	2013	Web-based
GATE Crowdsourcing Plugin [91]	2014	Web-based

Table 2.1: NLP Annotation Tools

GATE is one of the earliest and most complete offerings. As a language development environment with annotation as one of its many features, allows for processes such as supporting annotation with customised automated pipelines to reduce the workload. The second version of GATE was created in Java (originally C++/Tcl) [92]. GATE has a very natural word processor like annotation of markables, highlighting them with a colour coded background (shown in Figure 2.1).

Figure 2.1: GATE - viewing and annotating markables [8]

Prior to rich web applications, early annotation tools in general were often desktop based Java applications that took the approach of offering a large single package with a variety of annotation options [9, 89].

MMAX is a multi-modal annotation tool supporting a variety of different annotation types. MMAX2 is a progression of MMAX, with similar visualisation. Both were implemented in Java for cross platform compatibility. Both are popular tools that are likely familiar to computational linguistics practitioners.

One of the most interesting features of MMAX for this work is their visualisation of markables, particularly when those markables are nested or overlapping, as these are properties of mentions (see Section 2.1.3).

Figure 2.2 shows nested and overlapping markables. In MMAX these are denoted by pairs of colour coded brackets. Hovering over a markable bracket highlights the bracket at the other end of the markable to make it easier to find the accompanying bracket. This visualisation is customisable. [93]

Figure 2.2: MMAX2 - Nesting and overlapping markables [9]

Annotation tools progressed into web-based applications [10, 11, 11, 91, 94], motivated by the need for large scale collaboration [11], and the move to micro-work crowdsourcing platforms such as CrowdFlower and Mechanical Turk [91]. At first these were read-only and mimicked the tools used before [94], but progressed into interactive modern applications featuring more graphically intensive visualisations [10, 11] than their predecessors.

BRAT is a popular web based annotation tool. BRAT leverages the modern capabilities of the web to offer an intuitive and attractive visualisation of a variety of annotations. BRAT shows markables in colour coded boxes. Spans of text are marked by clicking and dragging, in a similar behaviour to a word processor.

BRAT's interface also features in WebAnno, another web based annotation tool. WebAnno has its own backend that adds large scale project management, a different packing mechanism and built in crowdsourcing management. [11]

GATE also added support for large scale project management with GATE Teamware [90]. This takes GATE to the web, adds support for the different roles (e.g. annotator - provide annotations,

Figure 2.3: BRAT [10]

Figure 2.4: WebAnno [11]

managers - setup project and annotation guidelines, admin - manage services and accounts) required in large scale annotation, a distributed data store and provides an interface that is friendly to the non-expert.

Furthermore, in line with the progression to web based annotation and crowdsourcing the GATE offering was yet again expanded to enable easy design of tasks and deployment into crowdsourcing platforms directly from GATE [91] (shown in Figure 2.5). This is a more simplified interface with a more focused purpose.

This section has briefly looked at some of the most popular annotation tools used, how they visualise nested sequence labelling, and the evolution of those tools. There are several points that really stand out in the advancement of the tools as clearly desirable traits and several elements of functionality that are pursued by the designers. Many of these are clearly important to pursue in the design of a human computation approach. In summary, these are, non-expert friendly UI, web-based system, pre-annotation provided by pipeline, direct integration into microwork

Redacted - Copyright compatibility unknown

Figure 2.5: GATE crowdsourcing plugin [12]

crowdsourcing, project management and roles.

CHAPTER 3

HUMAN COMPUTATION

In the previous chapter we looked at Natural Language Processing (NLP) tasks, including the requirement for large amounts of annotated text to train statistical models (section 2) and how (section 2.2) this text is often hand-annotated [39] using pre-built annotation tools including MMAX2 [95], web-based crowdsourcing focused WebAnno [96], or the wiki style web-based *GMB Explorer* [97]. 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.

In this chapter we look at Human Computation, a method of combining computing effort with human effort, for tasks in which humans have a strong, and often natural ability to complete the task (sometimes through common knowledge). Or, "using human effort to perform tasks that computers cannot yet perform" [98].

For example, one proven method of reducing the time to gather natural language annotations is crowdsourcing [40]. It has been shown that this does not scale very well and that when attempting to build large corpora gamification approaches can be cheaper [19], provide more accurate results and better contributor engagement [42]. There are various human computation methods, each with different strengths and weaknesses depending on the task.

Human computation comes in many forms, with some work proposing a taxonomy. One dimension by which the genres is separated is how the task is incentivised or motivated. Citizen science is typically motivated by altruism or interest in the task, crowdsourcing on platforms such as Mechanical Turk is motivated financially, GWAPs are motivated by providing entertainment or fun [99].

In this chapter we look at some of approaches that have been used for natural language

resourcing with a particular focus on Games With A Purpose, as they appear to offer a lot of untapped potential.

3.1 Crowdsourcing

Howe gave the following definition for crowdsourcing [100]:

> crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call.

This presents an ideal opportunity for human computation by outsourcing a task that is divisible into small micro-tasks and paying individuals to complete the task for you [101].

There is a methodology that has evolved in crowdsourcing for annotation purposes. Over time, based on experience, prior investigations and an analysis of crowdsourcing projects, best practices have been proposed [102]. Project definition has been shown to be a crucial starting point to crowdsourcing. This includes determining the optimal crowdsourcing method and how/when to pay workers, which beyond cost effectiveness, can have variety of effects on the task quality [103] and speed of completion [104]. The next stage is the preparation of user interfaces and data. Choice of interface has also been shown to impact result quality, with binary choice interfaces for example, potentially resulting in spam [105]. Training and instruction in project execution has been shown to be particularly important, especially in what form that instruction is provided [106]. Crowdsourcing typically uses a non-expert/unreliable audience and receives multiple noisy judgements for a single data point. These are then aggregated to provide a single, high quality, judgement for each data point. Various methods have been proposed ranging from the simplistic majority voting, that considers all annotators to be equal, to more complex methods that model annotator ability and other features [107–109].

Multiple platforms offer paid services that seek to cater to some of the steps in the crowdsourcing process, particularly recruitment and remuneration. One very popular platform is Amazon Mechanical Turk Platform (AMT or MTurk). Taking the definition above, it's clear there are two key roles in crowdsourcing, the large network of people who may choose to complete tasks and the individuals or organisations advertising tasks to be completed. In MTurk, these are referred to as *workers* and *requesters* respectively. In MTurk, a single task completed by a single worker is termed a Human Intelligence Task (HIT). MTurk recruits workers by advertising HITs in a web portal. Pre-filtering of workers comes in the form of *qualifications*. A qualification in MTurk allows attributes to be stored in MTurk against a worker's profile. These can either be set by MTurk based on things such as the previous performance, or set by the requester. A qualification can also be linked to a set of test questions specified by the requester. MTurk offers a set of templates for requesters to use to deploy their tasks, or the opportunity to design and integrate a custom interface.

Crowdsourcing has already been demonstrated to be an effective and sometimes inexpensive method of performing text annotation for tasks that can be expressed as multiple choice (e.g. word sense disambiguation or event choice) [110]. However, many text annotation tasks, such as sequence labelling in the aforementioned mention detection and named entity recognition, have more intricate labelling requirements that are not natively supported by crowdsourcing platforms and require custom interfaces [12–14, 105, 111]. Methods of crowdsourcing and aggregating non-expert annotations in sequence labelling, particularly named entities, is an active area of research [12, 112].

As the popular crowdsourcing platforms do not natively support sequence labelling, Lawson et al. used a custom web interface (Figure 3.1) integrated into MTurk. They crowdsourced the annotation of named entities in a 20,609 document corpora comprising of emails (including mailing lists and newsgroups). The entire subject and body of an email was shown for each HIT. The initial pilot identified a tendency of workers under tagging resulting in low recall. They attributed this to the fact that an empty response is valid (although possibly not correct) and fixed payment model rewards workers regardless of the number of annotations. To address this, each HIT paid a low \$0.01 base rate, but gave workers a "bonus" (a mechanism in MTurk that allows the requester to send a specific amount of money in addition to the HIT reward). These ranged from \$0.01 – \$0.02 per entity, depending on the type of entity. They had 798 workers complete 169,156 HITs. [13]

Redacted - Copyright compatibility unknown

Figure 3.1: Lawson et al. - Crowdsourcing Interface [13]

In an experiment designed to compare two crowdsourcing platforms, Finin et al. collected named entity annotations for Tweets using both Crowdflower and Mechanical Turk. Each HIT

shows a single tweet tokenized. The worker may select from person, place or organisation radio buttons to individually identify each token as being of those named entity types. An additional checkbox allows the worker to express an uncertainty over their selected label. 251 HITs were submitted to MTurk and 30 to CrowdFlower. [14]

Redacted - Copyright compatibility unknown

Figure 3.2: Finin et al. - Crowdsourcing Interface [14]

Concluding Remarks There have been multiple efforts to crowdsource named entities [12–14, 112]. The key challenge is that whilst the crowdsourcing platforms do support classification tasks [110], they do not natively provide an interface for sequence labelling tasks. As such, experimenters typically provide their own. Apart from the challenge of creating a bespoke UI, a separate challenge is that unlike multiple choice tasks, it is possible that there are no annotations available. Workers may try to take advantage of the fact that being able to proceed without input is valid as a means of taking the quickest route to the reward or, at least not exhaustively explore all the annotations. Technically, given a sentence and some unknown set of mentions, a work. This has the end effect of dramatically reducing recall [13]. One approach used to reduce this has been to pay workers using the "bonus" mechanism in addition to the standard payment for completing the task for identifying sequences [13]. This may raise the cost of the task, make it appear less appealing (if the base rate is lower) or over-justify identifying sequences which in turn hinders precision.

3.2 Citizen Science

Citizen Science has been defined as "partnerships initiated by scientists that involve non-scientists in data collection" [113].

Multiple citizen science projects have been highly successful. This chapter is not an exhaustive list of citizen science projects, unlike we attempt for GWAPs for language resourcing later in Section 3.3.3. Instead, in this chapter we will look a few citizen science projects selectively that exhibit interesting characteristics or practices relevant to this work.

Citizen science projects rarely present a solely scientific interface. They often use ideas from gamification [114] and games with a purpose [16, 115]. In many cases it is difficult to separate them [47, 115]. The general distinction seems to be that the core motivation of citizen science is the project itself, rather than entertaining gameplay [47, 116]. Aside from the addition of gamification elements into crowdsourcing for motivation, the topic of motivation itself in citizen science projects is a subject of extensive research in its own right [116–121]. This is one of the core reasons for discussing citizen science projects here. Apart from their inseparable overlap with the other more game-like forms of human computation, is their informative attention to detail with regards to participant motivation, particularly when given specifically in relation to their gamification choices [115]. This interest has led to a proposed set of metrics from a project that cover player engagement [122] (we discuss these later in relation to our proposed metrics discussed in Section 5.1.3). These also refer to their marketing strategies, some of which have been very successful, which is also of interest in promoting GWAPs [119]. Generally speaking, investigation into the application of gamification methods in citizen science seem highly transferable to Games With A Purpose.

There have been crowdsourcing efforts that involve sequences, similar in some respects to the task we look at in Chapter IV. These are of interest, particularly in relation to their successes and interface [16, 123].

The rest of this chapter will look at these projects in more detail.

3.2.1 Zooniverse

Zooniverse is both a web-based portal and software framework for the implementation and deployment of citizen science projects. Zooniverse was home to 20 projects in 2014 [124]. Many of the projects on Zooniverse are very successful. In this section we will selectively look at two of these projects that are of particular interest to us because of their particularly high levels of contribution, attention to detail with regards to assessing player motivation and a careful and deliberate use of gamification elements.

3.2.1.1 Galaxy Zoo

Galazy Zoo [15] is a project to classify pictures of galaxies from the Sloan Digital Sky Survey (segmented imagery of galaxies of the northern sky) by their morphology or shape. The visual nature of these galaxies provides insights to their physical characteristics or movement.

Volunteers are shown an image of a galaxy, and in the original version, can choose one of six categories to best describe the image (shown in Figure 3.3).


Redacted - Copyright compatibility unknown

Figure 3.3: Galaxy Zoo (version 1) [15]

The results of the original version collected classifications for nearly 900,000 galaxies [125]. Galaxy Zoo exhibits a power law distribution of user contribution, with a small number of users completing more than 100,000 classifications each [15].

In the second version of Galaxy Zoo, more detailed classification was added for 300,000 galaxies. Volunteers provided information about the galaxy based on a decision tree of features that they effectively navigated by answering questions about the image [126].

Aggregation is performed using a weighted method that considers a volunteer's competency based on their agreement with other volunteers with minor differences between the two versions [15, 126].

Galaxy Zoo had a very successful initial marketing campaign. After the initial launch on a BBC radio programme the news quickly spread to a variety of news outlets. Tens of thousands of volunteers joined shortly after. [119]

Galaxy Zoo setup half-hour interviews with 22 of the volunteers (of their 160,000 at the time), carried out by either phone or instant messaging. They were asked a fixed set of questions (time constraint permitting), the responses of which were coded to provide 12 categories [119]. These categories were implemented as a survey in the form of a Likert scale asking how motivating each of the factors was. 10,992 of the 174,764 volunteers responded to the survey after data cleaning. By far the most motivational factor was contributing to science. This was followed by an interest in astronomy and the possibility of seeing galaxies that no one had seen before [120].

3.2.1.2 Old Weather

Old Weather is a citizen science project to digitally transcribe the images of handwritten historic (19^{th} century) ships logs. The motivation for gathering this data is to improve climate prediction

models and historical research. The target user base for the project was people interested in the climate change debate.

Although a citizen science project, Old Weather includes a selection of gamification elements to further motivate players and survey their players to test the impact of these [114].

Unlike other citizen science projects, Old Weather does not provide a web-based interface to interact with, but relies on players downloading and uploading spreadsheets to contribute. [127]

The gamification elements are largely based on rank. Volunteers progress through a series of titles for each ship they work on based on their contribution (once they have completed n items) to the logs for that ship. The top transcribers then compete over the top title, Captain. The goal of this mechanic was to encourage a player to dedicate their time to a specific ship and set of logs so they became familiar with those logs. A volunteer's position in the crew for a given ship is also shown.

Their survey showed that, like other citizen science projects, the intrinsic motivation of scientific contribution was a key motivator. The extrinsic gamification elements had a mixed impact.

As of December 2010 Old Weather had nearly 8,000 contributors [128]. There have been multiple projects, more recently, 16,400 contributed at least one page to the Royal Navy WW1 logbooks with all users contributing over a million in total [129]. However, the distribution of user contribution followed the common pattern, with the minority of the players contributing the majority of the work [129].

3.2.2 Phylo

Phylo is a web-based puzzle like citizen science game. It is of particular interest, as it has the objective of being game-like, and although in a different field, has a core mechanic that involves sequences.

Figure 3.4: Phylo [16]

From a design perspective Phylo attempts to decouple the scientific problem from the game

so provide an entertaining tetris-like game that does not require scientific understanding. The core game mechanic is the task itself, which involves aligning coloured sequences.

Phylo was marketed with media coverage and achieved the following player recruitment over the first seven months of activity [16]:

- 12,252 registered players
- 2,905 players who logged in multiple times
- 365,722 puzzles played
- 254,485 puzzles completed
- Registered users complete an average of 12.5 puzzles
- Returning users complete an average of 45

Phylo appears to exhibit the same power law distribution of contribution with the 10% top contributors contributing 80% of the solutions. [16]

3.2.3 Mark2Cure

Mark2Cure [123] is a web based system for crowdsourcing NER (NCBI Disease corpus [130] documents - a collection of documents annotated for disease mentions, in the first experiment). The system includes four short tutorials relating to the interface and the task itself. Feedback is given by paring the player with a partner and showing the player their partners annotations in comparison to theirs. Players may be assigned a document with a gold available. If so, the gold is used as the opponent. Players are awarded points based on agreement ($F_1 * 1000$) with their opponent (gold or previous player's game). If no gold or previous play through is available for a document then the player is awarded 1000 points and no feedback is given.

Mark2Cure recruitment method included Tweets, a mailing list of 100 interested potential users, an article in San Diego Union Tribune and a topical podcast. During a 28 day period, 212 users annotated 10,278 abstracts. The distribution of participant contribution saw the minority of the players contribution the majority of the work.

Average annotation quality was F_1 0.761, comparable with a previous MTurk crowdsourcing attempt.

3.3 Games With A Purpose

The term GWAP (Games with a purpose) was originally proposed by von Ahn [131] to describe a method of soliciting human computation as a byproduct of playing a game.

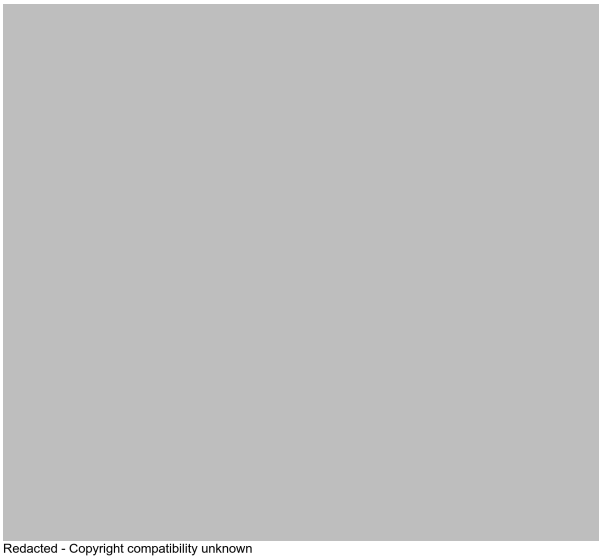

Figure 3.5: Mark2Cure Interface

The first key challenge of GWAPs was, given that the games were typically designed to solicit a label from a player, for which the true label was not known, *how can the game confirm a win state?* Von Ahn addressed this with by awarding a win state based on player agreement. To fit the notion of player agreement into games, Von Ahn et al. drew on their experience of creating such games to provide three patterns [43]:

output agreement game two randomly chosen players, are given the same input, and win by producing the same output, without communicating or seeing the other player's input.

input agreement game two randomly chosen players describe the inputs they are given, which may be the same or different. They can see each others descriptions, and must decide whether they were given the same input.

inversion problem game combines these two ideas, one player describes the input, and the other player, without seeing the input, must guess what it is from seeing the descriptors.

Variations on these continue to feature heavily in GWAPs.

The original GWAPs targeted image labelling tasks [132–134], but the concept was later deployed to tackle far more ambitious tasks in domains such as biology [16, 135, 136] and language resourcing (LR) [19, 24, 30, 137].

It has been said that one of the biggest opportunities for GWAPs to excel is in training their players [48]. Adding a training or learning element to crowdsourcing has shown to increase accuracy, but is difficult to do [138]. Games, however, have been shown to be an effective tool for teaching [139] and learning has been said to be a key part of the fun of a game [140].

From their inception it was evident there were various challenges to GWAPs with respect to design and evaluation. In relation to the patterns, the primary concerns were that players might communicate through a channel other than the game itself, or that in a multi-class labelling task that labels might converge to the most obvious set [43]. Naturally, as designers attempted to ask more of the GWAP paradigm, it became clear there were more challenges, but also new opportunities [48].

One such design challenge is the conflicting interest of tools and toys. Challenge in a game is artificially introduced in the form of internal goals, for the sake of the game. Tools however, are designed to reduce the challenge of achieving the external goal or purpose [141]. This dissonance in design is nicely summarised by the idea that if games were good tools/applications, they would simply be a "Win game" button [142]. In Games with a Purpose, by improving the *Game* one could negatively impact achieving the *Purpose* or by improving the users ability to achieve the *Purpose*, make the *Game* less entertaining. The former, addition of game mechanics that negatively impact the players ability to achieve the task, has been termed "orthogonal game mechanics" [48].

Costs are another challenge. There are few comprehensive cost analyses to draw upon from industry titles, but it is known that modern games typically have large budgets. Since the 1980's companies have been spending millions of dollars in the marketing of their games alone [143]. More recently costs often run into the tens of millions of dollars for development [144–146] and tens if not hundreds of millions for marketing [144, 147]. However, reduced cost over manual labelling was the main motivation given for the development of the original GWAPs [132, 133], been given as a motivation [24] and featured in GWAP evaluation [19, 27]. When creating a GWAP, one must keep the initial cost of game development low and development time fast, ruling out starting with a large project. If the project is overly expensive or takes a long time it may be faster and cheaper to use alternate methods (e.g. crowdsourcing). For example, the cost of annotation with *Phrase Detectives* (discussed in 3.3.3.4), one of the earliest GWAPs for gathering natural language annotations (more specifically anaphora), was equal to the cost using microtask crowdsourcing. However, the final projected cost for completed annotation of the corpora is 50% of the estimated cost of using crowdsourcing [19]. Whilst *Phrase Detectives* has evidently struck a good balance with their GWAP, had the creators invested much more in game development it may have been more cost-effective to use alternate methods.

In this section we will look at some of the major GWAPs and how they address these issues and opportunities.

3.3.1 Image Labelling

3.3.1.1 ESP & Peekaboom

In the ESP game set out to gather labels for images by having people label the images in a game. A pair of players play at a time, and win points by guessing common labels for the image. A set

of taboo words, that are not allowed to be used, are taken from previous rounds to gather new labels [46]. This is a type of output agreement game [43].

Similar to ESP, Peekaboom has users label objects in images. A pair of players play cooperatively. One player reveals small areas of parts of an image to show only the object they want the other player to guess, whilst the other player guesses the object they are trying to reveal. The player revealing the image may indicate to the player guessing whether they are "hot" or "cold". There are additional mechanics known as pings and hints. Additionally, in bonus rounds players are given the image and the name of the object, and both awarded points according to how close they click. [148]

Players of Peekaboom appear to experience what may be defined as Flow [149], with reports of a sense of loss of time reported.[148]. Apparently, every top player played over 53 hours. Furthermore, there was player feedback like:

> The bad point is that you look at your watch and eight hours have just disappeared!

Also, a loss of self-consciousness/environment awareness with users reporting playing until they sustained repetitive strain injuries [148].

3.3.2 Biology

3.3.2.1 FoldIt

FoldIt is another example of a GWAP with a very positive outcome [150].

Players interact and collaborate to directly manipulate protein structures and develop new strategies exploring not just validation, but all possible search spaces. The nature of the problem lends itself to a rich visual interpretation that fits well in a computer game context. Puzzles are introduced slowly, and carefully designed to be accessible to people without knowledge of the area. Technical constraints represented as geometry challenges. Some of the more common gamification elements include an interactive tutorial, points, leaderboard/rank system and player status. However, FoldIt also incorporates some less common design choices including chat/forums, collaboration, task ownership, giving players tools, and actively retooling based on observing players approaches. A survey revealed that contributing to important scientific discovery was a motivating factor, alongside immersive gameplay, exploration, points, and social interaction. FoldIt has been very successful in motivating players. The system seems to have around 657,650 registered users [1] which compares very well to other systems, especially considering the fact the game represents a highly complex task. [47]

[1] https://fold.it/portal/players

3.3.3 Language Resourcing

Table 3.1 lists a selection of GWAPs, their launch dates and the type of data they attempted to gather. In the rest of this section they are described in greater detail.

3.3.3.1 1001 Paraphrases

A phrase can be worded in a variety of semantically equivalent ways. Understanding these variations is particularly relevant for machine translation. Arguably the first GWAP to resource a linguistic phenomena, “1001 paraphrases” (shown in Figure 3.6) collected paraphrases from its players.

Players are shown a phrase and a set of obscured equivalent phrases that are revealed over time. They are required to guess the obscured phrases.

Upon a correct guess, the game proceeds to the next item. Upon an incorrect guess, more words of the obscured phrases are revealed to the player and the potential number of points they can win for that item is decreased.

Figure 3.6: 1001 Paraphrases[17]

Participation (over 15 months):

- 1,300 visitors (not all contributors)
- 20,944 distinct paraphrases

Game	**Year of release**	**Task**	**Data Collected**
1001 Paraphrases [17]	2005	Paraphrases	20,944 paraphrases [17]
Verbosity [18]	2006	Common-sense facts	7,871 facts [18]
JeuxDeMots [151]	2007	Lexical Relations	12 million relations [152]
Phrase Detectives [19]	2008	Coreference (anaphora)	Over 3 million judgements [70]
Sentiment Quiz [29]	2008	Sentiment	65,021 answers [29]
PlayCoRef [153, 154]	2009	Coreference	455 [20]
Place the Space * [20]	2009	Tokenization	N/A
The Shannon Game * [20]	2009	Word Substitution	N/A
PackPlay [21]	2010	NER	N/A
Jinx [22]	2010	WSD	N/A
MoleHunt [23]	2011	OCR checking	2.5 million tasks
Wordrobe [24]	2012	Various	41,541 answers [24]
Dr. Detective [25]	2013	NER	155 annotation sets [25]
Infection [26]	2014	Concept-Concept Relations	6,505 annotations from game (13,854 from crowdsourcing)
The Knowledge Towers [26]	2014	Image-Concept Relations	6,323 annotations from game (13,764 from crowdsourcing)
Puzzle Racer [27]	2014	Sense-Image Relations	16,479 images
Ka-Boom! [27]	2014	WSD	2,595 images
Quiz Bowl Coreference [28]	2015	Coreference	615 documents
RoboCorp [155]	2016	NER	3923 sentences [155]
uComp Language Quiz [29]	2015	Sentiment	9,320 answers (further 55,791 from crowdsourcing)
ZombiLingo [30]	2016	Dependencies	107,719 annotations [156]
Word Sheriff [31]	2016	Related words	246 games
TileAttack! [1]	2017	Candidate Mentions	
[NO NAME PROVIDED] [157]	2017	Common-sense facts	
Argotario [33]	2017	Fallacies	
RigorMortis [34]	2018	Multi-word Expressions	68 players [34]

Table 3.1: GWAPs for NLP - timeline and key characteristics
* while these games feature language annotation they were not designed to collect annotations

3.3.3.2 Verbosity

Verbosity is a two player cooperative web-based game with a purpose that gathers common sense facts. The design is inspired by the existing game Taboo™.

Two players are chosen randomly to play together, and assigned the roles, “Narrator” and “Guesser”. The “Narrator“ is given a word that is hidden from the “Guesser” and several templates they may fill in to describe properties or “facts” relating to that word. From these completed templates the “Guesser” must guess which word the “Narrator” was given.

The scoring system is cooperative and does not make use of penalties. That is to say that players’ are rewarded equal points for correct actions in the game, and not punished for incorrect actions.

Verbosity evaluated the accuracy of their responses by selecting a 200 of the 7871 facts gathered, and verifying them by 6 individuals. They reported 85% accuracy.

Verbosity made use of in game agents when players were not available.

Figure 3.7 shows the “Narrator” role playing.

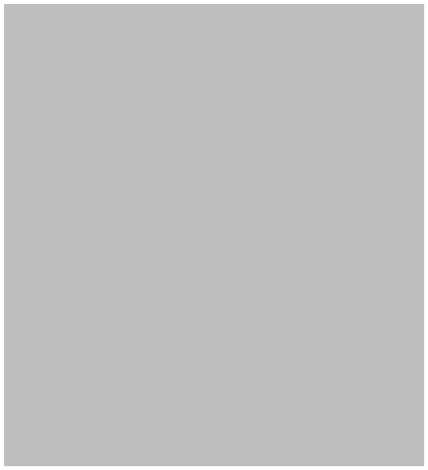
Redacted - Copyright compatibility unknown

Figure 3.7: Verbosity [18]

Participation:

- 267 players over 1 week
- Advertised on another game
- Some played over 3 hours
- average time of 23.58 minutes
- average of 29.47 facts per player
- 7871 facts in total

3.3.3.3 JeuxDeMots

The French game JeuxDeMots (translation word games/play on words/game of words), is a two player web based cooperative game in which players suggest relations based on a word they are presented with (discussed further in 2.4). The player suggests as many words as they can of the specified relation to the specified target word in under the time limit (around a minute). Answers are validated based on player agreement, with more points given for greater agreement. However, unlike ESP, players are not playing at the same time. The consequence of this is that at the time of playing, the first player receives no score, but is instead later notified by email. Like ESP, the game makes use of taboo words to encourage new suggestions unknown to the system. According to Lafourcade, despite the time delay, this still did seem to fulfil players desire for social interaction with players requesting the email addresses of the other players they had reached agreement with on puzzles.

Audience:

- 100 players/1 month
- Word of mouth only
- 20,000 relations

3.3.3.4 Phrase Detectives

Phrase Detectives [19] is a one player web based game in which players label anaphora relations. Phrase Detectives makes use of a tutorial before players are progress on to the main game. The main game is split into two tasks, “Name-the-Culprit” (Figure 3.8) and “Detectives Conference”, which are annotation and validation respectively. These two game modes serve as a means of quality control [158]. In the annotation mode, the player directly annotates the antecedent of the anaphora relation, mark if the item has not been mentioned before, or whether the highlighted item is a property. If all players agree on the antecedent then no further action is required for that text. Otherwise, the text is displayed in the validation game mode. In this mode the players confirm an existing annotation. The pre-processing required to support the annotation (i.e. identification of candidate mentions) is provided by an automated pipeline. To address possible errors passed on by the pipeline Phrase Detectives provides the possibility for players to skip the task, and specify the nature of the error.

Phrase Detectives is described as a GWAP. Whilst Phrase Detectives does incorporate game-like aesthetics, game-like mechanics such as points/leaderboards, levels and giving the player the role of playing the a Detective, the player is clearly directly performing an annotation task that one may argue could be described as a high quality, comprehensive example of gamification as opposed to a GWAP.

Figure 3.8: Phrase Detectives - Annotation [19]

Aside from incentivising play through entertainment value, Phrase Detectives also made use of financial prizes awarded to top players.

There is an fairly comprehensive study of cost-effectiveness for Phrase Detectives in which expert annotation, crowdsource annotation and the application of Phrase Detectives are included. This analysis highlights the main difference between the other approaches and GWAPs, in which one hopes the high initial investment of the development of the GWAP will hopefully pay off in the long term, verses the ongoing constant investment of other methods such as crowdsourcing. The costs per completed markable (sufficient annotations for aggregation) are given as:

Expert Annotation $3

Supervised and Trained non-experts $1.2

Crowdsourcing $1.2-1.3

Phrase Detectives $0.47

Audience:

- 8,000 players registered/December 2008 to January 2012
- 3,000 went beyond initial training phase
- 5,000 hours of work
- 450 annotations per hour
- average lifetime play of 35 mins 5 seconds

3.3.3.5 PlayCoRef

The PlayCoref game (shown in Figure 3.9) [153], is a game designed to annotate texts with coreference information for both English and Czech. The game can operate as a one or two player game, in which players annotate coreference chains. The document is presented to the players for annotation sentences at a time, and players attempt to annotate as many instances of coreference possible within 5 minutes (or the document is complete). The creators compare PlayCoRef with Phrase Detectives. They present various differences including their preference for instructions over tutorial and two-player play over one player.

The scoring combines multiple F-Measures based on agreement with other players, manual annotators, and a pipeline, weighted by the player annotating 12 sentences. PlayCoRef states its goal for scoring is to motivate as accurate contribution as possible by tying the score as close as possible to the accuracy of the annotation.

The effectiveness of the game is evaluated against expert annotated data sets from PDT 2.0 (Prague Dependency Treebank 2.0) [2] and MUC-6 (the sixth Message Understanding Conference) [3]. Annotations are undirected - they may be anaphora or cataphora.

"PlayCoRef" was released on a portal accompanied by two other games. These games were not intended to gather annotations, but rather motivate players to visit the portal. However, they did feature text annotation as a core game mechanic. "Place The Space" (Figure 3.10) is a tokenization game in which players insert spaces, and "The Shannon Game" (Figure 3.11) is a game in which players insert missing words.

3.3.3.6 PackPlay

PackPlay [21] is a suite of games focused on named entity identification (shown in Figure 3.12. These are "Entity Discovery", "Name that Entity" and "Vocabulary Builder".

In the multiplayer *Entity Discovery* game players highlight an entity then click on a class. In the absence of a second player, an automated agent is used that selects a previous players responses. The player is not aware they are playing against an agent.

[2] http://ufal.mff.cuni.cz/pdt2.0

[3] http://cs.nyu.edu/faculty/grishman/muc6.html


Redacted - Copyright compatibility unknown

Figure 3.9: PlayCoRef [20]


Redacted - Copyright compatibility unknown

Figure 3.10: Place The Space [20]

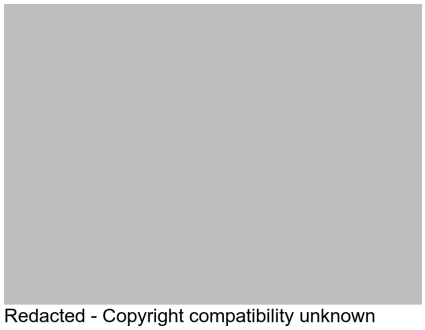
Redacted - Copyright compatibility unknown

Figure 3.11: The Shannon Game [20]

- 8 players
- 29 games (mean 3.62)
- 291 annotations (mean 40.85)

Name that Entity is a multiple choice game. Unlike Entity Discovery players are not asked to the text selection, only select the class. This game serves to check the validity of the answers given from the Entity Discovery game.

- 8 players
- 20 games (mean 2.85)
- 195 annotations (mean 27.85)

Figure 3.12: PackPlay [21]

3.3.3.7 Jinx

Jinx [22] (shown in Figure 3.13) is a 2 player cooperative game that looks to gather information relating to word sense disambiguation. Judging from the example in the paper this appears to relate to homonyms (words with identical spellings and pronunciations, but different meanings), and homographs (words with identical spellings and different pronunciations, but also different meanings). In the game, both players see a sentence with a single word and suggest related words to disambiguate the word. For example, in the sentence ”The wind was bad.”, a player may suggest the term “weather”, indicating that the homograph ”wind” as in “wind the cable in a loop”, was not meant. This hypernym would clarify that ambiguity. Players receive points when they guess an identical related word, scoring more points the earlier in the round they make their guess. Players are presented these words in repeated 30 second rounds. Using WordNet synsets, 54% of the suggested tags uniquely identified the original word shown to the player, giving a positive outcome for the method. Seemakurty et al mention an interest in extending the game to include taboo words, similar to ESP, to encourage users to make new suggestions.


Redacted - Copyright compatibility unknown

Figure 3.13: Jinx [22]

3.3.3.8 Mole Hunt & Mole Bridge

Mole Hunt and Mole Bridge are both games for OCR (optical character recognition), created for the purpose of digitising newspaper archives from the National Library of Finland. [23]

The Mole Hunt game takes its inputs as single images of word tokens, with a possible answer. The output is boolean with the player marking an item as correct or incorrect. The game has a whack-a-mole appearance, with moles popping up holding the image, and the digitized word. The players answers are confirmed at the end of a round of play. An animation is shown in which a mole moves from left to right beneath some flowers. Each flower represents an annotation. If the annotation was correct, the flower blooms, if incorrect, the flower is eaten by the mole.

The Mole Bridge game takes its inputs as single images of word tokens without suggestions, and outputs digitized tokens. The player is required to type the word without seeing any digitized possible answer. The players answers are confirmed during play. Moles cross a bridge. When an annotation is made, it forms a part of the bridge with a wood appearance whilst unconfirmed. Once confirmed to be right or wrong, this either turns to metal or explodes. This effects whether the moles successfully cross the bridge.

To offer the player feedback on whether they are right or wrong, answers are compared with other players answers. When those existing answers are not available, validation is used instead (i.e. a known token will be displayed to the user). A similar technique is used to filter out problem players that are not contributing correct answers during the early rounds of the game.

In under two months the games attracted 4,768 players, with the majority coming from Facebook. The data received had over 99% accuracy.

The median play time was 9 minutes 18 seconds. The most active 1% of users contributed approximately 33% of the work.

3.3.3.9 Wordrobe

Wordrobe (shown in Figure 3.16) is a collection of web based 1 player games with a common design. The core game mechanic sees a player presented with a few sentences, asked to choose an annotation from a series of multiple choices, and express their confidence in their answer using a slider control. Players are computed based on agreement with other players (majority voting)

Figure 3.14: Mole Hunt [23]

Figure 3.15: Mole Bridge [23]

and the bet the player placed when providing the annotation. The answers to the multiple choice questions are generated by an automated pipeline. In some cases, the automated pipeline may generate errors and the correct answer will not be available. To address this, the player is offered the opportunity to skip questions.

- 41,541 answers
- 962 players

3.3.3.10 Dr. Detective

Dr. Detective is a 1 player web based game in which players annotate domain specific named entities in medical texts. Dr. Detective models a documents difficulty as being the normalized vector of the number of sentences, the number of words, the average sentence length, the number

Figure 3.16: Wordrobe [24]

of item types and the readability of the document (using the SMOG measure [159]) [25]. The selection process is then to find the item with the smallest difficulty increment from all items that have a difficulty greater than or equal to the current item, excluding the current item. In this work, the authors mention that they believe computing difficulty based solely on textual metrics was a weakness and that the system would benefit from a domain specific metric of difficulty.

Assigns items based on a heuristic of their difficulty that is based on the SMOG measure, the number of words in the sentence, number of sentences in the document, average sentence length and the number of UMLS concepts present.

The scoring system is set out with the goals of rewarding players that perform in a way that is beneficial; award high agreement, but also new contributions; penalise incorrect contributions and reward a score proportional to the task difficulty. To this end, the scoring calculation is split into 4 components that model, agreement, the novelty of the contribution, the consistency of the user in consecutive tasks and their loss if the item is incorrect.

The only data player data available for Dr. Detective appears to be from a pilot study conducted with medical professionals. They report the following player engagement stats:

- 155 annotations sets collected
- 11 participants in total (10 playing full game version)

3.3.3.11 Infection

The goal of the game Infection [26] is to add to a knowledge base/semantic ontology like WordNet. Infection ties together associations between concepts. The game is a top down shooter game in which the player must defend a city from zombies, without killing humans (or the uninfected). The challenge for the player is to determine who is infected. The player is given a passphrase,


Redacted - Copyright compatibility unknown

Figure 3.17: Dr. Detective [25]

that their character shouts. This is given as a challenge to other characters, who either reply with a related phrase (human), or non-related phrase (zombie). The player can then kill the zombie indicating a negative annotation, or allow the zombie to proceed indicating a positive annotation. Killing humans, beyond a certain threshold, results in the player losing. Unlike ZombiLingo [30], there is no extensive explanation required here. The task has clearly broken down to a level that most people could easily understand, and the process of annotation is binary with players being presented with optional answers, rather than having to go looking for them. This design takes a similar strategy of omitting complexity in favour of attempting to recruit a large number of non-expert players similar to Wordrobe [22].

The design of these games set out a new and perhaps ambitious goal of making the games sufficiently general that only the game data has to be changed, to allow the game to be used for a different annotation task.

Interestingly, in an analysis of another game that injects unrelated mini games around the task, Vanella et al describe the game [160] as follows [27]:

> the annotation task is a chore the player must perform in order to return to the game, rather than an integrated, fun part of the game's objectives, which potentially decreases motivation for answering correctly

Vanella et al clearly are considering game design and give an objective discussion of similar attempts and mistakes made following on from their already mentioned previous work in which

they mentioned their desire to be more game-like [27].

Figure 3.18: Infection [26]

Infection was marketed via social networking sites and online forums in two forms. One with paid prizes, and one without. In summary, the resulting stats were as follows:

- 252 players (89 free group, 163 paid group)
- 6505 annotations (3150 free group, 3355 paid group)

A further 13854 annotations were crowdsourced from 290 workers using the crowdsourcing platform Crowdflower.

3.3.3.12 The Knowledge Towers

The Knowledge Towers is similar to a dungeon crawling role playing game. The goal is to associate words with images. The player proceeds through rooms facing enemies and finding rewards in chests as is typical of the genre. However, unlike games of that genre, the player must fill their inventory related to a word specified by the game. They have limited space in their inventory, encouraging them to find the most relevant images [26]. Having gathered all the relevant images, which represent their annotations, they proceed to face the dungeon boss.

The key difference in terms of design mechanics between the two games appears the be the presence of a time element in the Infection game. The Knowledge Towers gives players as much time as they want to complete the challenge. The effect this had on either motivation or accuracy in comparison is unclear.

The Knowledge Towers was marketed via social networking sites and online forums in two forms. One with paid prizes, and one without. In summary, the resulting stats were as follows:

- 197 players (100 free group, 97 paid group)
- 6323 annotations (3005 free group, 3318 paid group)

A further 13764 annotations were crowdsourced from 1097 workers using the crowdsourcing platform Crowdflower.

Figure 3.19: The Knowledge Towers [26]

3.3.3.13 Puzzle Racer & Ka-boom!

Puzzle Racer is 1 player web-based game inspired by games such as Temple Run and Subway Surfers. Unlike others games, Puzzle Racer tries to be game-like to ensure it was accessible to all (similar to FoldIt's approach [47]), as per the common definition of GWAP [27].

In Puzzle Racer, there are three stages per challenge. In the pre-race stage a player has to identify a common theme from three images. In the race stage the player has to repeatedly navigate a series of gates, selecting one of three images that continue this theme. These are a mix of golden gates, which validate player understanding, and mystery gates, that allow the player to contribute influence the score that indicates a link between these picture and the word sense. At the end of the race, the post-race stage allows the user to write a word that they feel describes the theme of the images throughout, for double points. Additional game elements noted include leaderboards and unlockables.

The output of Puzzle Racer is a WordNet word sense to image mapping.

Ka-boom! is a word sense disambiguation game in the style of Fruit Ninja [27]. Players are presented with photos indicative of the word senses. They are given a sense to match, and destroy photos that do not match the sense to reject them.

Jurgens et al arrive at the conclusion that this demonstrates some cost reduction in using GWAP. However, they also note other factors may have affected this. They used just 126 undergrad students, presumably from the same institution, which they paid with gift cards (prize based incentive - top ranking players only), and compared them with CrowdFlower users (crowdsourcing workers).

For Puzzle Racer:

- 126 undergrad students for participants
- 20,253 ratings across 16,479 images

For Ka-boom!

Figure 3.20: PuzzleRacer [27]

- 19 players - fluent English speakers
- 2594 images

Figure 3.21: Ka-boom! [27]

3.3.3.14 Quiz Bowl Conference

Quiz Bowl Coreference is a single player game in which players annotate coreference chains in data used for Quiz Bowl questions. In a Quiz Bowl game players attempt to guess which entity is described by the Quiz Bowl clue. As such, unlike the typical newswire data, Quiz Bowl questions are deliberately rich with co-referent phrases. The web application, was advertised to participants in a Quiz Bowl tournament. Over the course of one month, there were 615 documents tagged by 76 users. The top 5 annotators tagged 342 of the 651 documents.

An Active Learning approach is used to reduce the amount of data needed to be tagged to produce a useful corpora by switching between annotation and training to discover the documents that are most useful to the classifier, and annotating those first.

Quiz Bowl Coreference does not describe itself as a GWAP, or have any game mechanics, but is conceptually similar in that it does claim to entertain its players to motivate annotation. [28]

- 615 documents
- 76 users (one month period)

Figure 3.22: Quiz Bowl Coreference Game [28]

3.3.3.15 uComp Language Quiz

uComp Language Quiz is a web based 1 player game for sentiment analysis.

From October 2015 to March 2017, the Language Quiz attracted 2,688 users, 959 of which it converted to active users that submitted a valid answer. Players are recruited via the Crowdflower crowdsourcing platform.

- 2,688 users (of which 1,916 were crowdsourced)
- 2,150 conversions (to registered users)
- 65,021 valid answers (9,320 from organic players)

3.3.3.16 ZombiLingo

ZombiLingo is a web based one player GWAP launched in 2014 with the goal of annotating dependency syntax structure in French. The creators identify multiple motivational game-centric design patterns, but prefers a people-centric approach known as MICE [30] - a persuasion framework for the CIA to recruit agents [161]. The idea behind MICE is that people typically


Redacted - Copyright compatibility unknown

Figure 3.23: uComp [29]

have some weakness or vulnerability that can be leveraged and that either Money (providing safety, shelter, luxury), their Ideology (appealing to their beliefs), Coercion (typically black mail), or their Ego (the excitement associated with the role) can be applied to leverage that [161]. This contrasts heavily with Von Ahn's approach to simply leveraging the players "desire to be entertained" [43]. In ZombiLingo, Money is related to the in-game currency, Ideology is related to the zombie theme, Coercion is related to the scoreboard, and Ego is related to the in-game avatar. Despite this novel application of the MICE framework the resulting game mechanics are not dissimilar from other games. From the launch of the game up to February 23rd 2017, 987 players produced 214,082 annotations. 20 of these players had played on more than 5 days or completed more than 500 annotations [162].

3.3.3.17 Word Sheriff

Word Sheriff is a game to crowdsource words that are similar, by definition.

It is a multiple player game with two roles. The *narrator* role, for which there is one player, is presented with or phrase. They provide a series of clues to the *guessers*, for which there is one or more players. For each clue, each guesser may provide one guess as to what the target word or phrase is. The first correct guess wins. [31]

The original experiment saw 246 games players by approximately 100 players.

In 2017, the team delivered a summary of what they felt were the shortcomings of their initial implementation as a lack of social interaction, mobile device compatibility and unconvincing AI players which they planned to address with Facebook integration, an updated user interface with

Figure 3.24: ZombiLingo [30]

Figure 3.25: Word Sheriff [31]

responsive web design (Figure 3.26) and an AI with a more diverse vocabulary.

3.3.3.18 Argotario

Argotario (shown in Figure 3.27) is a web based game that annotates fallacies. It is described as a serious game. Whilst most GWAPs seek to provide some educational experience, if only for giving the best possible game experience for players and gathering the highest quality annotation, education is given as one of the main purposes of Argotario.

Argotario has both one and two player modes. The two player mode Argotario, the first player makes a fallacious argument, of a fallacy type specified by the game. The second player has to guess that fallacy type.

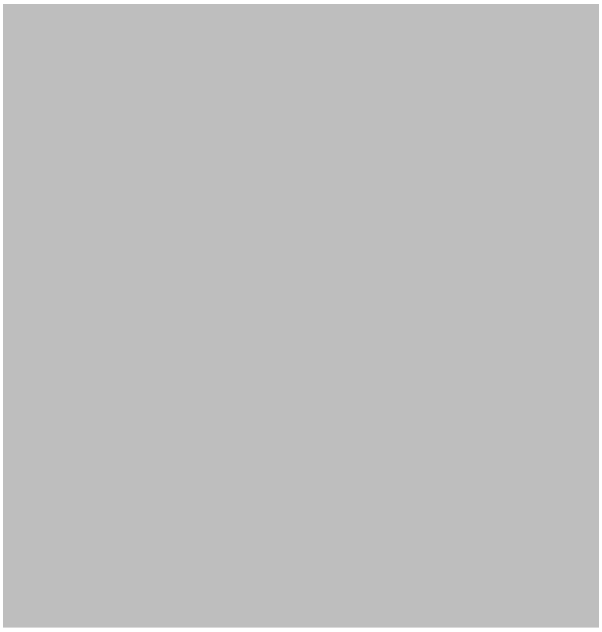
Redacted - Copyright compatibility unknown

Figure 3.26: Word Sheriff 2.0 [32]

There are several game elements used in Argotario. These include the typical gamification mechanics such as points, leaderboards and avatars, but also the representation of different game modes in a virtual world/treasure map.


Redacted - Copyright compatibility unknown

Figure 3.27: Argotario [33]

3.3.3.19 RigorMortis

In RigorMortis [34], players use a TileAttack like interface to label multi-word expressions (MWEs) in French (e.g. *"in order to"* or *"by the way"*). The experiment uses a small expert hand

labelled corpus of 10 sentences.

The task was publicised on social networks and language processing mailing lists. They recruited 68 players.

They showed volunteers with no prior training can identify at least some MWE's.

Figure 3.28: RigorMortis [34]

3.4 Concluding Remarks

This chapter has examined different types of human computation with particular attention to applications of human computation for natural language annotation, especially GWAPs.

Regardless of which method is used, a common pattern that emerges is uneven contribution both in GWAPs with the minority of the volunteers providing the majority of the work [19, 23, 30] and citizen science [15, 16, 123, 129]. Some feature mechanics that appear to deliberately entertain this distribution of work, such as Old Weather's application of gamification to add competition for the top captain role [114].

There is a large crossover between citizen science projects often feature gamification elements, or are describes as citizen science games. The latter bare resemblance to the GWAP model or are sometimes described as GWAPs in citizen science [115].

In terms of design, the presentation of citizen science projects ranges from quite utilitarian [123] to quite game-like [16, 47]. The more game-like approaches, such as Phylo, decouple their task from the game reducing the need for scientific understanding instead relying on the volunteers pattern matching skills [16]. However, they have a natural graphical representation that is not clearly available from some of the other tasks. In between, are the projects that maintain a utilitarian interface, but with the addition of gamification [114].

With respect to GWAPs specifically, several common challenges, patterns and trends have emerged.

One trend is the shift from unskilled homogenous tasks to highly skilled tasks. This has been said to be GWAPs greatest opportunity to excel over other methods through learning opportunities that naturally occur in games [48]. This brings us to another trend in GWAPs, the addition of game-like training. This includes tutorials [19, 30], skill based systems [30] and dynamically assigning tasks based on their complexity [25]. This also presents the question of,

which other concepts may be taken from games to address the problem of varied task complexity as GWAPs are used to tackle more ambitious tasks.

Multiple interesting patterns have emerged in game design. The earlier games, and many since, have favoured more of a lighter gamification approach over creating full GWAPs [19, 25]. These games often openly present their tasks, while others try to hide this with the goal of making contribution more accessible to non-experts by reducing the annotation task to multiple choice questions [24]. The original GWAPs took inspiration from, or adapted existing game designs [163], this has been revisited with more game-like adaptations and designs [26, 27], although so far such attempts have always targeted tasks that closely relate to image labelling. Another unique design is rather than having the players produce work as a byproduct of play, having them work to enable play [155]. Whilst many interesting approaches have been taken, it seems important to note that few of them have surpassed small scale experimentation, making it difficult to evaluate their success. More game-like GWAPs clearly have a multitude of benefits and that is evidently a direction that is being explored. However, the mechanic of text annotation in truly game-like GWAP seems to remain a challenge.

Furthermore whilst many different and interesting designs have been proposed since the original GWAPs, the majority of games, including those that propose new designs [155], have tended to focus their evaluation towards the accuracy of their final results rather than the games ability to recruit or retain players [25–27]. In contrast, the more game-like citizen science projects often look closely at what motivates their volunteers through use of surveys, interviews and statistics [114, 119, 120]. Although in that type of game the core motivation for participation does appear to be scientific contribution and an interest in the topic [114, 120]. This demonstrates a need for an assessment of the key performance indicators in GWAPs, and the development of a set of metrics to measure them.

In terms of accuracy, the range of approaches taken, differences in annotator groups/group sizes, tasks and corpora, make it difficult to compare different human computation approaches. As a result, it's difficult to say conclusively whether one approach fares better than another in terms of the quality of resources they can produce.

In terms of experimentation and player recruitment, several approaches have been taken. By far the most popular approach is a small scale study of directly contacted participants, from sources such as the local institution [26, 27], a relevant organisation [25], or players of a previous game [18]. Another approach is the use of crowdsourcing to recruit players into the game with the goal of converting them to unpaid players [29]. Some games have used traditional marketing approaches such as Google Ads [33]. Marketing methods are clearly an important consideration to GWAP developers and can have a great impact in terms of their user base and the final results they achieve. This demonstrates that the aforementioned metrics would benefit from being able to reflect the effect of a specific marketing campaign or audience as part of their measurements.

Part III

The Proposal: A Road Map

This part serves as a map of sorts for the way in which the various experiments and projects discussed in this work are arranged and the methods use to evaluate them.

This map identifies the various experiments and projects that make up this work are deliberately arranged and the methods used to evaluate them.

There are three chapters. The first describes the idea that underlies the arrangement of the three games as a gamified pipeline. The second discusses the adaptation of F2P metrics we used for evaluating our GWAPs. The third chapter discusses our informative preliminary experiment into a game-like GWAP for tokenization.

CHAPTER 4

GAMIFYING THE PIPELINE

As discussed in Section 2, Natural Language Processing involves a number of steps (with some exceptions, e.g. end-to-end systems, see Chapter 2), each transmitting a series of interpretations. Both Natural Language Processing and annotation is typically carried out using a pipeline of steps relating to a set of dependent. This led to the development of NLP pipelines with a collection of modules such as the popular Stanford CoreNLP pipeline [50], a subset of the available steps consist of tokenisation module, sentence splitting module, a part-of-speech tagging module, a named entity recognition module etc. Each of these steps requires information from the previous step.

For example, a pipeline might look like:

Tokenization → Part-of-Speech Tagging → Mention Detection → Anaphora

However, most games focus on one aspect of interpretation and use an automated pipeline to gather the prior interpretations that are required to support their in-game annotation. Phrase Detectives for example, as previously discussed, is a game for marking anaphora. Anaphora are linked mentions, so those mentions must be discovered first. In this case they are provided by an automated pipeline [70] (the Berkeley Parser [164] for the mentions, and other pipelines for previous steps). Developing an adequate pipeline, particularly when attempting to introduce other languages, has been said to be one of the most challenging parts of developing such a game [165]. An automated pipeline cannot be expected to be 100% accurate, so Phrase Detectives features a comment box, so that players can inform them of problems with the mentions. Some 10,000 comments are submitted a year, but only 3,000–4,000 can be processed. This significantly impacts throughput and creates bottlenecks.

In this section we explore the notion of gamifying the entire pipeline with games supporting tokenisation, tagging and nested sequence labelling. The majority of labelling interactions required text annotation tasks.

Aside from catering to the range of annotation tasks, we also enable progression in, and between these games. Having achieved a high level of understanding in pos-tagging for example, a non-expert is better prepared for a task such as noun-phrase segmentation.

Having introduced this pipeline we discuss the prototype for the first game in pipeline, "The Logging Game".

4.1 Related Work

The main example of a pipeline is the General Architecture for Text Engineering (GATE). GATE provides the complete architecture for integration of automated components to build full text processing pipelines, use modules independently of the pipeline, and exporting those applications designed in GATE for external batch use. It also supports annotation or production of resources, and interaction between the two (i.e. correcting an automatic pipeline rather than providing annotations from scratch). GATE is ready to use, with an easy to use graphical user interface, and high performing components that can be easily replaced as the state of the art advances. [8]

GATE also includes a plugin that enables integration of annotation steps into platforms such as Crowdflower [91].

Various other pipelines have been created, each with different core goals. Stanford CoreNLP for example [50], aims to be streamlined and powerful (state-of-the-art), but as a trade-off does not offer features offered by competitors (such as annotation), a GUI to design pipelines or easily replaceable components. It's goal is a system that offers high performance and ease of use from the command line or API [50]. NLTK takes the ease of use goal further by sacrificing both features and state-of-the-art performance to provide clear easy to understand components and pipelines that can be used to teach in a classroom setting [166].

Right at the other end of the spectrum are pipelines that are less turn-key, but instead try to provide a comprehensive model for incorporating multiple components and facilitating the communication (e.g. UIMA [167]). UIMA does not provide the components itself, and has been criticised as being an "empty toolbox" [50].

Another example is the Groningen Meaning Bank (GMB) project. They used NLP tools to get an approximation of annotation at different levels, then pieces of information gathered from crowdsourcing annotations they refer to as Bits Of Wisdom (BOW) [168]. One of the methods used to crowdsource BOWs is a suite of GWAPs called Wordrobe that gather annotations at various steps in the pipeline by presenting all tasks as multiple choice questions [169] (discussed in Section 3.3.3.9).

4.2 A GWAP Pipeline For Training and Annotation

As previously mentioned, to avoid bottlenecks, it is not sufficient to design a game to cater to one task in the pipeline, we need a comprehensive set of games that can feed into each other. Collectively our games provide annotations that could be consumed by the following game to annotate a corpora for tokenization (*The Logging Game*), segmentation (*The Logging Game*), tagging (*WordClicker*) and nested sequence labelling tasks (*TileAttack*), which covers the majority of annotation tasks.

Each GWAP in our pipeline not only focuses on a different labelling task, but for research purposes, a different set of GWAP interests, exemplifying a key ideas in addressing the challenges in GWAPs for language resourcing. This section will serve as a map of sorts to help the reader navigate the games, their interests and how they link together. Broadly speaking, there are 3 dimensions to the pipeline:

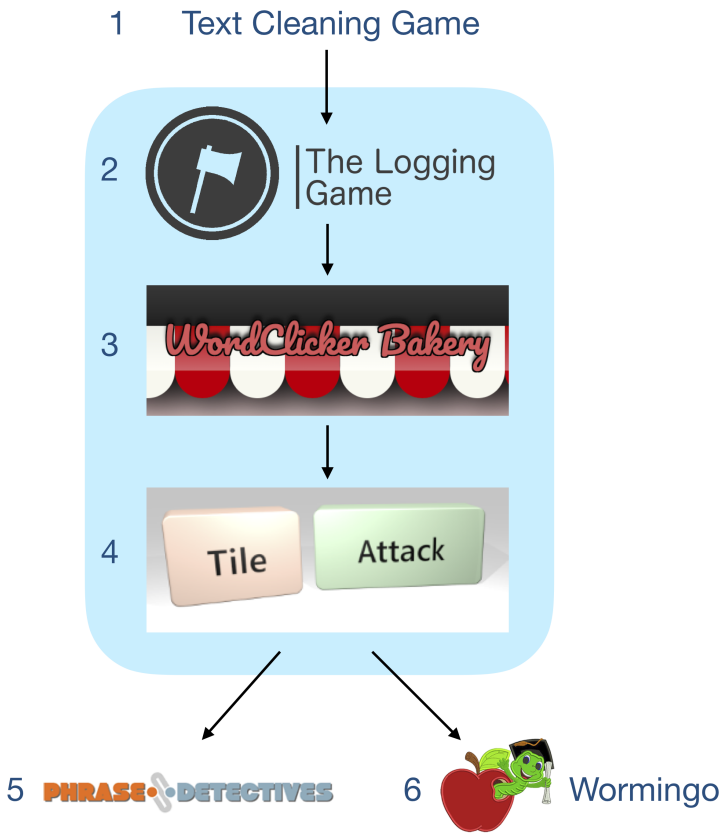

Figure 4.1: Pipeline

Firstly, we focus on efficiently using a GWAP to gather accurate annotations from non-experts comparable with expert annotators. To achieve this we present novel strategies to aggregation and automated pipeline correction. We test these at scale using micro-task crowdsourcing with

our game *TileAttack*.

Secondly, there is the notion of training and resource allocation. The games follow a sequence of increasing complexity to familiarise players with the simple and fundamental tasks before progressing them onto tasks of increasing difficulty. This is crucial to ensure that the right players/workers are completing tasks in line with their current competencies to maximise resources and make best possible use of non-experts. We demonstrate this through *WordClicker* and *TileAttack*. *WordClicker* is designed with training and progression on to another game in mind, and invites players to progress to *TileAttack* once they have reached a certain level. In *TileAttack* we look at how to progress non-experts in a task through increasingly difficult cases to encourage learning and maximise the returns. Both *TileAttack* and *WordClicker* feature interactive tutorials.

Thirdly, there is the investigation of game design. A GWAP cannot be considered truly successful, as per the original definition, unless it is effective at gathering players to perform annotations as a by-product of play. However, like many tasks, text annotation is not an easy or obvious fit into games. In our design of *WordClicker* we identify these challenges and present a pattern to address them.

CHAPTER 5

TOWARDS A NEW SET OF METRICS FOR GWAPS

As discussed in our review of the GWAPs in Section 3.4, evaluation, even in GWAPs proposing novel design ideas, is typically focused on accuracy over player engagement, recruitment or retention. We believe this could be a barrier to understanding and advancing design in GWAPs.

In this chapter, we look at other methods proposed and reported. We then discuss the core aims and objects of GWAPs, as given by GWAP projects, to identify their key performance indicators. Finally we propose a new method ourselves, the adaptation of Free-to-Play metrics.

5.1 Related Work

5.1.1 Von Ahn's Proposed Metrics

Von Ahn saw GWAPs as a means of running algorithms (games) on human (processors) [131]. He wrote of the design of those games being like algorithms, in that they can be proven to be correct, and their efficiency analysed [131]. The idea of examining the evaluation of GWAPs, or design of metrics for such a purpose, is not novel. Following the success of the original GWAPs and two others (The ESP Game [46]; Peekaboom [148]; Phetch [134] and Verbosity [18]), their creator, Von Ahn, created a template for the creation and evaluation of GWAPs to serve as a starting point to formulate a generalized approach to applying his method [170]. Comparing the interests of a GWAP to that of an algorithm, he defined three original metrics [170]:

Throughput average number of problem instances solved per human-hour;

ALP Average lifetime play is the average (across all people who play the game) overall amount of time the game will be played by an individual player; and

Expected contribution throughput multiplied by ALP.

By Von Ahn's admission, these did not comprehensively capture all aspects and in general, much work remained to be done. When discussing these, Von Ahn identified two immediate shortcomings. They did not cover **popularity**, or **contagion** [43].

5.1.2 Performance Indicators of GWAPs

As the paradigm has evolved and been applied in different ways, in this section, we examine the literature for the current and recent key performance indicators that are reported by GWAPs.

The most commonly reported measure is the **quantity** of data, both in terms of work/user labels [19, 21, 26, 27, 151], and the resulting annotations data from those annotations [17, 19, 24, 25, 27, 28, 30, 151, 153, 155, 157]. The next most commonly reported is the quality of that data, or **accuracy** [19, 21, 24, 26–28, 30, 153, 155, 157]. **Cost** is occasionally mentioned, particularly in comparison to alternative human computation methods (such as microtask crowdsourcing) [19, 26, 27]. The **number of users** is often mentioned, but this is typically only in terms of methodology, discussing the experiment participant pool [17, 22, 24–26, 31, 153], rather than in terms of the games ability or a marketing strategy to recruit participants, that is reported less [30, 33, 151, 155, 157]. There are fewer measures relating to the game elements, or the users enjoyment in general. The **number of games played** by a user is rarely reported [21, 31, 151, 157], as is **session length** [33, 151] and **retention** [33].

Where such measures are provided, they are generally evaluated over a given **time period** [17, 19, 28, 30, 151, 155, 157].

The currently reported metrics are largely covered by those proposed Von Ahn, with a lesser focus on engagement and enjoyability factors. We aim to cover all the performance indicators given with our proposed framework.

5.1.3 Metric Frameworks of Similar Systems

A matrix of metrics has been proposed and applied to evaluate the success of citizen science projects. As previously discussed, citizen science projects differ from GWAP projects in that participants (often described in this case as volunteers) typically participate willingly through personal interest rather than any added incentive (e.g. entertainment in GWAPs; money in micro-task crowdsourcing). The two main dimensions to the matrix are the projects' "contribution to science", and "public engagement". The contribution to science includes measurements such as, the number of publications resulting from the project (publication rate), their citations (academic impact), the resource savings, how complete tasks are, the level of equality in the distribution of user contribution (distribution of effort) and the number of players that continue after the tutorial (effective training). The "public engagement" dimension includes the number of volunteers (project appeal), the median time interval between a registered volunteers first and last contribution (sustained engagement) and median number of classifications per volunteer (public contribution). The aforementioned framework was used in the evaluation of the Zooniverse project [122].

5.2 GWAP and Free-to-Play Objectives

In this section we will discuss the evaluation similarities between GWAPs and games using the Free-to-Play revenue model (we discuss the design similarities that likely motivated these metrics in Section 10.2). We look at metrics from F2P suitable for evaluating GWAPs.

Revisiting the original metrics proposed by Von Ahn we spoke of in the introduction to this section (Section 5), Von Ahn gave these metrics in three contexts [43]. In addition, two areas were proposed for future work, without metrics:

Throughput a measure of **efficiency**

Average Lifetime Play a measure of **enjoyability**

Expected Contribution a measure of **quality**

future work a measure of **popularity**

future work a measure of **contagion**

This gives us the start of a list of some of things our metrics should be expected to cover. Like Von Ahn, we aim to propose metrics that generalise to a variety of game approaches for the purpose of comparison, rather than evaluation of a single game [43]. More sophisticated metrics could give a more comprehensive game specific picture, but this proposal is concerned with the selection or adaptation of those that generalise to support between-game comparison (e.g. Activation related metrics are deliberately not included or adapted - see Section 5.2.2).

Free-to-Play metrics are quite comprehensive, but one convenient method of organising them is "Pirate Analytics" (AARRRR) [171]. Here, already, we find great similarity to those originally proposed by Von Ahn. This section will describe a selection of the F2P metrics grouped by their interest and related to Von Ahn's originally proposed metrics.

5.2.1 Acquisition

This group of metrics tracks different ways of counting the number of visitors to game, and the costs involved in acquiring them. This is similar to Von Ahn's **popularity**. Examples include:

5.2.1.1 CpA

Cost per Action (or Acquisition) is the cost of the **acquiring** a new customer, or having a customer perform some action (typically a conversion related action – e.g. turning a guest into a registered user, turning a free-tier player into a paying player). It is often specific to an advertising campaign, used to evaluate the cost effectiveness of different promotion methods. For example, *the cost of 100 customer registrations through prizes*, *the cost of 50 customers making a purchase through web based advertising*. Popular variations include *Cost per Install* [172], *Cost per Engagement*

[173], *Cost per Loyal User* (the cost of n customers who launch the app 3 or more times a year) [171] and *Cost per Click* [174], *Cost per Impression* [174].

$$CpA = \frac{\text{Cost of Campaign}}{\text{Users who completed action}} \tag{5.1}$$

5.2.1.2 DAU and MAU

Daily Active Users (DAU) and Monthly Active Users (MAU) are simply the number of unique users that have played the game at least once during that time period. Despite only varying in the time periods they measure, individually, they are used for quite different purposes. DAU is a more user-centric **acquisition** focused measure [171] looking at how many unique users play day by day, whereas MAU is typically used to measure the general growth of the game [174].

These are often used together in a ratio of DAU/MAU, sometimes referred to as the sticky factor [172, 174] or stickiness [175], to provide a measure that is similar to **retention**. This is calculated by taking the average DAU over all the days in the month, and dividing that by the MAU [175].

5.2.2 Activation

This group relates to the flow of users through the game (e.g. movement from tutorial to game, or between tasks). This was not covered by Von Ahn's proposed metrics, possibly as the focus of those metrics was to compare different approaches [43] rather game specific details, and this is quite application/game specific. However, it may be relevant now for more game-like approaches, especially with the increase in tutorial and skill based games (Section 3.4).

5.2.3 Retention

This group measures the games ability to keep players playing, with measures that closely resemble "Average Lifetime Play", such as "session length". This group matches with Von Ahn's **enjoyability**.

5.2.3.1 Cohort Analysis

Cohort Analysis, comes in many variants. Broadly speaking it relates to customers staying or returning over some period of time.

Classic retention is the players that play following a specific time interval after the first play.

Range retention is the players that returned during some interval following their initial play (e.g. first week).

Rolling retention is the players that returned any time after some initial interval that follows their original play session.

There is already some resemblance in the metrics used in Free-to-Play games, to those proposed by Von Ahn [170]. In the Von Ahn games, they typically give a percentage of players that returned on another date. This is a less detailed version of what is described in the Free-to-Play metrics as rolling retention (with an interval of one day).

5.2.3.2 Session Length

Session length measures how long users spend in an app in a single session. This can be positive or negative depending on the context and design goals. A long session may indicate an overly complex process, or an app that users want to spend lots of time in [171].

5.2.3.3 Churn

Churn is effectively the inverse of retention, the number of players that have been lost over some time period [174]. This is closely related to, and suffers similar challenges to, LTV. In a non-subscription setting it is not clear when users have left [175].

5.2.4 Referral

This group tracks players inviting friends - this maps to Von Ahn's **contagion**.

5.2.4.1 K-Factor

K-Factor is a measure of virality [174]

k = invites sent by each customer $*$ conversion percentage of invites

Values above 1 indicate an exponential positive growth. Values below 1 indicate a decline.

For example, if each user invited 10 friends, and 5 of those were converted to contributing users, $k = 10 * .5 = 5$

5.2.5 Revenue

This group tracks the financial gain from the game, whilst GWAPs do not extract value from their players in financial terms, these metrics map directly to work, or as described by Von Ahn, **quality**. This also includes an additional factor frequently given in the literature, **quantity**.

5.2.5.1 Average Revenue Per User (ARPU)

The Average Revenue Per User is the total **revenue** the users have generated divided by the number of active users. This is often used in conjunction with the aforementioned CpA to give a more complete picture of the success of a marketing campaign [171]. In the context of Free-to-Play

games, whilst being able to inexpensively acquire users is good (have a low CPA), those users need to return a profit. A marketing campaign can only be considered truly successful if the average revenue of those users is less than the cost of acquiring them (i.e. the ARPU is less than the CpA). Shown in equation [174]:

$$ARPU = \frac{\text{total revenue}}{\text{number of active users}} \tag{5.2}$$

In some cases the list of users taking into consideration is filtered down to just the paying users rather than the total active users ARPPU (as there are often many non-paying customers) [172, 174, 175].

Another variation on ARPU is ARPDAU. As opposed to narrowing the focus of the users, ARPDAU narrows the focus of the time. ARPDAU considers the total revenue and active users on a single day [172, 174, 175].

5.2.5.2 Lifetime Value (LTV)

Lifetime Value (LTV), sometimes CLV (Customer Lifetime Value) is a customer focused **revenue** orientated metric that predicts the profit a customer will generate in their lifetime. This can be as simple as the current profit generated by the customer minus the cost of their acquisition [172]. However, generally, probabilistic models attempt to forecast the profit from the customer to get a long term projection of a customers value. Such models fit, broadly, into two groups. Historical/retrospective CLV looks only at past transactions. Predictive CLV models future actions. Selecting a predictive model is contextual on the nature of the product and sale (e.g. whether their are repeated contractual subscriptions). Free-to-Play games are not subscription based (e.g. telephone service), nor can purchases be easily predicted (prescription renewal). As a result, there is no explicit notice of cancellation provided by the player/customer by which their end of life can be determined, nor is there any obvious point of purchase. Additionally, purchases are continuous/unobserved rather than discrete [176]. The player can make a purchase at any time [177]. This type of customer activity fits the criteria for the Pareto/NBD model [176].

5.3 Our Proposed Amendments

In addition to the selection of F2P metrics discussed above we recommend some additions/amendments. These are adjusted to feature the dimension of items and judgements.

5.3.1 Cost per Action (CpA)

Examples of other actions for analysis of GWAPs include:

CpJ Cost per Judgement - the average cost to provide a useful judgement

CpI Cost per Item - the cost to acquire a completely annotated items. For an item to be completely annotated, multiple annotations are typically required; some form of aggregation is then applied. A well designed system may require fewer annotations to achieve completion if for example, task to user assignment is better, players are trained better, or the task itself is presented more efficiently.

5.3.2 Lifetime Judgements (LTJ)

This metric would fall under the **quantity** metric interest given by Von Ahn. This is the total judgements (count of individual annotations) made in the game per player.

$$LTJ = \frac{\text{number of judgements}}{\text{number of players}} \tag{5.3}$$

5.4 Concluding Remarks

Von Ahn's discussion regarding the design of GWAPs stressed the importance of evaluation metrics, comparing them to an algorithm [170]. He identified some shortcomings in the metrics originally proposed. However, despite the increasing complexity of GWAPs over time, not only did the shortcomings remain, the original metrics went largely unreported in future GWAP approaches.

Inspired by the metrics of F2P and their commonalities with GWAP interests, this Chapter has discussed a new set of metrics to give a more comprehensive overview when evaluating GWAPs.

This will be a key building block for the rest of this work, as we use these metrics to evaluate our game design approaches.

CHAPTER 6

AN EXPLORATORY STUDY OF GAMIFICATION FOR NLP

In this Chapter we primarily focus on token labelling (Part V) and text segmentation (Part IV). However, we also carried out some preliminary experiments with game designs targeting tokenization. We explore one of those designs in this Chapter.

6.1 The Tokenization Game

"The Logging Game" was one of many prototypes designed for tokenization and segmentation with the goal of creating a more game-like GWAP experience. A pilot study was run on a small group to test the effectiveness of this early design and its basic usability. This did not involve accuracy. During this early development phase a user survey was preferred to implementing analytics. Many game-design and evaluation lessons were learned from this prototype and carried forward to the development of later games. However, they could equally be applied here to enhance The Logging Game.

6.1.1 Game Design

The Logging Game is a single player, web-based, casual game in which players can mark segments such as token boundaries, sentence boundaries and named entities. A screenshot of the game is shown in Figure 6.1.

The players are shown a subsection of text on a log, and asked to create selections that marking the area of interest with various tools in a way analogous to a normal text selection mouse cursor. The selection is encompassed by a translucent tube. This is colour coded to offer feedback about the selection (e.g. green (shown in Figure 6.1) or red depending on whether the selection was correct or incorrect respectively).

Figure 6.1: The Logging Game

The functions and operation of the tools were as follows:

Axe Chopping at the beginning and end of the selection created a selection.

Saw Sawing at the beginning or end of an existing selection, and sawing in a different location, moved the boundaries of a selection.

Hammer Hammering an existing selection on or between the boundaries of that selection, removed the selection.

The tool wheel shown at the bottom middle of Figure 6.1 allowed players to switch tools. The bottom middle button was the action button, and the two smaller buttons adjacent to the action button switched to that tool. A tooltip shown beneath the action button reminded the player of the purpose of the currently selected tool.

The arrows in the bottom left and right allowed the player to move the log left and right respectively. The smaller double arrows adjacent to those large arrows allowed the player to move the log in that direction, to the nearest white space character to save repeated pressing. Additionally, the player could touch and drag the log to their desired position.

The players score was given on a board to the right, with 1 being an optimal solution to the problem. When the player was satisfied with their solution, they could tick the *tick* button in the top right to proceed to the next challenge.

So that the player could easily see the text in context, the full sentence was displayed in a translucent box overlay at the top middle of the screen.

6.1.2 Experimental Setup

As the experiment sought to test the user experience rather than perform any annotation, the set of texts available for classification were fixed so that each user was presented with the same text.

10 participants were recruited in a deliberately informal setting. This is a small number of participants compared with a typical scientific experiment, but is considered adequate for usability testing purposes, where initial problems in pilot applications are typically discovered swiftly and additional users offer diminishing returns [178].

Players were first invited to complete an in game tutorial, before playing the game itself, to ensure they had similar knowledge upon which to base their judgements of the game.

After playing, players were given a two part survey (shown in table 6.1). The first part of the survey was bespoke, and related directly to the game itself. The second part of the survey was more generic, using the SUS (System Usability Scale) [179] survey questions. Players also had the opportunity to leave written feedback, and provide verbal feedback.

6.1.3 Results

6.1.3.1 Survey Results

Table 6.1 shows the results for the survey.

6.1.3.2 Written Results

Three written comments were given. I will refer to these individuals as persons A, B and C.

Person A wrote:

> Not clear with the controls, hard to determine words

although this person gave “Agree" or "Strongly Agree" for all control related questions in the survey

Person B wrote:

> Order of tools unclear - apparently it’s a wheel?

- gave "Neither Agree or Disagree" for actioning and navigating tools

Person C wrote:

> The axe was cool

6.1.3.3 Verbal Feedback and Observations

Verbal feedback was given questioning whether it would be faster to perform the task using a cursor similar to those present most GUI’s for text entry.

Questions	Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
I understood the axe created new selections.	0	0	1	4	5
I was able to create new selections with the axe.	0	0	0	2	8
I understood the saw edited selections.	0	2	3	5	0
I was able to edit selections with the saw.	0	1	3	4	2
I understood the hammer deleted selections.	1	1	1	5	2
I was able to delete selections with the hammer.	1	1	1	2	5
I found it easy to navigate the sentence.	1	1	1	3	4
I understood the silver and red stand beneath the log was the current position.	1	1	3	3	2
I understood the red tube around the log indicated the marking was incorrect.	2	0	2	2	4
I understood the green tube around the log indicated the marking was correct.	1	1	0	3	5
I found it easy to navigate and action the tools.	0	1	1	5	3
I think that I would like to use this system frequently.	3	2	4	1	0
I found the system unnecessarily complex.	1	5	2	1	1
I thought the system was easy to use.	2	0	2	5	1
I think that I would need the support of a technical person to be able to use this system.	4	5	0	0	1
I found the various functions in this system were well integrated.	0	1	4	3	2
I thought there was too much inconsistency in this system.	2	5	1	1	1
I would imagine that most people would learn to use this system very quickly.	2	3	2	3	0
I found the system very cumbersome to use.	1	0	5	1	3
I felt very confident using the system.	1	3	3	2	1
I needed to learn a lot of things before I could get going with this system.	2	4	0	3	1

Table 6.1: Survey Results

Multiple people mentioned how they enjoyed chopping the log lots. But didn't appear to take an interest in the activity itself.

One player suggested stepping back further from the problem and thinking about the representation. The example they gave was how protein folding in FoldIt could be broken down to a geometry problem, and 3D games are essentially geometry, providing a good link to represent the problem as a game.

Another player made a similar comment about the chopping metaphor and suggested I instead think about a game in which I show all possible segmentations, and get the user to pick one so that the process of contributing was not obstructed by the game mechanics.

This was a reoccurring theme, and it appeared the general consensus was that the upon which to base their judgements of the game. Chopping the log as a metaphor for segmentation was far too literal and got in the way of the task, making selections, edits and deletions take longer.

6.1.4 Discussion and Conclusions

Regrettably, there is a clear inconsistency between the survey, written/verbal feedback received, and observed play.

There are few obvious trends in the questionnaire:

- The creation of selections with the axe scored positively. This does match with the observation of participants. Players quickly grasped the function of the axe.
- The log turning green for a correct selection seemed a very effective communication
- Navigating tools scored positively in the survey, but when observed players struggled with this functionality. This also contradicts both the written and verbal feedback given.
- Players thought the system was inconsistent. User observation and survey results would assume this was in relation to the tools rather than navigating the log
- On the whole users thought it was easy to use and wouldn't need technical assistance or to learn a lot of things. This contradicted the observation of participant play
- The first question in the SUS part of the survey revealed that participants wouldn't really want to use the system frequently

It appears this experiment split the users into two groups. Firstly, there were those that were interested in the problem, but not the game, and seemed to find the game a barrier to effectively contributing towards the problem. Second, there were those that were interested in the game (e.g. like chopping the wood repeatedly with the axe), but had little interest in the annotation task. This ultimately meant that players were either quickly frustrated, or briefly entertained by the novelty of the game, then bored. Neither outcome would be likely to result in a positive contribution to the task. There is clearly a toy facet (chopping wood) and a tool facet (annotating noun phrases) to the application [180]. It would appear that it is on that line that the application has polarised the players. Users struggled with the interface which attempted to add fun through deliberately complicating the user interaction (a very pattern common to games), sometimes referred to as their ludic efficiency [181].

In conclusion, The Logging Game is a promising prototype for tokenization and text segmentation. In it's attempt to be a more game-like GWAP it suffers from many of the challenges such games do. Despite this, we believe these can be remedied and that The Logging Game, has great potential. We discuss these challenges and others, and how we addressed them in *WordClicker* in Part V.

Part IV

Text Segmentation

The chapters in this part discuss work towards the development of GWAP for text segmentation. In the first chapter we introduce the GWAP, *TileAttack* and discuss our method of aggregating mentions. In the second chapter we discuss our first large scale experiment. In the third chapter we discuss progressing workers through increasingly complex tasks in line with their competence to provide a rewarding experience, train players and optimize resource utilization.

CHAPTER 7

TILEATTACK

This chapter introduces the game *TileAttack*, a game I created as a base for experiments. *TileAttack* gathers **mentions** in text used to support coreference resolution (Section 2.1.3). The game supports any text segmentation task, whether markables are nested or non-nested, aligned or not aligned, and is therefore applicable at least in principle to a variety of text annotation tasks, including e.g., Named Entity Resolution (NER) and tagging, but this goes beyond the scope of this thesis.

TileAttack is a web-based (available at https://tileattack.com) 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 7.1). In the game, players perform a text segmentation task which involves marking spans of tokens represented by tiles. 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*, *TileAttack* uses the "output-agreement" format, in which two players or agents are paired, and must produce the same output, for a given input [43].

The game is parameterised so that the effect of different setups can be studied. Aside from being able to share or like the game on Facebook, there is further integration that allows players to log in to the game via Facebook. Before being taken to the game, players are shown a short introduction that includes an explanation of the items they will be marking, the interface, the controls and properties of the game unique to the specific experiment taking place. For example, when there is a timer, they are told how long they will have.

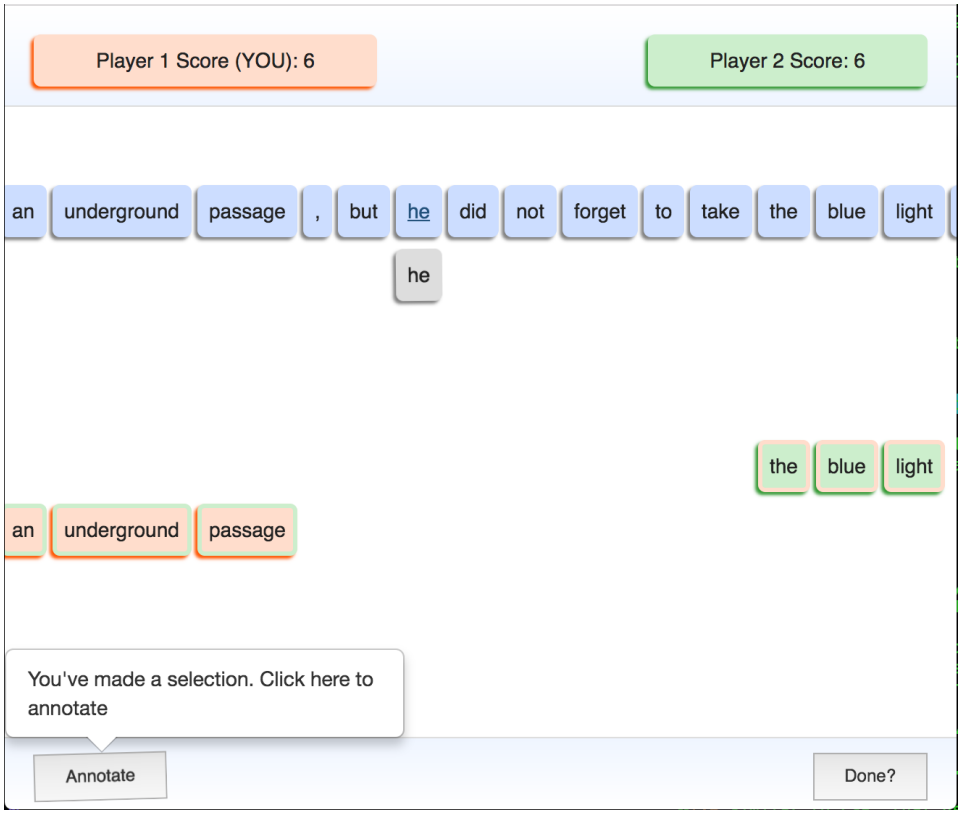

Figure 7.1: TileAttack game

7.1 Interface

The game deliberately omits any specific design themes that may appeal to a subset of the players in order to focus only on the game mechanics being tested. This clean Scrabble inspired template for the game provides a canvas for future experiments relating to individual user personalisation or theming the game in line with current trends (e.g. spaceships, zombies, football).

A mobile-first responsive interface has been used with quick methods of interaction. Selections can be made with minimal taps over large tiles to make it easy to tap the tiles on a mobile device. The sentence can be scrolled on the phone by swiping to the left or right. When displayed on a small portrait screen the scores resize and are stacked vertically.

Great care has been taken in the selection and application of visual game design concepts to effectively communicate operation of the game through the interface using multiple channels including colour, object movement and text. For example, items that are in an interactive state display a subtle animated wobbling effect. This can be seen when the player makes a selection in the preview selection bar and buttons, when appropriate. Consequences of positive actions are shown using a horizontally moving glinting effect (Figure 7.3). This can be seen when the players match moves. A simple colour scheme provides context to the user as to which aspects of

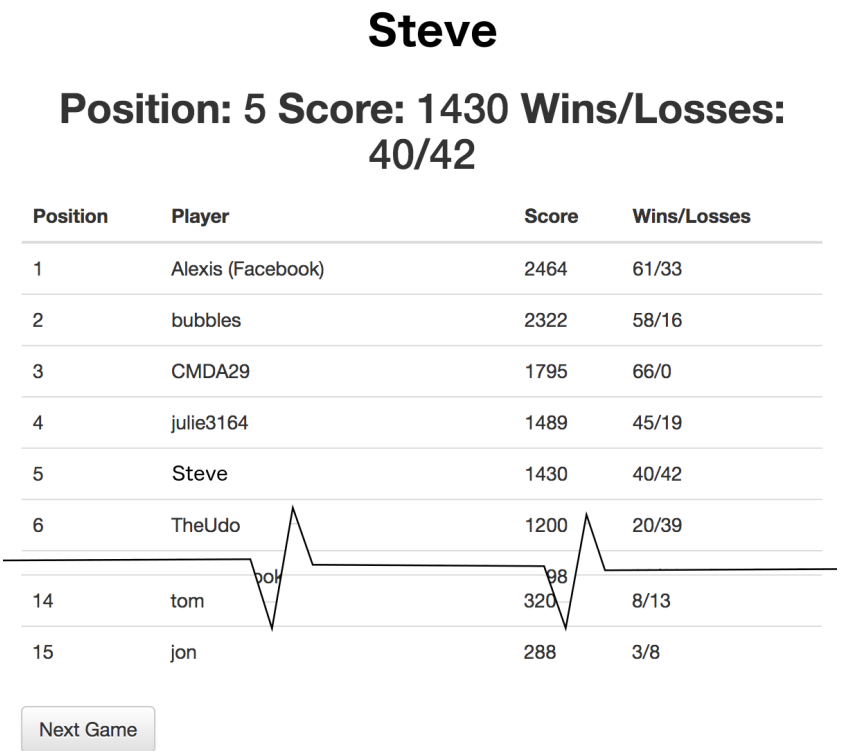

Figure 7.2: Leaderboard (midsection cut for brevity)

the game relate to them, and which relate to their opponent.

7.2 Tutorial

Following the documentation, but before the game, players are shown a two round tutorial (mandatory for crowdsourced workers). The tutorial shares the same user interface as the game. In the tutorial the player marks all of the available phrases to proceed to the next round. Correctly identified mentions show with a glinting effect.

They are informed of what entities are present in the sentence and how many mentions there are. This is shown with both written and pictographic cues. As they correctly identify mentions the counter is reduced until all mentions of that entity are discovered, and the text is then changed to grey. 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. They receive immediate and direct feedback to inform them of their progress, and a summary of how many mistakes they made at the end of the round.

First first round uses only one entity with two references. The mentions are very simple example, a *definite article; noun* phrase and single *it*. The sentence is: *{The music} was so loud that {it} couldn't be enjoyed.*. The player is shown a picture of a speaker emitting sound to illustrate

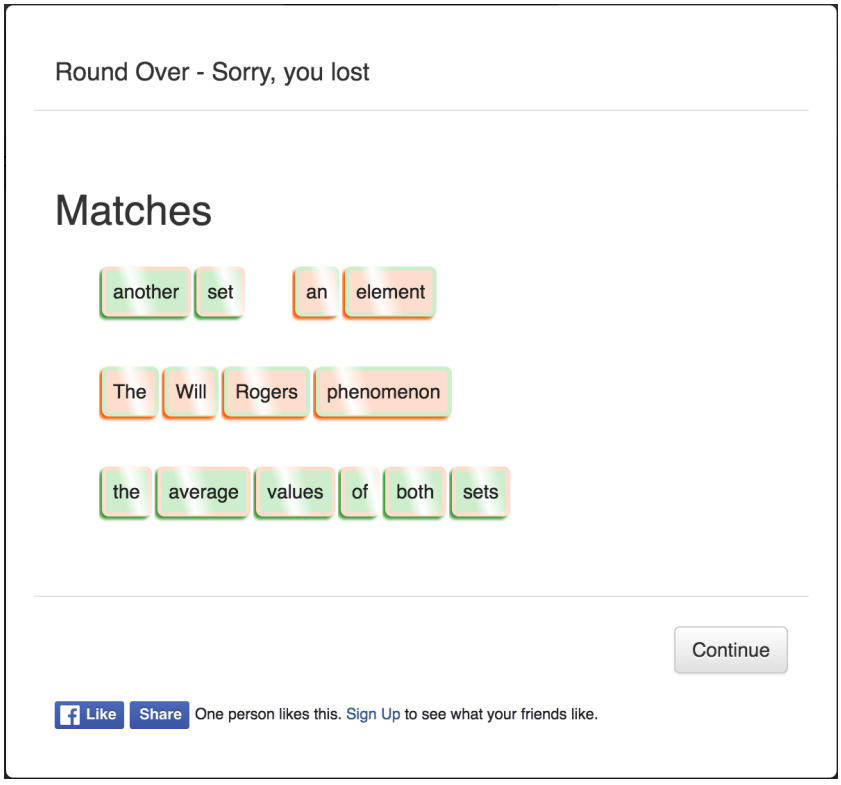

Figure 7.3: End of round summary

the entity they are looking for (shown in Figure 7.4).

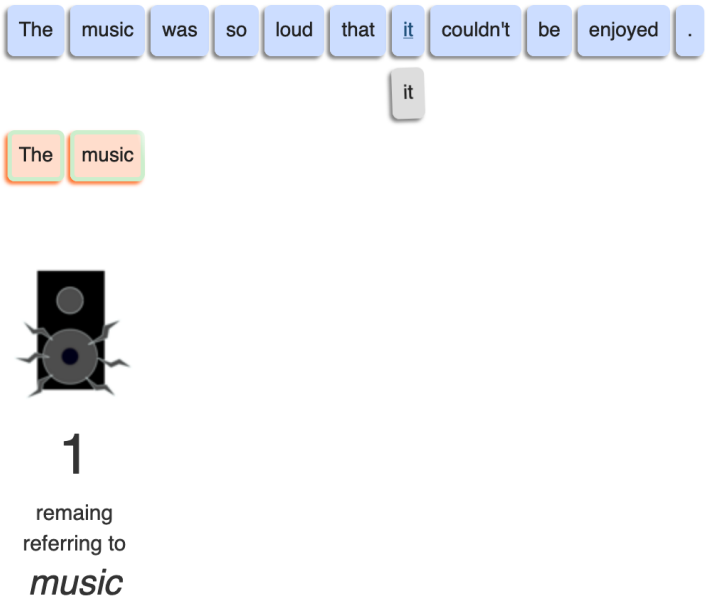

Figure 7.4: First tutorial round from *TileAttack*

The second sentence contains two entities and three mentions. The sentence is: *{A wolf} had been gourging on {an animal {he} had killed}.*. This is illustrated with pictures of a wolf and a deer given the appearance that it has been bitten and is deceased (shown in Figure 7.5). This is more complex than the first sentence, featuring a pronoun, noun-phrase with a post-modifier and a pronoun.

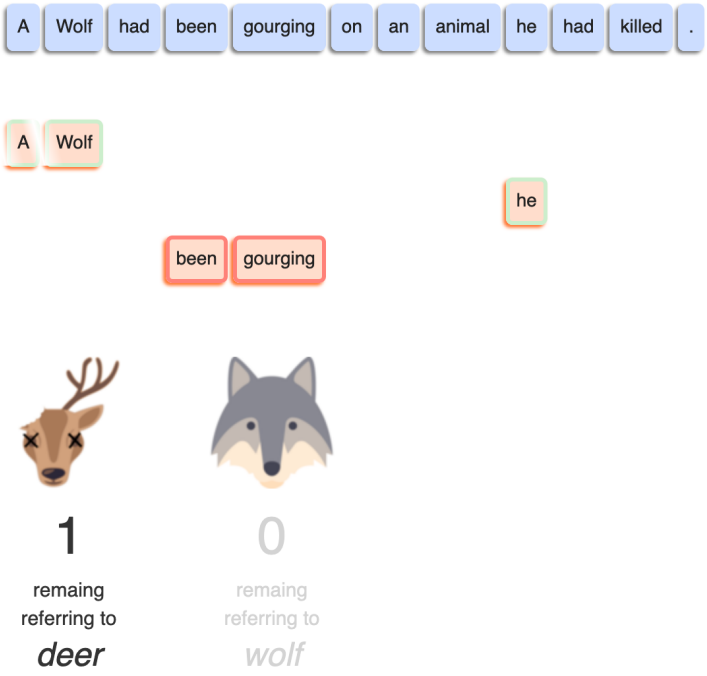

Figure 7.5: Second tutorial round from *TileAttack*

7.3 Gameplay

In each round, the player is shown a single sentence to annotate. The players can chose 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 opponents 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 (Figure 7.3). 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 (Figure 7.2), where they are shown their current position, score, wins, losses and the current top fifteen players. From this page they may click the *Next Game* button, to start another round.

7.4 Opponents

The game uses one of three artificial agents to fulfil the role of the opponent player. They are selected in the following order of priority, descending to the next unless the condition is met:

Silver AI Replays the aggregated result of all player games so far - if there are a sufficient number of games available to aggregate for that item

Replay AI Replays a recorded previous game - if a previous game is available for that item

Pipeline AI Plays the moves from an automated pipeline. The pipeline used varies with the experiment

Opponents form an essential part of TileAttack. Beyond the notion of awarding points based on agreement, they allow various means of communicating potential labels with players about which the system is currently unsure, and continuously improving upon current knowledge. As is often the case with GWAPs, the player does not literally play against another player, but rather a replay of a previous player's actions. This addresses the challenge of ensuring multiple players are available at once.

7.5 Crowdsourcing

TileAttack is designed to be a GWAP, so using crowdsourcing to recruit players may seem counter-intuitive. However, collecting judgements from organic players tends to be slower than using a crowdsourcing service. Given that not all research questions are concerned with game-design, recruitment or player engagement, but rather producing as accurate as possible annotation from non-experts, some experiments collect results through paid crowdsourcing.

To achieve this, TileAttack is integrated into the Amazon Mechanical Turk 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, workers are awarded an MTurk qualification to track their performance.

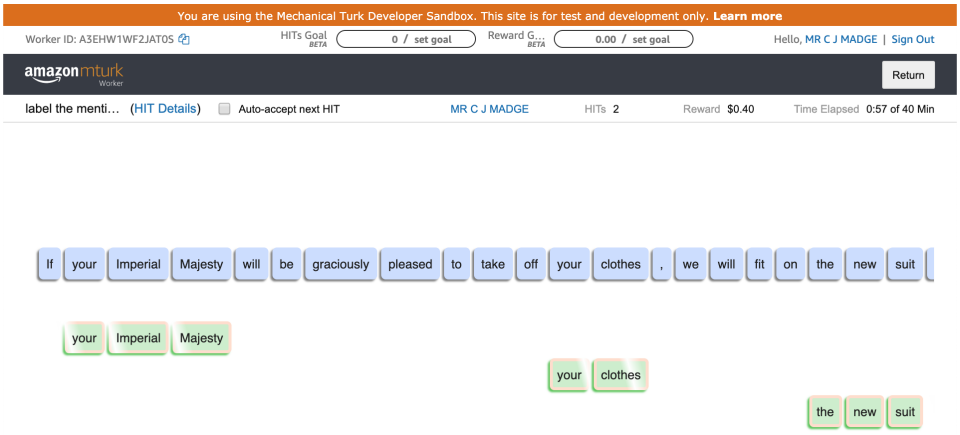

Figure 7.6: TileAttack integrated into MTurk (worker sandbox)

7.6 Aggregating Mentions

The boundaries labelled by non-experts can be expected to be quite noisy compared to expert annotations; but we can also expect the quality of the **aggregated** judgements to be comparable to that obtained with experts, provided sufficient non-experts are consulted [40]. We are not aware however of any previous proposal to aggregate such annotations when they are nested. In this Section we introduce the two methods we used: a baseline on one based on taking the most popular judgement among the annotators (majority voting); and a probabilistic approach. Both these methods require a way for clustering together the mentions to be compared; we propose one such method in the first Section.

7.6.1 Head-based mention boundary clustering

To apply aggregation, it is necessary to determine which judgements (boundary pairs) are competing. We do this by clustering all annotations sharing the same **nominal head**. The heads from the sentence are extracted using the dependency parse from the DEP pipeline. Typical

nominal heads include nouns, proper nouns, pronouns and expletives, but other types occur as well.

The heuristics to find them are based off the part of speech tag and dependency parse. The full set of rules is shown in Table 7.1.

POS Tag	Dependency Rule
PRP*	$\notin$ { punct, cc, advmod }
NN*	$\notin$ { advmod, partmod, prep, cc, nn, discourse, punct, amod, num, det, cop, aux }
CD	$\in$ { nsubj, nsubjpass, pobj, dobj, poss, ROOT, appos, acomp }
DT	$\in$ { nsubj, nsubjpass, pobj, dobj, conj, npadvmod }
JJ	$\in$ { nsubj, nsubjpass, dobj, appos }
EX	$\in$ { expl }
VB*	$\in$ { nsubj, nsubjpass, dobj, pobj }
$	$\in$ { dep, dobj, pobj }

Table 7.1: Head finding rules
POS Tag are part of speech tags [182]; Dependency [183] rules; * indicates wildcard match

For example, in the case of an elliptical construction such as *I need a folder for my notes, as I have collected too many* $_{notes}$ *to carry*, the adjective *many* would be identified as a head in the absence of the noun *notes*. This is the case for which the adjective is the direct object of the phrase ("JJ" and "dobj" in Table 7.1).

The dependency tree is aligned with the candidate mention as follows.

Given a player-generated candidate mention, we find first of all subtrees of the dependency tree that completely cover all the tokens in the candidate mention. The highest leftmost head of those subtrees is then considered as the head. If no nominal head is present in those subtrees, the candidate mention is not considered for aggregation.

For example, consider the sentence *John's car is red*. Suppose the players proposed the candidate mentions *John's car*, *John*, and the (incorrect) mention *John's car is*. Further suppose that the (automatically computed) dependency tree is as in Figure 7.7:
Then *John's car* can be aligned with the subtree whose head is *car*; *John's* can be aligned with a subtree with head *John*. Both of these heads are nominal, so the two candidate mentions are considered for clustering. *John's car is* would be aligned with the two subtrees with the roots *car* and *is*, shown in Figure 7.7 by the red box. The highest leftmost head and therefore the head that would be used is *car*. Relaxing the alignment criteria this way is important to allow the pipeline to guide the clustering while not constraining newly proposed boundaries to the pipeline's overall interpretation (which may be incorrect).

If no viable heads are discovered in the selection then it is not considered. For example, if the player chose *is red*, this would be omitted.

We next take an example that looks at prepositional attachment, a common example of where pipelines often produce incorrect interpretations or there is an intrinsic linguistic ambiguity.

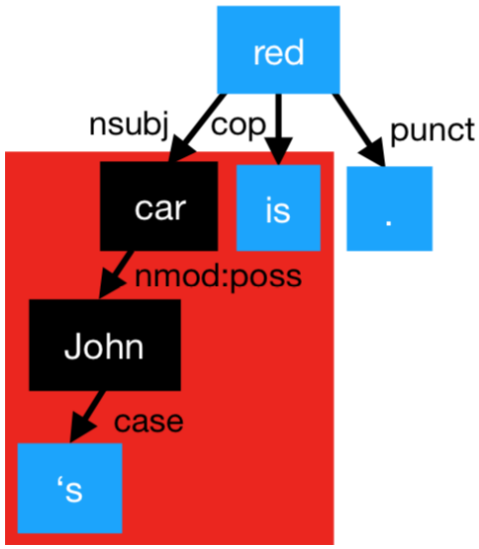

Figure 7.7: Finding a head for a proposed boundary

Figures 7.8 and 7.9 show a high prepositional attachment and low prepositional attachment dependency parse, respectively. In the high prepositional attachment version of the sentence, the "knife" is attached to the verb "killed", interpreted as the instrument of the action. In the low prepositional dependency parse, "knife" is attached to "man", where it is a possession.

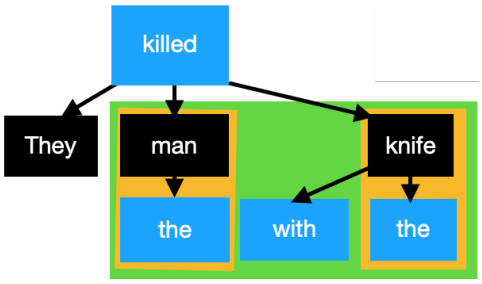

Figure 7.8: Finding a head with ambiguous prepositional attachment: High attachment

Here we can see regardless of the dependency parse (high or low), the annotator can volunteer either interpretation and the appropriate head is still discovered. In this scenario, we would be able to vote between alternate prepositional attachment interpretations.

7.6.2 Majority Voting

Majority Voting was used as a baseline aggregation method. Following clustering, majority voting is applied to each cluster, choosing the boundary that has the highest number of votes among all those sharing the same nominal head. Ties are broken randomly; the process is rerun five times.

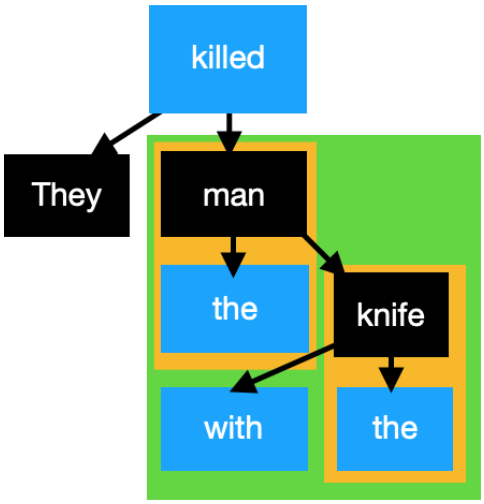

Figure 7.9: Finding a head with ambiguous prepositional attachment: Low attachment

7.6.3 A Probabilistic Approach

The majority vote baseline implicitly assumes equal expertise among annotators, an assumption shown to be false in practice [184]. A probabilistic model of annotation, on the other hand, can capture annotators different levels of ability [185]. This Section, the work of Silviu Paun, describes an application of the model proposed by [108] to the boundary detection task.

Bayesian models of annotation [35] are a mechanism to infer from the annotation data parameters characterising the annotation process such as the annotator accuracy, bias and class prevalence. Dawid&Skene, to our knowledge, is the first model based approach to annotation. This is shown in Figure 7.10, where π is the prevalence of a class, $\beta_{j,k}$ is the annotator ability for a class, $y_{i,k}$ is the observed class and c_i is the inferred class.

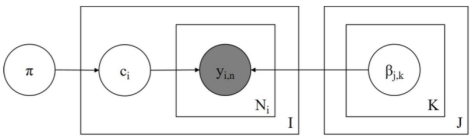

Figure 7.10: Plate diagram of the Dawid&Skene model [35]

Each cluster contains a number of candidate boundaries supplied by the players. The goal is to identify the correct boundary for each cluster. A multi-class version of the Dawid&Skene model cannot be applied since the class space (the boundaries) is not consistent (i.e., the same set) across the clusters. However, a binary version of the model can be applied after some careful data pre-processing. Concretely, for each boundary we obtain a series of binary decisions as a result of a "one vs. the others" encoding performed at cluster-level. For example, given a cluster whose annotations are the boundaries "a, b, a, a", we have for the "a" boundary a collection of "1,

0, 1, 1" decisions, while for the "b" boundary we have "0, 1, 0, 0". A Bayesian version of the binary Dawid&Skene model is then trained on these boundary decisions. The model infers for each boundary a decision indicator which can be interpreted as whether the boundary is correct or not. After some simple post-processing, we assign for each cluster the boundary whose posterior indicator has the most mass associated with the positive outcome.

CHAPTER 8

TileAttack for Mention Detection: An Evaluation

This chapter brings together several of the ideas and technologies discussed with the goal of demonstrating a method of crowdsourcing high quality candidate mentions labels (Section (2.1.3)) from non-experts using TileAttack (Chapter 7).

There are multiple components to this system. Firstly, rather than have annotators start completely from scratch, it makes sense to improve upon an automated pipeline. TileAttack allows a pipeline to be used as an opponent, providing a method to convey information to the non-expert human player that may be informative, but cannot assumed to be correct.

There is some disparity between the goals of existing mention detection systems leading to some to focus on having a high recall and others a high F_1 (Section 2.1.3). It is not clear which behaviour would be most desirable when using such pipelines as opponents, in this setting. As such, this chapter will not only propose two new pipelines (Section 8.1) and test them against the existing state-of-the-art, but go on to use them in two different configurations.

Having used TileAttack to crowdsource non-experts to correct existing mentions produced by automated pipelines, the remaining step is extract the wisdom of the crowd to produce a final high quality result. Whilst there are a number of aggregation methods available for text segmentation tasks, they do not support nesting. To address this, a novel method of aggregating nested mentions is proposed used (Section 7.6).

8.1 Two automated mention detectors

The first ingredient of our proposal are two strong mention detection pipelines to serve both as baselines and as AI opponents for *TileAttack*. These were developed by **Juntao Yu**. The first pipeline first parses the input sentences using a dependency parser and then extracts mentions from the dependency parse; heuristic patterns; we call this the DEP pipeline. The second pipeline

is a modified version of the neural named entity recognition system proposed in [186]; we call it NN pipeline. Both pipelines are trained on the Penn Treebank (PTB).

8.1.1 DEP pipeline

Juntao Yu's DEP pipeline first parses input sentences using a dependency parser, then applies a rule based mention extractor that extracts mentions from dependency trees using heuristic patterns. In preliminary experiments we compared two frequently used dependency parsers: the neural network based parser by [187], and the Mate parser [188]). The [187] parser is the current state-of-the-art dependency parser, but it is slower than the less accurate Mate parser. In our preliminary experiment we found that the small difference in parser accuracy affected the performance of our mention detector only slightly. We decided therefore to use the Mate parser to maximise efficiency.

The second part of the pipeline is a rule based mention extractor. The extractor follows a three steps approach. It first extracts mention heads using heuristic patterns based on part-of-speech tags and dependency relations. The patterns are automatically extracted from the Phrase Detectives 1.0 [70] corpus, which was annotated by experts and follows the same mention annotation scheme as our game. We extract all the part-of-speech tags and dependency relations pairs of the mentions' head in the corpus, and use the most frequent patterns. The second step of the extractor is to find the maximum span related to a given mention head; for this we use the left/right-most direct or indirect children of the mention head as the start/end of the mention. The last step checks if any of the mentions created by the step two overlaps with each other. When overlapping mentions are found they are replaced with the union of those mentions. Please note the nested mentions are not counted as overlap mentions, hence will not be processed.

8.1.2 NN Pipeline(s)

Juntao Yu's second pipeline does not use a dependency parser; instead, it uses part of the neural named entity recognition (NER) system proposed in [186]. The [186] system takes a sentence as the input and outputs a sequence of IOB style NER labels. This is an early version of the system that was later published [189].

The system uses a bidirectional LSTM to encode sentences and applies a sequential conditional random layer (CRF) over the output of the LSTM. The CRF is effective when handling sequence labelling tasks such as NER, but it is not suitable for predicting mentions, as mentions can be nested. In our detector we represent mentions with the representation of tokens at the start and end positions of the mention. For each token we create a maximum l candidate mentions. Let s, e be the start and end indices of the mention, and x_i the LSTM outputs on the i_{th} token. The mention is represented by $[x_s, x_e]$. In addition, we add a mention width feature embedding (ϕ) and apply a self-attention over the tokens inside a mention ($[x_s \ldots x_e]$) to create a weighted mention representation w_{se}. After creating the mention representation $[x_s, x_e, w_{se}, \phi]$, we use a

Configuration	P	R	F_1
OntoNotes [68]			
Stanford	40.38	89.46	55.65
DEP	36.60	83.79	50.95
NN High F1	73.53	74.01	73.77
NN High Recall	51.53	87.53	64.87
News			
Stanford	71.55	67.28	69.35
DEP	86.03	72.33	78.59
NN High F1	79.33	86.16	82.60
NN High Recall	71.65	91.29	80.29
Other Domains			
Stanford	77.52	80.11	78.79
DEP	84.72	81.78	83.22
NN High F1	79.92	87.48	83.53
NN High Recall	73.35	93.04	82.03

Table 8.1: Mention detectors comparison.

feed-forward neural network (FFNN) with a sigmoid activation function on the output layer to assign each candidate mention a mention score. During training we minimise the sigmoid cross entropy loss. During prediction, mentions with a score above the threshold (t) are returned. The threshold can be adjusted to create models for different purposes. In particular, in this work we experimented with two models: one optimized for high recall, the other for high F1. We use the same network parameters as [186] except the two parameters introduced by our system. We set maximum mention width to 30 i.e. $l = 30$, and set $t = 0.5/0.95$ for our high-recall and high-F1 versions respectively.

To achieve high-recall rather than high F1, β is adjusted to 2 in the targeted F-score to give an F-score that prefers recall over the balanced F_1 score.

$$F_\beta = \frac{(1+\beta^2)\cdot \text{true positive}}{(1+\beta^2)\cdot \text{true positive} \cdot \beta^2 \cdot \text{false negative} + \text{false positive}} \tag{8.1}$$

8.1.3 Results

We use as a baseline the Stanford mention detector included in the Stanford CORE-NLP pipeline [74]–arguably, the most widely used mention detector for coreference with the CoNLL dataset [68]. The pipelines have been designed based on different corpora, but the goal of this work is to demonstrate the use of a pipeline to support annotation out of original domain. In the interest of completeness and understanding how the pipelines perform in this respect, we compare against three datasets. Table 8.1 shows the comparison between our pipelines and Stanford's in these three datasets. The first dataset is OntoNotes (CoNLL 2012 shared task) [68]. The second, *News*, is the Penn Treebank (PTB) [190]. The third and final is a dataset of our own creation that covers various genres (described in further detail in Section 8.2.1).

Both of our pipelines constantly outperform the Stanford pipeline by a large margin.

8.2 Experimental Methodology

In order to evaluate our approach, we tested the mention boundaries obtained using the two proposed pipelines and by aggregating the judgements collected using *TileAttack* in several different ways over datasets in different genres.

As said above, our approach to human checking of system-produced mentions is to treat automatic mention detectors as artificial agents that human players 'play against'. But we also pointed out that the mention detectors used for coreference resolution *systems* are optimised to achieve extremely high recall–the assumption being that the extra mentions will be filtered during coreference resolution proper– and that this optimisation may not be optimal when using an automatic mention detector for *annotation*–in our case, treating it as an agent from which the other players will derive feedback. In this context, a mention detector optimised for high overall F may be preferable, as it may provide better feedback to the human players. We tried therefore two versions of the NN pipeline in this experiment: one optimized for high recall, and one for high F_1. The two configurations are shown in Figure 8.1.[1]

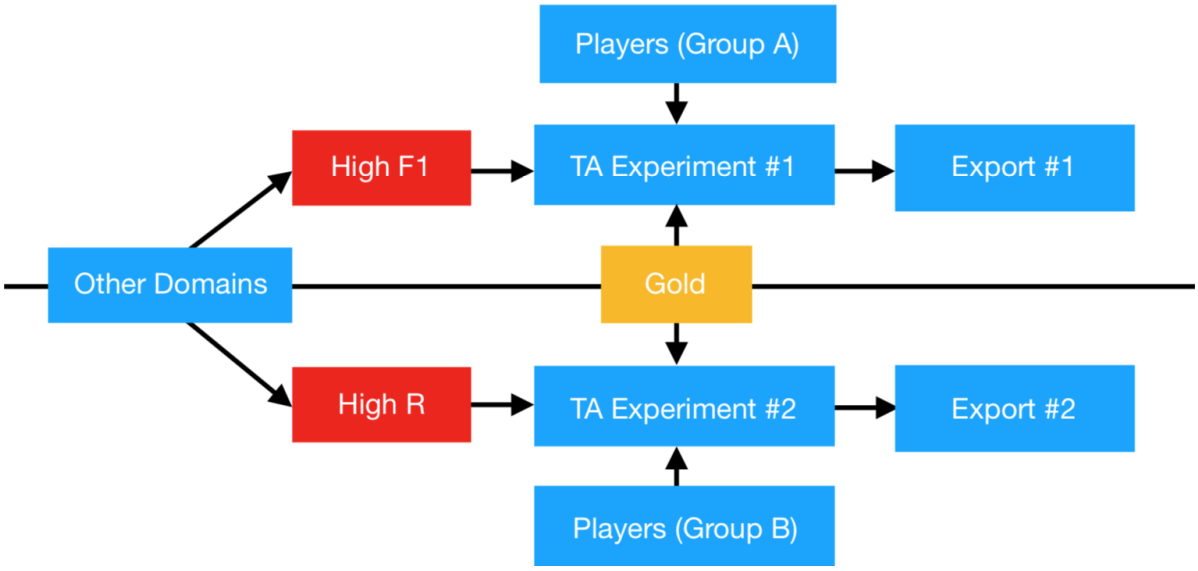

Figure 8.1: Experiment Setup

The regular players of *TileAttack* are typically experts in language or language puzzles, and many of them are linguists or computational linguists. As a result, the quality of the mentions they produce tends to be very high, as shown in Table 8.2, which reports the aggregated results

[1]The DEP pipeline is not optimised either way.

	High Recall Pipeline		
	precision	recall	F_1
MV(users)	90.284	87.536	88.889
P(users)	91.928	89.13	90.508

Table 8.2: Regular players accuracy on "Other domains"
MV: majority voting; P: probabilistic

of these players on the sentences from the 'Other Domains' dataset when playing against the 'High recall' pipeline. Our players obtain an aggregated F of 90.5, which is very high.

However, collecting judgements from the players tends to be slower than using a crowdsourcing service. So given that in this work we were not concerned with comparing the effectiveness of crowdsourcing platforms and GWAPs, we collected the headline results for this experiment using judgements from participants recruited through Amazon Mechanical Turk (MTurk) using the *TileAttack* MTurk integration. This was done for purely practical reasons–namely, ensuring we would collect sufficient data in a reasonably short time.

MTurk qualifications are used to pre-filter suitable workers in two forms. Workers periodically annotate a sentence for which there is a gold standard. Their average performance over these gold standard rounds is assigned to a qualification. This qualification is a requirement HITs to be visible. Should this value drop below the threshold, future HITs will not be visible to the worker. This helps eliminate spammers. Secondly, each experiment treatment is assigned a unique identifier. A qualification is awarded to the worker with that unique identifier to ensure that the player has not previously participated in an experiment with an incompatible treatment that would void their contribution in this experiment setup. When experiments are run they use qualifications to explicitly exclude workers that have participated in experiments that are not compatible.

The participants using *TileAttack* are shown the game documentation, then taken to the mandatory tutorial. Having completed the two tutorial rounds they are then asked to annotate three sentences. At the end of each round, the participant is given feedback in the form of a comparison of their moves, to the acting agent (discussed in the description of *TileAttack*). Having completed the tutorial and three sentences, the participants are then remunerated 0.40 USD for their participation (effectively 0.08 USD/sentence). This value was based on the observation that a single game of TileAttack typically takes less than 30 seconds (which equates to approximately 9.60 USD/hour), exceeding US minimum wage [191]. When accepting future HITs participants are not required to repeat the tutorial but are, instead, asked to annotate five sentences.

8.2.1 Datasets

Two datasets were used for evaluation. Most coreference datasets consist primarily of news text; for this reason, our first dataset, referred to below as "News", consists of 102 sentences from five

randomly selected documents from the Wall Street Journal section of the Penn Treebank [192], annotated with coreference as part of the ARRAU corpus [69].

The second dataset, referred to below as "Other Domains", is 180 sentences from a collection of our own creation consisting of documents covering different genres, ranging from simple language learning texts and student reports, to Wikipedia pages and fiction from Project Gutenberg. These sentences were hand labelled by three expert linguistic researchers. We could not find a suitable measure of inter-annotator agreement for this task, so annotators met to discuss disagreements and find a consensus.

We aimed to have at least 100 sentences in each corpora. The exact number of sentences was not predetermined, but rather determined by the amount of time the workers took to complete them. The experiment was run for a fixed period and a each sentence had to be completed in game at least 8 times to be included.

8.2.2 Results

8.2.2.1 News dataset

102 sentences were annotated by 131 participants. Each sentence was annotated at least 8 times (maximum of 11). For evaluation purposes, a boundary is considered to be correct iff the start and end match *exactly*.

The results in Table 8.3 compare the results obtained using the four pipelines or application the two different aggregation approaches on the user (u), our DEP pipeline (d), NN (High F_1 and Recall configurations) and Stanford Pipeline (s). The presence or absence of the annotations for the users or pipelines is indicated by a preceding plus (+) or minus (−) respectively. MV indicates application of the majority voting aggregation method, and P the probabilistic aggregation method.

The Table confirms, first of all, that the domain-trained pipelines outperform the domain-independent Stanford one, as expected. Second, that in this genre human judgments only match the domain-dependent pipelines when probabilistic aggregation is used. Third, that aggregating user judgments and domain-dependent pipelines we see an improvement in F1 of up to 2.536 percentage points, but again only with probabilistic aggregation.

In Figures 8.2 and 8.3 we plot F_1 to look at how many non-expert annotators are required to rival the performance of the pipelines using the respective aggregation methods. In Figure 8.2 only the participants are shown. The Figure shows that in this genre the domain-specific automated pipeline (trained on this domain) outperforms the participants, but already at five annotators, aggregated with the probabilistic aggregation method, we are very close to the performance of the domain specific pipeline. And in Figure 8.3, which shows the results aggregating participants with the pipelines (and in which the first two participants are the two automated pipelines), we can see that we only need to aggregate 3 participants to the domain-specific pipeline to exceed its performance.

	Precision	Recall	F_1
Stanford	72.222	71.367	71.792
DEP	85.122	75.135	79.817
NN High F_1	78.090	83.151	80.541
NN High Recall	69.447	**88.833**	77.953
MV(+u -d -s)	80.293	70.786	75.240
MV(+u +d -s)	82.884	74.855	78.665
MV(+u +d +s)	77.542	78.794	78.163
MV(+u -d +s)	75.101	76.233	75.662
MV(+u +NN F_1)	85.578	77.706	81.452
MV(+u +NN R)	83.194	75.541	79.183
P(+u -d -s)	84.737	74.704	79.405
P(+u +d +s)	80.700	81.916	81.303
P(+u +d -s)	**86.770**	78.364	**82.353**
P(+u -d +s)	78.025	79.117	78.568
P(+u +NN F_1)	86.587	78.247	82.206
P(+u +NN R)	85.697	77.814	81.566

Table 8.3: Comparing pipeline and aggregation methods

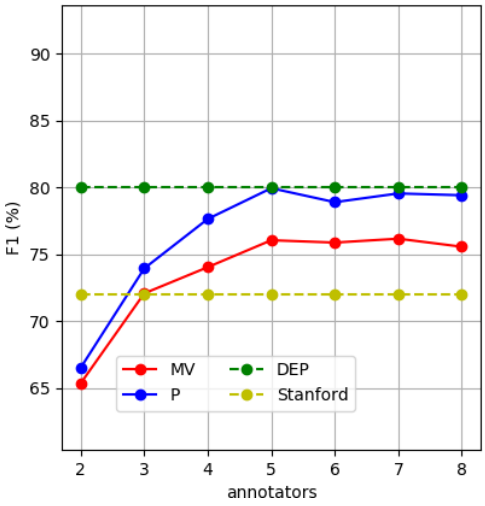

Figure 8.2: Human annotators F_1

8.2.2.2 Other Domains

431 participants in the High Recall Group and 120 participants in the High F_1 Group labelled 180 sentences.

Table 8.4 shows the results for both configurations of the pipeline with the highest score marked in bold. We can see that operating out of their original domains, the automated pipelines can not be expected to achieve the same performance as in the News genre. However, they do appear to serve well as agents to train participants to perform annotations, as participants

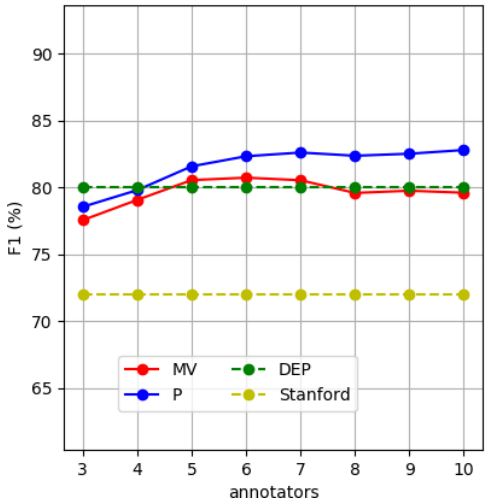

Figure 8.3: Aggregated users and pipelines (first two annotators are automated pipelines) F_1

	High Recall Experiment			High F1 Experiment		
	precision	recall	F_1	precision	recall	F_1
Stanford	77.524	80.111	78.796	77.524	80.111	78.796
MV(users Stanford)	82.152	87.065	84.537	82.260	87.065	84.595
P(users Stanford)	82.438	87.483	84.885	82.523	87.344	84.865
DEP	84.726	81.780	83.227	84.726	81.780	83.227
MV(users DEP)	**88.434**	87.204	**87.815**	**87.729**	86.509	**87.115**
P(users DEP)	87.870	86.648	87.255	87.588	86.37	86.975
NN	73.355	**93.046**	82.036	79.924	87.483	83.533
MV(users NN)	81.472	89.291	85.202	80.000	89.013	84.266
P(users NN)	81.807	89.430	85.449	84.363	**89.291**	86.757
MV(users)	87.977	85.349	86.643	86.533	84.006	85.251
P(users)	88.270	85.633	86.931	82.523	87.344	84.865

Table 8.4: Results on the 'Other Domains' dataset (rounded to 3 dp)

annotate to a high level of accuracy.[2]

8.2.2.3 Error Analysis

We analysed the nature of the errors produced both before and after aggregation. There were many errors to consider, so we took an approximate rule driven approach to characterise as many as possible.

Before aggregation, by far the most common error (1254 cases) is participants marking individual nouns as noun phrases (e.g., marking *the [cat]* instead of *[the cat]*. This suggests a misunderstanding of how the game is played that may possibly be addressed by improvements to the tutorial (Section 7.2). Similarly, in 606 cases participants mark named entities/strings of

[2]As already pointed out, workers do not perform as well as players recruited to *TileAttack* by more organic means (see Table 8.2).

proper nouns rather than the encapsulating noun-phrase.

The next most common error (529 cases) is annotators neglecting to include post-modifiers when selecting noun phrase boundaries (e.g., marking *[the cat] in the hat* instead of *[the cat in the hat]*. This is often the most popular judgement, and as such, chosen by MV. A real example of this is in Figure 8.4. In Figure 8.4, whilst five annotators did identify the correct boundaries (in green), matching the gold standard (in gold), more (six), only marked the reduced boundaries (in red) "A consortium of private investors". This sequence, missing the post-modifier, was consequently chosen by majority voting. The probabilistic method (in silver), however, expressed more confidence in the five annotators and provided a correct final judgement.

Count		0	1	2	3	4	5	6	7	8	9	10
		A	consortium	of	private	investors	operating	as	LJH	Funding	Co.	said
gold	0, 9	A	consortium	of	private	investors	operating	as	LJH	Funding	Co.	
silver	0, 9	A	consortium	of	private	investors	operating	as	LJH	Funding	Co.	
6	0, 4	A	consortium	of	private	investors						
5	0, 1	A	consortium									
5	0, 9	A	consortium	of	private	investors	operating	as	LJH	Funding	Co.	

Figure 8.4: Example of post-modifier phrase

In 122 cases participants omit the determiner.

Following probabilistic aggregation of the users' annotations (excluding pipelines), the most popular error remained identification of individual nouns as opposed to complete noun phrases, but with only 133 cases. There were 63 cases of proper names being identified without the complete noun phrase.

False negatives tend to be quite long. The average sentence length in the datasets is 31.5 (1dp) tokens, and the average markable length is 4.5 tokens, but the average false negative length is 10.3 tokens. It would appear users tend to miss the longer noun phrases.

In the texts from "Other Domains", one of most common errors produced by the automated pipelines in in cases of coordination, as in

Sammy chose ten [books and the library] said he could borrow them for one month.

where "ten books" and "the library" should be separate markables.

Another common error for automated mention detectors was prepositional phrase attachment, a well known challenge for parsers. Our automated mention detectors tend to prefer low attachment, as in

So John and Caroline filled up a [green bin with mandarins].

The example above highlights another common error with the mention detectors, missing the determiner - most commonly, quantifiers and indefinite articles.

Lastly, proper nouns near the start of sentences are often incorrectly grouped with the capitalized first token which is incorrectly also identified as a proper noun (e.g. *[First Art] sat in the car...* rather than *First [Art] sat in the car...*)

8.3 Related Work

8.3.1 Gamifying all steps of a pipeline

The GMB project includes multiple gamified interfaces as part of a platform called *Wordrobe*. These gamified interfaces are supported by prior judgements provided by an automated NLP pipeline and the *GMB Explorer* [168]. These judgements are used to generate questions for the games, which then produce corrections referred to as "Bits of Wisdom" , which in turn are automatically fed back through the pipeline and into other games and finally aggregated using majority voting [193].

The *Wordrobe* suite of games [193] include multiple games that go on to produce similar annotations to that of *TileAttack* (e.g. Named Entity Recognition). However, all tasks are represented by a single common multiple choice format. Whilst this fits efficiently into a common game design that generalises throughout all tasks, it does constrain annotator choice. In contrast, *TileAttack* targets a single yet core NLP annotation task (sequence labelling) with a broad set of applications. We do not constrain user input based on any prior judgement beyond tokenisation.

8.3.2 Aggregating markable annotations

Whilst there has been a great deal of work and evaluation on aggregating judgements from noisy crowdsourced data, this is generally focused on classification based annotations [194] and does not generalise to sequence labelling tasks like NER, IE or markable annotation. Dredze et al proposed both a "Multi-CRF" approach to aggregating noisy sequence labels, and including judgements provided by an automated pipeline, in a NER task [195]. Confidence in annotators is not modelled in this method. However, it has been extended to incorporate the reliability of the annotator with a similar method that also combines Expectation Maximization with CRF in an NER and NP chunking task [111]. Nguyen et al apply HMM and LSTM methods to aggregating judgements in NER and IE, including a crowd component in both models representing each annotators ability for each label class [112].

Whilst variations of CRF and HMM have demonstrated a great improvement over majority voting approaches, models to date have not taken into account the nested nature of sequences that occur in tasks such as markable identification.

8.4 Conclusions

In this chapter, we introduce a hybrid mention detection method combining state-of-the-art automatic mention detectors with a gamified, two-player interface to collect markable judgments. The integration takes place by using the automatic mention detectors as 'players' in the game. Data from automatic mention detectors and players are then aggregated using a probabilistic aggregation method choosing the most likely interpretation among those in a nominal head-centered cluster.

We showed that using this combination we can achieve, in the news domain, an accuracy at mention identification that is almost three percentage points higher than that obtained with an automatic domain-trained mention detector, and over seven percentage points higher than that obtained with a domain-independent one. We also test the approach in genres outside those in which the automatic pipelines were trained, showing that high accuracy can be achieved in these as well.

These results suggest that it may be possible to gamify not just the task of annotating coreference, but also the prerequisite steps to that. This shows, in answer to RQ3, that a GWAP can be very effective in correcting the output of an automated pipeline.

CHAPTER 9

PROGRESSION

In the previous chapter, sentences were presented to annotators at random and we saw how with aggregation and as little as five non-experts we could improve over a state-of-the-art pipeline. In this chapter, with a view to answering our RQ4, we look at improving individual annotator accuracy by introducing game-like progression. More specifically, we examine the application of a progressive case selection that sees sentences assigned to annotators in-line with their increasing level of competence.

Within traditional games design, incorporating progressive difficulty is considered a fundamental principle. However, despite the clear benefits, progression is not such a prominent feature of Games-With-A-Purpose (GWAPs), nor one that is commonly evaluated. There is little evidence of the effects of progression, despite the clear benefits it can bring for training non-expert annotators to produce more complex judgements. Most current methods utilise either text based, or readability based heuristics for estimating item difficulty. There is often a substantial disconnect between readability and the complexity of the task itself with respect to the text in question. In this work we present an approach to progression in GWAPs that generalizes to different annotation tasks with minimal, if any, dependency on gold annotated data. Using this method we show a statistically significant increase in accuracy over randomly showing items to annotators.

In Human Computation annotators typically have very mixed ability [40]. Traditionally, the result of this has been that in both projects based on plain crowdsourcing, and projects based on Games-With-A-Purpose (GWAPs), responses from annotators that fail to pass a periodic assessment against a gold standard [27], or pass an initial test [196], are simply disregarded, without attempting to train these annotators to carry out those labelling tasks. This approach is generally complemented by aggregation methods that learn the various annotator abilities based on their agreement [108, 185, 197] and task complexity [198] and use these parameters to weigh

an annotator's contributions.

More recently, in the interest of maximising resource utilization, crowdsourcing methods have been proposed to match annotators to specific tasks. Such methods have been found to result in better resource utilization by taking into consideration the workers specific skills, availability, and cost [199–201]. Researchers have also come to realize that whereas some human computation tasks only require very simple judgements, in other cases the pool of workers with the required background is restricted. Early GWAPs focused on context-free, decomposable tasks, all of a level of difficulty that was accessible to annotators of all skill levels, such as image labelling [46, 133, 134]. However, later GWAPs have become increasingly ambitious, used for language annotations that require deep linguistic knowledge [20], understanding the context of sentences or sometimes paragraphs [19], carry out tasks that vary in complexity [24] and sometimes domain specific knowledge [25]. Such tasks further motivate introducing some progression in the worker's task: starting with easier assignments before progressing to more complex ones when the worker has demonstrated to have acquired enough practice and/or understanding. Yet many crowdsourcing projects (using GWAPs or microtask crowdsourcing) appear to employ some form of progression, we are not aware of any work in the area proposing some form of progression and demonstrating its benefit. This is the main objective of this chapter.

Assigning to workers tasks at the appropriate levels also has benefits that go beyond the optimization of resources. Despite the advertised motivation for participating in crowdsourcing being the financial incentive, studies have shown some evidence that *fun* is one of the leading intrinsic motivators [202] and in some cases, may be even more motivating than money [203]; and this is uncontroversially the case for GWAPs [43]. This provides a further motivation for employing some sort of progression in GWAPs. Ensuring that players have the appropriate level of challenge has been shown to increase motivation [141], learning [139, 204] and enjoyment [205, 206]. Collectively, these would appear beneficial in recruiting workers, training them to perform complex tasks and retaining them over a long period of time.

Last, but not least, the type of progression explored here is very appropriate for the target players of the particular GWAP used for this study, a language annotation GWAP in which workers are asked to identify noun phrases in text, and whose primary target are players interested in linguistics or in improving their English through playing. Target players can start with simpler types of noun phrases and then progress to more complex ones once they demonstrate to have understood the more basic concepts.

In this chapter we present a method for task assignment in GWAPs aiming to present workers with tasks that match their current competence, which is dynamically reassessed possibly leading to progression to more complex tasks. We apply the method to our natural language sequence labelling GWAP, *TileAttack*, demonstrating that it results in significantly better labelling performance than random assignment of tasks to workers.

9.1 Traditional Progression Approaches

9.1.1 Training and Progression in GWAPs

Whilst historically human computation, particularly in the form of microtask crowdsourcing, focused on unskilled homogeneous tasks, such methods now aspire to address increasing more challenging tasks. This is seen to be the future of crowd work [207]. However, training is very challenging to design in microtask crowdsourcing. In contrast, games incorporate learning and provide a variety of training mechanisms that can be carried over into GWAPs. For this reason, it has been said that devising suitable methods for training players is an opportunity for GWAPs to surpass methods such as microtask crowdsourcing for complex tasks [48]. The dual motivation of progression as a means of training and providing engagement has thus been identified from the very early GWAPs for language resourcing [151]. This section will look at some methods of training and progression currently used in GWAPs. Whilst all of the progression systems described seem perfectly suitable for the tasks they attempt to address, we discuss the potential positives and negatives of selecting such an approach for a different task.

The first approach to progression found in the literature we refer to as **switching**. When switching, a system toggles back and forth between the player labelling unknown items and being assessed against gold annotated examples. When annotating gold examples they are given feedback on their label. As their performance increases, the player sees fewer gold examples, and spends more of their time labelling. In this sense, the system could be described as a progression. However, it does not account for varying difficulty items. The other apparent negatives to this would seem to be the requirement for a gold, and the reduced resource utilisation of testing a player against a gold, in which time they are not providing labels. The strengths to this system is that only one player is required at a time, a departure from the original methods [43] which can permit for more game-like interfaces [27]. We discuss here two prominent examples of what we have referred to as "switching". In the game *PuzzleRacer* [27] players provide annotations tying images with word-senses. They do this by racing through puzzle gates. Each gate has a series of images associated with it for the user to race through. The assessment/gold gates damage the players health when answered incorrect as a means of feedback. The gates through which the player provides a label have no resulting action regardless of if they are answered correctly or not. A model of the confidence of the annotator is held to determine which gate to show. Quizz [208] is a multiple choice style gamified crowdsourcing system that experimented with recruiting players/workers through targeted advertising rather than the traditional micro-payment approach offered by platforms such as Amazon Mechanical Turk. Quizz users annotate by answering multiple choice questions in a variety of domains. A Markov Decision Process is used to learn which of the two to present to a user next. This system is also designed to optimize retention.

The next method is an example of real progression, that we refer to as **domain agnostic**

progression. In the game Dr. Detective [25], players annotate domain specific named entities in medical texts. Dr. Detective models a documents difficulty as being the normalized vector of the number of sentences, the number of words, the average sentence length, the number of item types and the readability of the document (using the SMOG measure [159]). The selection process is then to find the item with the smallest difficulty increment from all items that have a difficulty greater than or equal to the current item, excluding the current item. The authors mention that they believe computing difficulty based solely on textual metrics was a weakness and that the system would benefit from a domain specific metric of difficulty. The weakness to this system is that it makes the assumption that the readability of the text is linked to the complexity of the task. A very short sentence could incorporate complex linguistic phenomena in a language resourcing task, depending on the nature of the task. A positive to this method is that it does offer progression and does not require modelling a domain specific measure of complexity for a sentence.

ZombiLingo [209] uses a **skill-based domain specific progression**. ZombiLingo includes a variety of tasks for different labelling phenomena. Different phenomena relate to different skills. The initial measure of item difficulty is based on the type of linguistic phenomena that occurs in the item, and is derived from an automated pre-processing pipeline and the corpora the text comes from. This difficulty continuously evolves based on user responses. A player must complete a tutorial for each phenomena before they are allowed to annotate. The strength to this system is that it is likely to closely model the complexity of the task. Whilst this would seem a well suited approach for ZombiLingo, it is not clear that it would generalise beyond this GWAP. The first weakness to this approach is that many labelling tasks may not be decomposable into a skill set required to complete them. The second, a reliable automated domain specific system must exist that can be used to identify the skills required to label an item. Such a pipeline or method of inferring complexity may not always exist, particularly if the task is gathering data for a new corpora.

In conclusion, there have been a variety of approaches taken to incorporating progression into GWAPs. However, as of yet it would seem there is no evaluation on the benefit of applying such mechanics.

9.1.2 Progression in Game Design

Within the context of traditional games, ensuring the player level of challenge is a very active area of research and discussion. Popular topics that are considered fundamental game design include difficulty scaling [210], user selected difficulty modes [211], dynamic difficulty adjustment [212].

When designing for challenge in games and looking at how to bring enjoyment, game designers typically look to the theory of “flow” [149, 206]. This involves presenting the player with in-game challenges that are commensurate with their increasing skill level to keep the player in the

psychological state of "flow"; an enjoyable state of elevated focus and engagement. When the challenge is insufficient, players may become bored, when the challenge is too great, players may become anxious. Designers try and keep their players in the narrow margin between these two states known as the "flow channel". More specifically, they attempt to take a meandering path through the channel (Figure 9.1) in which the player cycles between feeling the reward of applying their newly acquired skills and the challenge of acquiring new skills to meet the next challenge. In practice, this is often presented in levels in which a player perfects a skill or acquires an ability that makes the level they are currently at easier, shortly before progressing onto a new level where they face new challenges. [36]

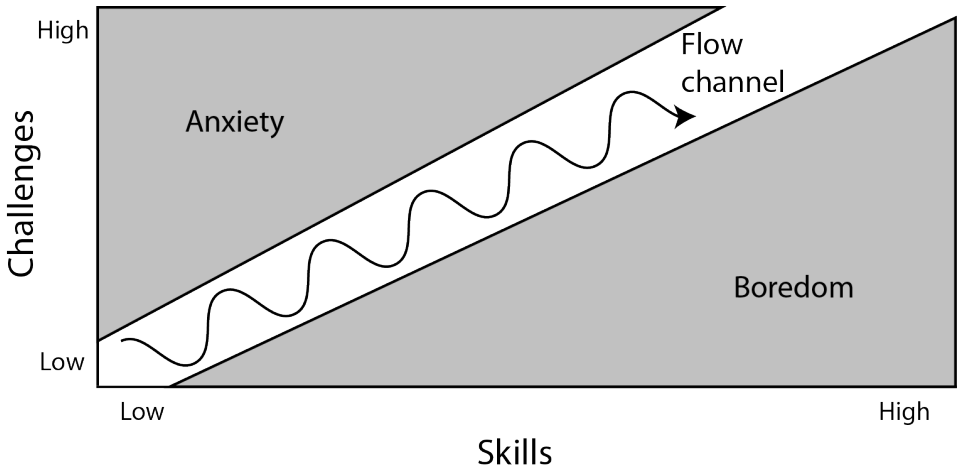

Figure 9.1: Flow Theory - Wave Channel [36]

9.1.3 Training and Progression in Learning Games

Learning games such as "Motion Math: Hungry Fish" [204], Quantum and Spunmore [139], have shown how challenge and flow are important in game-based learning, both directly in terms of the achieved learning outcomes and indirectly in terms of player engagement and satisfaction.

9.1.4 Task Assignment in Crowdsourcing

Crowdsourcing tasks rarely feature progression, or training. However, there has been multiple efforts in crowdsourcing to derive a measure of annotator skill to optimise task distribution and resource utilisation. Such methods often model annotator ability and item difficulty based on inter-annotator agreement. [199–201]

One such system is the "SmartCrowd" system. SmartCrowd attempts to find the best possible task for a worker based on the worker expertise (the level of knowledge with regards to certain skills), plus other factors such as, wage requirements and the worker acceptance rate. However, having assessed a users ability, SmartCrowd finds the best possible task for that ability and cost.

There is no progression. They do mention that it would be possible to add skill improvement into the model and discuss the merits of doing so [201].

9.2 Progression in *TileAttack*

9.2.1 Worker ability and document complexity

In *TileAttack*, each worker has a linguistic ability level, starting at 0, and the documents to annotate have a readability level. The workers' linguistic level is used to select an item from a document with a matching level.

9.2.2 Progressing to the next level

The progression principle used in the system is that a worker progresses to the next level once they have provided a sufficient number of high quality annotations at their current level. The key problem to be addressed is how to assess the quality of the annotations in a setting in which we do not necessarily have a gold. It is not sufficient to simply assume that once a worker has completed so many items to a certain accuracy they are ready to progress, as the reading levels assigned to the documents do not directly reflect the labelling complexity, and therefore, the detail required to assess the worker's competence.

Instead, the distribution of player accuracies against the aggregation of all worker labels for an item ("silver standard") is used as a picture of an items difficulty. A player is deemed ready to progress to the next level having completed 3 items with an accuracy (F_1) above Q3 of the interquartile range of this distribution.

To further motivate this choice in the context of this project specifically, the rest of this section will take a closer look at the relationship between the following sentences are of equal reading difficulty:

Item #2315: "Before that {a wooden bridge} helped {people} get across {that river}."

Item #2317: "{The other bridges} are {the Fairfield Bridge, {a very attractive bridge built in {1940}}}, {the Cobham Drive Bridge} and {the newest bridge} is {the Pukete bridge which is now {part of {the Wairere Drive Express way}}}."

However, the second sentences posses noun-phrases with multiple levels of nesting and more linguistic phenomena in relation to the task. To correctly label Item #2317, the annotator must be aware that prepositions, appositives and relative clauses can form part of an noun-phrase. In contrast, to correctly label #2315, they only need to understand the simple determiner; adjective; noun arrangement.

One possible proxy for item complexity is the number of mentions that occur in the item or the average mention token length. To provide some quantitative insight into how mention

complexity level varies alongside reading levels we chart the range of mention token length in Table 9.1 and Figure 9.2.

	# Mentions		Length Mentions	
L	μ (σ)	**min-max**	μ (σ)	**min-max**
0	3.35 (1.44)	1-7	1.84 (1.43)	1-12
1	3.07 (1.72)	1-9	1.93 (1.45)	1-11
2	3.66 (1.53)	1-8	2.19 (1.58)	1-8
3	5.66 (4.51)	1-37	2.81 (4.03)	1-78
4	7.76 (5.02)	1-30	3.60 (4.95)	1-64

Table 9.1: Document level compared to the average number of mentions per item (#) and the average mention length (in tokens) - from gold annotations

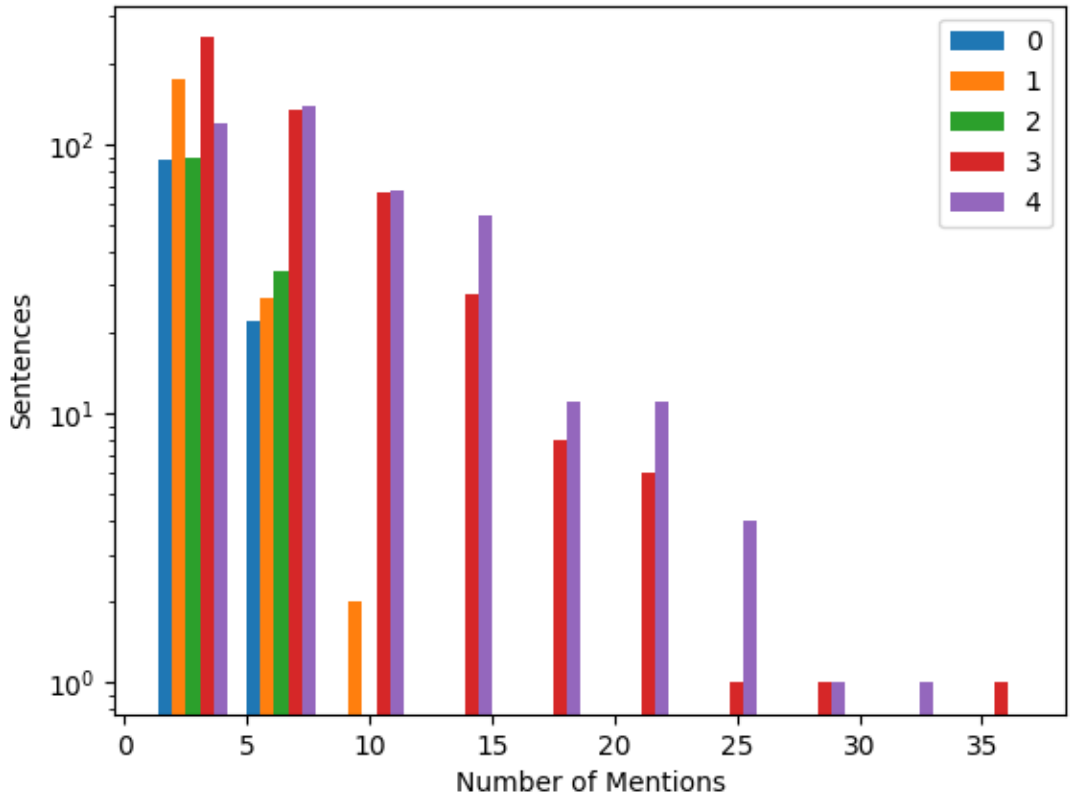

Figure 9.2: Mention length (tokens) for each level

We can see, as with the previous example, whilst the number of mentions and their token length typically increase with reading level, there are a very simple items (items containing single token markables and sentences with only one markable) at every level.

9.2.2.1 Aggregation

We can expect the non-expert labelled boundaries to be quite noisy in compared to expert annotations [40]. To extract "silver standard" annotations from the various non-expert judgements, once a sentence has been annotated 5 or more times, an aggregation step is performed. This step attempts to draw upon the shared wisdom of the annotators as a whole to extract a final judgement. Majority voting assumes equal skill among annotators, an assumption shown to be false in practice [184]. Instead, we use a probabilistic model to capture annotators different levels of ability. More specifically, a multi-class version of Dawid & Skene [108] in conjunction with method of clustering nested sequence labels. This has been found to be worst comparable, and at best excel beyond, majority voting approaches in this particular domain [3].

9.3 The Experiment

We ran an experiment to test the hypothesis that including a progression in *TileAttack*–starting by presenting workers with easier sentences before progressing to more complex ones once we have determined that they could reach a good quality of annotation with simpler documents–results in better accuracy than when presenting sentences in random order. In the experiment, participants were asked to mark noun phrases.[1]

A between-subjects experiment design was used with two groups. The first group is presented with items from levels at random. The second group uses the *TileAttack* progression mechanism discussed earlier.

Every 5 rounds an assessment round is shown. In this round the annotator's accuracy is assessed against gold annotated data from a separate corpora. The player must score greater than or equal to 30% F_1. If the player fails to stay above this level they are not allowed to continue. This is a low barrier put in place only to remove spammers from the task, not the less capable annotators.

9.3.1 Data

In order to get texts at different levels of difficulty, we used a combination of easier texts from English learning collections and 'real,' harder texts from actual coreference corpora. Specifically, the documents at the first three difficulty levels come from the "Read in Easy English" collection available from the FLAX public repository for English learning [2]. The 'real' text include a combination of Wikipedia entries, fiction, and student reports. These are the documents that we would expect to need to annotate for a real NLP corpus, and were considered to be of level 4.

[1] Specifically, the workers were asked to mark **mentions**, the noun phrases that would be identified by a mention detection system for the use of a relation extraction system or a coreference resolution systems [80].

[2] http://flax.nzdl.org/greenstone3/flax

9.3.2 Participants

Naturally, there is argument when evaluating GWAPs to use organically gathered players as participants, through means such as marketing the game, to stay as true to the natural setting of the application as possible. However, we believe that when testing for accuracy (as opposed to engagement, retention or recruitment) in a between-subjects experiment, to nullify as many individual biases as possible, the best option is to take a micro-task crowdsourcing approach to player recruitment. Taking this approach and applying minimal filtering (as mentioned above) allows us to gather a large and varied audience of participants in a short time period. We believe the lessons learned should transfer through to an organic player base. For this, TileAttack's MTurk integration is used.

9.3.3 Experiment Design

The Amazon Mechanical Turk Workers are shown the game documentation, then taken to the tutorial. They must complete the tutorial before they are allowed to perform the annotation task itself. Having completed the two tutorial rounds they are then asked to annotate three 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 three sentences, the participants are then remunerated 0.40 USD for their participation (effectively 0.08 USD/sentence). This value was based on the observation that a single game of TileAttack typically takes less than 30 seconds (which equates to approximately 9.60 USD/hour), exceeding US minimum wage [191]. When accepting future HITs participants are not required to repeat the tutorial but are, instead, asked to annotate five sentences.

9.4 Results

We take two perspectives in our results, the first focuses on the effect on users at the level of player games, the second looks at effect on the final results.

9.4.1 User Focused Perspective

We ran an experiment with **149 workers** in the *progression group* playing **3,875 games** and **156 workers** in the *random group* playing **5,669 games**. Both groups show the typical Zipfian distributions in terms of contribution (Figures 9.3 and 9.4).

We exclude any contributions from workers that did not play at least 3 games.

Table 9.2 show the average precision, recall and F_1 at the different levels for the two groups of random and progressive difficulty respectively. In levels 3 and 4 where the tasks are more difficult, we see a significant difference between the resulting agreement with the aggregation

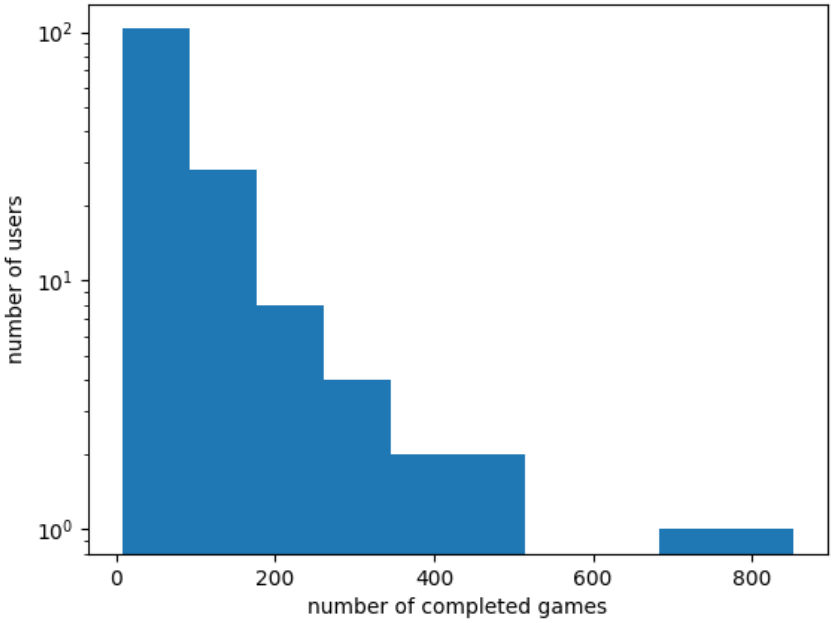

Figure 9.3: Distribution of worker contribution in *progression* group

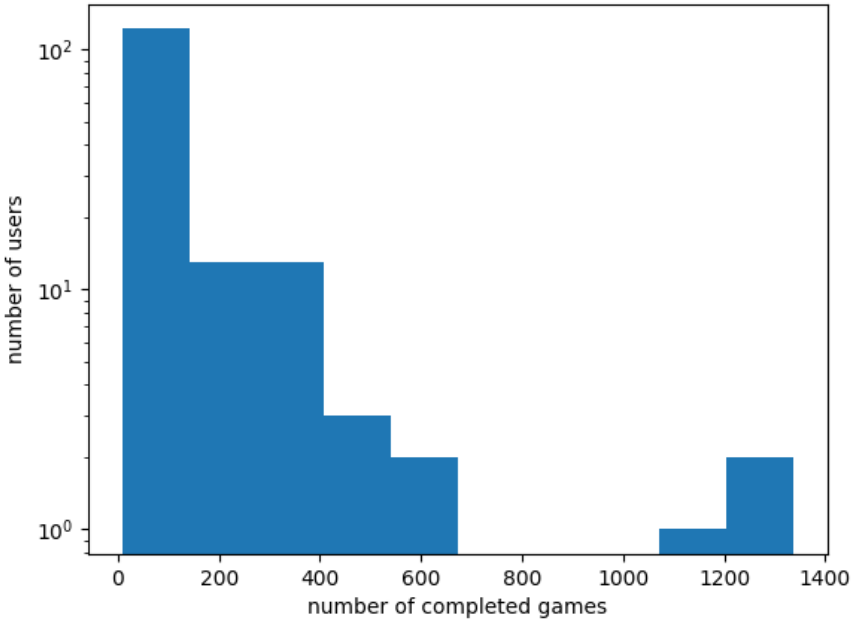

Figure 9.4: Distribution of worker contribution in *random* group

between the two groups. The groups that have been delivered tasks progressively in line with their ability score much higher. This is particularly evident with recall.

	Random Group				Progression Group			
L	**# games**	**Precision** μ (σ)	**Recall** μ (σ)	F_1 μ (σ)	**# games**	**Precision** μ (σ)	**Recall** μ (σ)	F_1 μ (σ)
0	1059	73.8 (0.266)	85.4 (0.212)	76.3 (0.217)	623	69.0 (0.300)	76.7 (0.256)	68.7 (0.253)
1	1289	69.4 (0.288)	86.4 (0.222)	73.5 (0.238)	592	69.8 (0.317)	78.1 (0.268)	70.2 (0.270)
2	1184	64.5 (0.279)	78.4 (0.244)	67.2 (0.230)	505	71.7 (0.263)	75.2 (0.239)	71.0 (0.223)
3	1424	64.7 (0.284)	74.0 (0.265)	65.7 (0.245)	1337	83.9 (0.220)	75.1 (0.241)	77.3 (0.210)
4	713	62.9 (0.273)	66.1 (0.265)	61.1 (0.237)	818	78.9 (0.235)	64.2 (0.258)	68.5 (0.227)

Table 9.2: Accuracy for worker games - *random* vs. *progression* groups **exact boundary evaluation** (rounded to 1 dp)

Figure 9.5 shows a box plot of recall for levels 2-4 - those for which there is statistical significance (see Table 9.3). On the whole, the *progression* group has a tighter distribution, with a lower standard deviation than the *random* group, in the more challenging levels. This is also visible in Table 9.2, particularly in the precision.

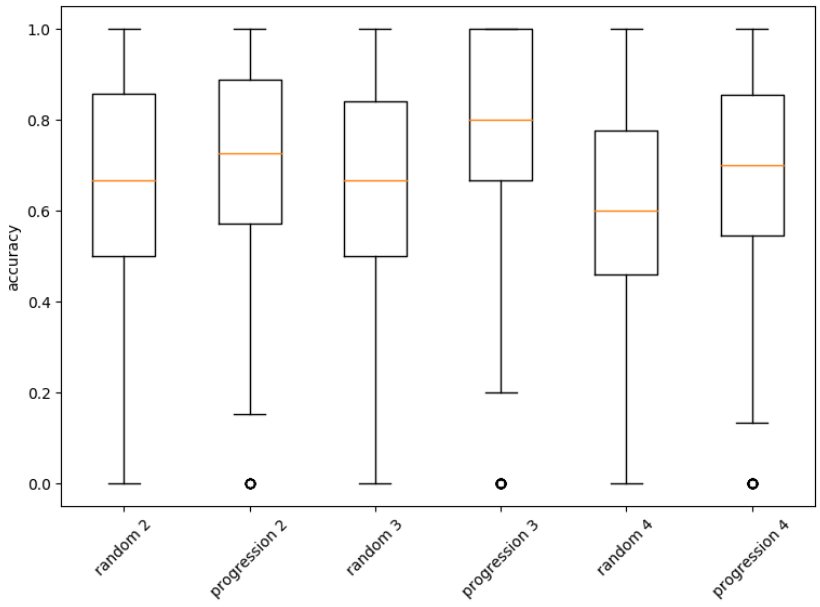

Figure 9.5: Player Game F_1 on levels 2-4 between *random* and *progression* groups

Figure 9.6 and Table 9.3 shows the difference in F_1 accuracy between the *random* and *progression* groups across the levels. Mann-Whitney U test is used to test for statistical significance. Whilst the *random* group appears to outperform the *progression* group in the lower levels (0 and 1), there is no statistical significance to this difference. This might be as a result of the fact that in the progression group, only inexperienced players ever tackle those problems, whereas in the

random group the players tackling level 0 sentences might do so as their first sentence, or their last, after gaining much experience. There is however statistical significance in the difference for levels 2-3, where the *progression* group outperforms the random group by a large margin in all levels; particularly in level 3 (11.56%).

L	**Random** F_1 μ (σ)	**Progression** F_1 μ (σ)	**Difference**	**P-Value**
0	76.3 (0.217)	68.7 (0.253)	-7.58	1.000
1	73.5 (0.238)	70.2 (0.270)	-3.32	0.973
2	67.2 (0.230)	71.0 (0.223)	+3.79	0.001
3	65.7 (0.245)	77.3 (0.210)	+11.56	0.000
4	61.1 (0.237)	68.5 (0.227)	+7.39	0.000

Table 9.3: F_1 for worker games - *random* vs. *progression* groups with Mann-Whitney U test **exact boundary evaluation** (rounded to 1 dp)

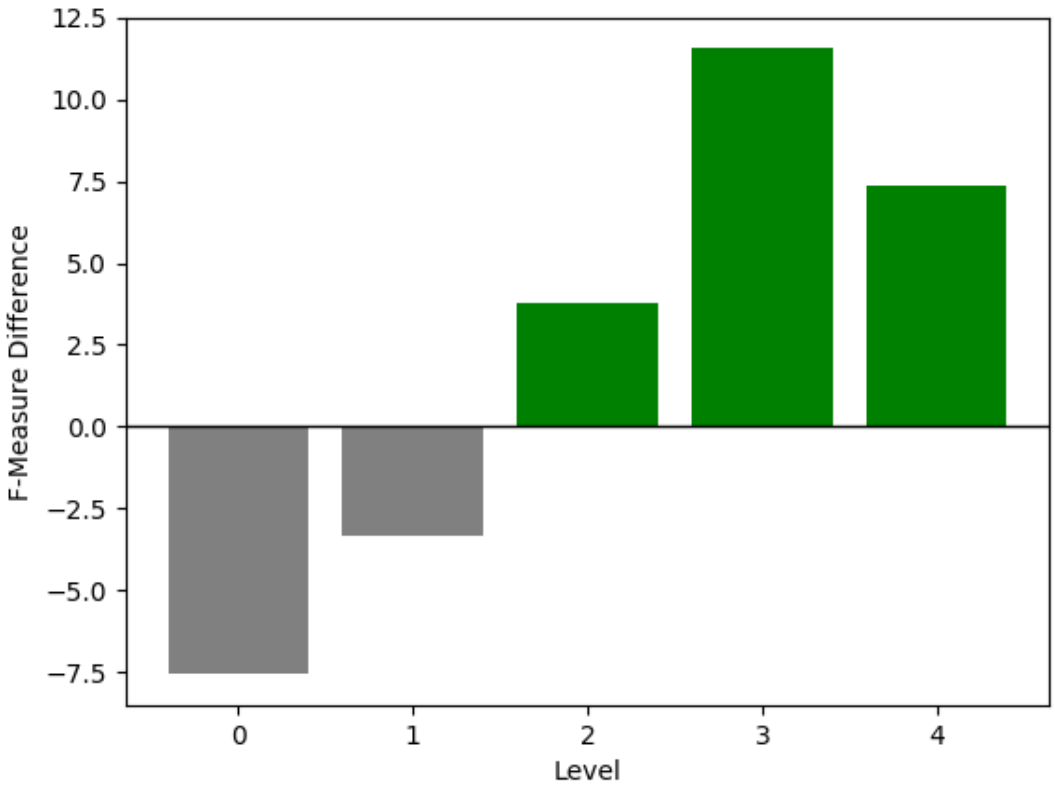

Figure 9.6: F_1 difference between *random* and *progression* groups

9.4.2 Output Focused Perspective

We consider only items with at least 3 games played. This is **688 items** for the random group and **657 items** for the progression group. We take at most the first 5 games for each item. No worker plays a game on a single item more than once. A probabilistic aggregation method is used (the very same used as part of determining an items difficulty. Both groups do well on level 0, then as the item difficulty increases, the accuracy begins to decrease. However, the progression group is far more resistant to the increase in difficulty. At the start, the random group does slightly better, there is a similar picture in the user-centric evaluation. This is probably due to the fact that some of the random players may have been playing for a long time and gained some expertise, whereas the progression players would have all been beginners at level 0. (Figure 9.7).

	Random Group				Progression Group			
L	**# items**	**Precision**	**Recall**	F_1	**# items**	**Precision**	**Recall**	F_1
0	55	90.3	86.4	88.3	55	88.5	87.5	88.0
1	103	85.2	85.5	85.4	102	86.7	85.6	86.1
2	62	82.8	78.4	80.5	62	83.7	79.3	81.4
3	256	79.9	75.3	77.5	240	90.1	80.6	85.1
4	212	78.5	66.8	72.2	198	90.8	74.9	82.1
all	688	80.5	73.1	76.6	657	89.5	79.0	83.9

Table 9.4: Accuracy at levels

Figure 9.7 shows the F_1 of aggregation at the respective levels for items labelled by both groups. As one might expect, with items labelled by the random worker group, as the difficulty increases throughout the levels, the accuracy decreases. However, in the items labelled by the progression group, whilst the accuracy of the items decreases for the first two levels in line with the increasing difficulty, the remaining levels are are far more resilient to the increasing difficulty.

9.5 Discussion and Conclusions

In this chapter, we presented a method of offering progression in a labelling GWAP for arbitrarily complex labelling tasks that support aggregation, vary in difficulty but do not benefit from easily identifiable distinct skills. We use broad, domain agnostic readability levels for identifying item difficulty, but our assessment of player ability is based on agreement against aggregation. We demonstrated this approach with a sequence labelling task of identifying candidate mentions and evaluated against randomly assigning items to players. The approach is tested via micro-task crowdsourcing in order to controlling the between-participants nature of the study, and nullify the individual biases present with organic players by gathering a much larger audience.

Our results demonstrate noticeable benefits to applying this strategy. On average, workers with the progression treatment perform considerably better on more difficult items than those

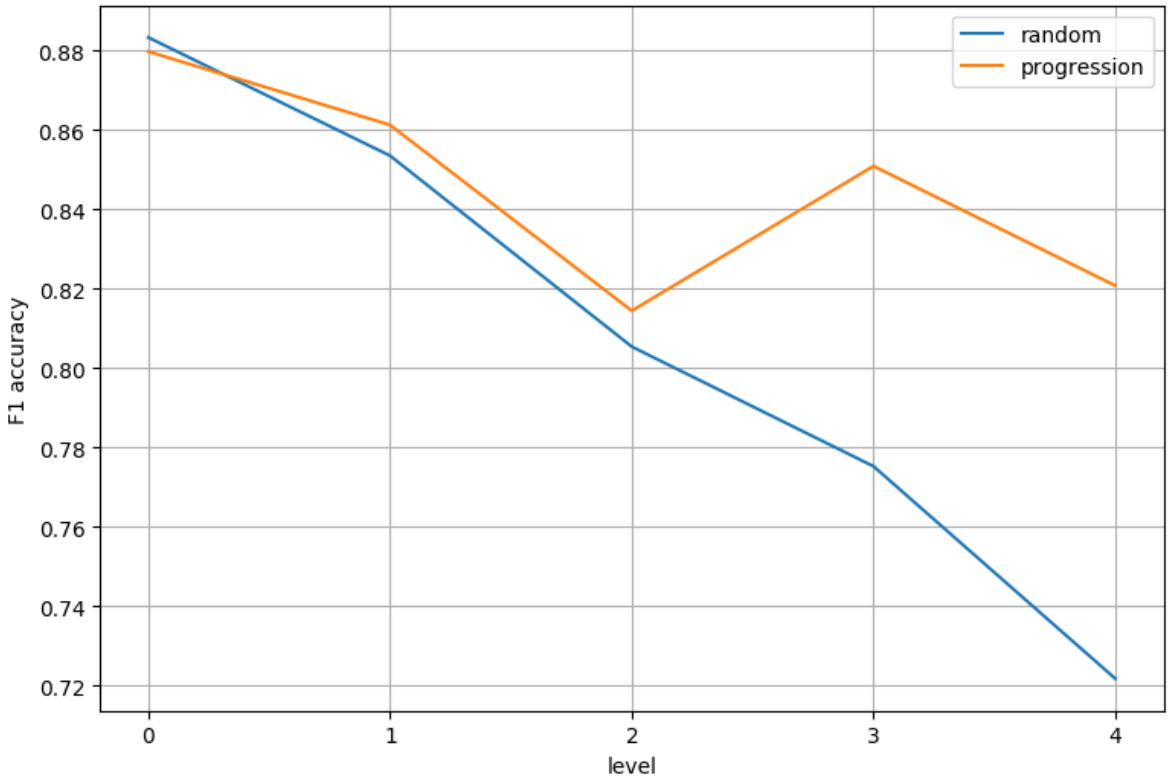

Figure 9.7: F_1 of probabilistic aggregation of annotations on items for *random* and *progression* groups

who play randomly (all with a high statistical significance).

There is a similar picture with the resulting output of the system. The aggregation of the labels provided by the progression group are much more resistant to the increasing difficulty than those provided by the random group.

With regards to our original research question, we have demonstrated in this chapter that progression can be very effective in improving individual non-expert annotator accuracy. In the next part we will explore the progression idea even further with progression between games.

Data for this experiment has been released at https://tileattack.com/data.

Part V

Token Labelling

In Part IV we discussed our work to develop GWAPs for marking noun-phrases by correcting the output of pipelines with large scale crowdsourcing recruitment and aggregation it produce high quality annotations at a sequence labelling level. In this part we look at labelling at a token level, this time focusing on organic player recruitment, game design and the notion of using a separate game to train players to perform more complex tasks. We also focus on RQ2, of whether it is possible to develop truly entertaining games for NLP.

Games-With-A-Purpose (GWAPs) for creating language resources [19, 24, 30, 137] have shown promise in terms of their ability to gather high quality annotations and in terms of scalability. However, player recruitment and retention remains a challenge with such games, that have yet to acquire or retain players at a scale comparable to the most successful GWAPs [131, 135]. The original GWAPs for AI by von Ahn, such as The ESP Game, were effective in presenting their tasks, as per the original definition, in such a way that the labels gathered were a byproduct of play [131]. In contrast, it has been said that language resourcing games such as PhraseDetectives [19] (Figure 9.8), are not really GWAPs as annotations are not a byproduct, but rather it is evident that the player is annotating text [45]. This can be said of the majority, if not all language resourcing GWAPs. Wordrobe for example, unlike PhraseDetectives, is a game which deliberately aims to hide the true nature and linguistic complexity of the tasks by presenting them as multiple choice questions and removing linguistic terminology [24]. However, it remains evident the player is annotating text. Similarly for other well-known game-like approaches to NLP resource creation such as Jeux-de-Mots and Zombilingo [30, 137]. Proper GWAPs have been proposed, but never really used for resource creation or reported high levels of player acquisition [27].

Figure 9.8: Language Resourcing GWAP: Phrase Detectives

The approach to making text annotation GWAPs more game-like followed in this work is

based on the general principle of starting from pre-existing and engaging game mechanics, just as is done in some of the most interesting GWAPs for AI [27, 170]. The question we addressed was: what type of existing game can incorporate the mechanics of text annotation? The common wisdom is that it is not possible to have game mechanics centred on annotation, and they must be on the side of an entertaining game. As discussed, such approaches have struggled. We are looking for a design that places the task of annotation organically at the centre whilst preserving an enjoyable experience.

We argue that the mechanics of 'Ville type Free-To-Play (F2P) games in general, and incremental games in particular, is particularly suited for designing GWAPs. 'Ville games, given their collective name by their common name suffix (e.g. *FarmVille*, *FishVille*, *YoVille*), are a group of highly successful games [213–215] originally targeting social network gamers (platforms such as Facebook) that share a similar novel design approach [215, 216] and monetization strategy [217], pioneered by the company Zynga. We present *WordClicker*, an incremental game whose mechanics is designed around text labelling. Tested in a training setting with audiences from three popular indie gaming portals, we show promising figures for both the entertainment value and learning. We believe the design and mechanics used are highly transferable to other games featuring annotation where game design is a challenge, such as serious games and language resourcing GWAPs.

Our first contribution is the proposal to adapt the so-called 'Ville game [218] mechanics for text labelling games. We believe this type of game design addresses a lot of the challenges to be addressed by text annotation GWAPs (or indeed, GWAPs for any type of annotation). In the paper we draw a parallel between the interests of annotation games and that of 'Ville style games and their F2P monetization strategy. We believe this type of design, in which entertaining games are created out of intrinsically repetitive activities, is uniquely suited for annotation games, in which the objective is to keep players performing unentertaining activities for a long period.

Our second contribution is the idea of using a game of this type to address the problem of getting the players of a GWAP to understand the phenomenon about which judgements are collected without boring them by asking them to read instructions. As pointed out by Tuite [48], the complexities in modern GWAPs, that attempt increasingly more difficult tasks, is also perhaps their biggest opportunity to excel beyond other types of crowdsourcing. Adding a training or learning element to crowdsourcing has shown to increase accuracy, but is difficult to do [138]. Games, however, have been shown to be an effective tool for teaching [139] and learning has been said to be a key part of the fun of a game [140]. Many language resourcing GWAPs already use a variety of training mechanisms borrowed from games. Phrase Detectives for example uses the traditional tutorial approach [19]. In ZombiLingo, tasks are split into different subtasks for each linguistic phenomena, training the player only on the subtask they are about to attempt [30]. We are proposing here a different approach: to develop a separate game specifically devoted to teach the linguistic knowledge required to successfully play a text annotation GWAP. Specifically, the

annotation game presented here was developed to train players to understand parts-of-speech categories at a lexical level (e.g. nouns, proper nouns, adjectives), so that they can proceed to successfully play a GWAP designed to label categories at a grammatical layer (e.g. noun-phrases). (At the moment our priority is to ensure that our game can teach, acquire, retain and motivate players to annotate text. Thus, while we believe the design ideas proposed here can be used to actually collect parts-of-speech labels, in the work reported in this paper we only carried out training-oriented experiment to evaluate the mechanics in this regard.)

We demonstrate the effectiveness of the two ideas just discussed through our third and final contribution, our training-orientated annotation game called *WordClicker*. We will discuss how we believe the selection of incremental game mechanics support the desired outcomes and within the constraints of the challenging design space.

CHAPTER 10

WORDCLICKER

This chapter introduces the game *WordClicker* (available to play at https://wordclicker.com). I created *WordClicker* as a base for experiments in relation to game design and training. *WordClicker* is a GWAP for token labelling and aims to offer an engaging game-like experience with that annotation as a core mechanic.

10.1 Related Work

10.1.1 Free-to-Play Games

The F2P (Free-to-Play) revenue model has become a popular method of reaching casual gamers on web and mobile platforms [37, 219]. These audiences would not necessarily consider committing to an initial purchase, but may consider small purchases to enhance experience as they progress in the game [37]. For games that employ this revenue model, F2P motivates a specific set of design objectives [220] which, we will argue, also apply to GWAPs. For example, the fact that there is no initial financial commitment from the player (unlike in games based on the traditional revenue model), means that the game needs to appeal to the player right from the start, as there is nothing to stop the player putting it down if it is too difficult to master for a casual gamer, or is not immediately entertaining. Consequently F2P games commonly feature a shallow learning curve. To make another example, to integrate the concept of in-game purchases, many F2P games feature a "double currency model" that allows players to purchase more of the in-game currency they have earned through in-game actions, with real money [219]. These purchases take place over a long period of time, so F2P games are often designed to have infinite or long lasting content and retain their players over a long time. Although the games are designed to be played over a long time, they are also designed for inclusive play, allowing casual gamers to pick them up and

put them down in many short play sessions. [220]

F2P mechanics, influenced by behavioural economics and behavioural psychology [216], are guided through extensive instrumentation, user data and other analytics [220]. These investigations originated a substantial body of knowledge concerning which game design patterns to apply based on the type of game and results the game producer is looking for. This is particularly useful in GWAPs, as whilst not all existing studies may translate directly between the two domains, it can be used to inform starting investigations. The key design element in Free-to-Play games, is their core game loop, with optional waiting step [37], (Figure 10.1).

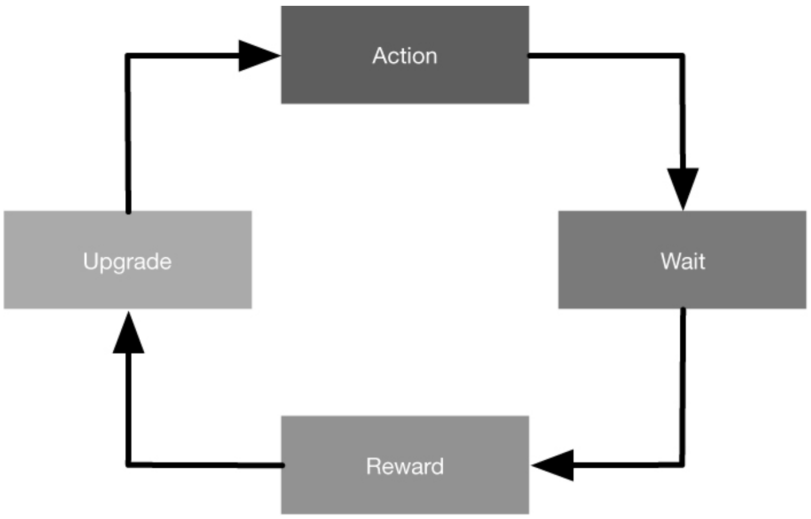

Figure 10.1: Free to Play - core game loop [37]

10.1.2 'Ville Games

So-called "'Ville Games" are, we would argue, a particularly relevant category of F2P games for GWAP design. Following the advent of social networks [221], various organisations, particularly Facebook [222], opened their platforms to the embedding and distribution of third party applications. Of these, was a set of multiplayer games that allowed friends to play together known as social network games (SNGs). Often using the Free-to-Play revenue model [217], offering inclusive play to casual gamers [217], and being web-based (accessible on a variety of devices [223]), they quickly shot to success [213–215]. One particularly successful group of such games is the 'Ville group of games created by Zynga that share the "Ville" suffix to their names (e.g. *FarmVille*, *FishVille*, *YoVille*). Over time, led by the successful 'Ville titles, the design of popular SNGs began to homogenize into a common set of studied design patterns [215, 216].

There are reoccurring design patterns that appear in these games [216]. In the gameplay, a player action results typically results in gathering an in-game resource which develops over time. A further player action realises a reward from **harvesting** the resource as in-game currency.

Resources that the player has failed to convert to in-game currency after some given timer period **wither**. Aside from losing the potential reward of the resource, the player loses their investment in the resource and possibly incurs an additional penalty. In terms of progression, they often use a standard XP style level system. Additionally, there are often quests that, under the guise of missions, train the player to perform actions in an interactive tutorial like fashion. They have a variety of resources, aside from the previously mentioned **game-specific resource** and **in-game currency**, there is sometimes **energy** that constrains the rate at which the player may perform actions. As SNGs, 'Ville games feature multiple social orientated mechanics. These include **gifting**, and a sort of **leaderboard** for viewing friends achievements. [216]

The core game loop of action (purchase resource), waiting (resource appreciates in value), reward (resource converted to currency), upgrade from F2P, is very evident in the design patterns of 'Ville games. [37]

10.1.3 Incremental Games

The aforementioned **'Ville Games** and their wider SNG genre, have been the subject of satire with critics creating games with deliberately bland core game mechanics, such as "Cow Clicker", that involves simply clicking once every six hours [224]. This widely mocked, but undeniably successful [225] game design pattern spawned a sub-genre of games that distilled the 'Ville paradigm known as "clicker games" in which the player repeatedly clicks to earn points which they can use to purchase items that enable them to earn more points [218]. Being satirical of F2P and 'Ville, these games have mostly left behind the F2P monetization strategy and social element, but still hold the key motivational design elements.

There are now many variations of "Incremental games", with some research proposing a taxonomy [226]. The key defining factor that separates them is the spectrum of interactivity with the player [218], The previously mentioned "clicker game" variety is the among the highest level of interactivity and the lowest "zero player games" in which the player's role is reduced to that of a spectator for the majority of the game [226].

Many games have continued these ideas more seriously and there have already been successful entirely text based "clicker games", such as "A Dark Room" (Figure 10.2) [38]. Exploiting behavioural psychology and decision making, these games appear to have, in part, changed our definition of what we believe a "good" game is [216].

10.2 Annotation Games and F2P

In this section we argue that there are systematic commonalities between the objectives of annotation games and 'Ville games that justify the adoption of a 'Ville game mechanic for GWAPs. We will start by looking at the relationships between the F2P games so show how design choices have cascaded through. In Figure 10.3, we select from and extend previous work on the taxonomy

Figure 10.2: Clicker Game - “A Dark Room” [38]

of incremental games [218, 226], with parts relevant to this work, to give a clear overview. We move on to summarising the similarities in a table, then discussing them individually in further detail.

Others have identified the potential of F2P mechanics in gamification to enhance motivation and increase retention [226]. The only application of any F2P paradigm in GWAPs we are aware of is *RoboCorp* [227]. However, whilst that work referred to the game mechanics of F2P, it is focused specifically on the notion of exchanging the “micro-payments” for work (“micro-work”) rather than incorporating the F2P game design principles. More specifically, rather than a directly integrated closed loop, this was an annotation task (identifying named entities in texts from the Polish National Corpus) that the player could perform to build virtual currency that could then be used to purchase upgrades to play a separate mini-game [227], whereas in this work annotation remains part of the game itself. We are not aware of any others that attempt to use F2P style mechanics, or the more specific incremental game mechanics we propose applying.

10.2.1 No initial payment/commitment

‘Ville games are commonly Free-to-Play (F2P), meaning the game is not sold, but rather the vendor receives revenue from players prolonged engagement by charging a small fee for purchases in game that typically enhance in-game mechanics, shorten game loops or add to the aesthetics. Similarly, GWAPs also receive no immediate benefit from the user initially being able to play the game, but rather a long term reward in terms of the player performing work as they continue to play. For this reason, they are both designed to be infinite and prioritise the retention of players. Serious games used in schools or in place of learning materials are unlikely to be able to charge students.

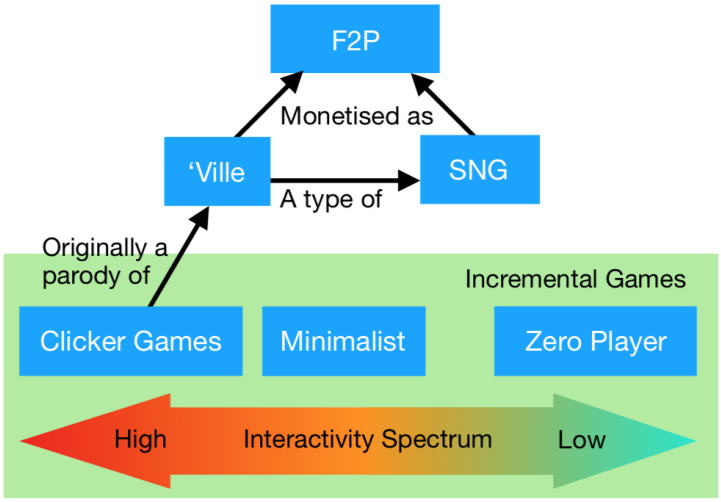

Figure 10.3: Taxonomy of relevant genres

Incremental/F2P games	**GWAPs/Annotation games**
F2P: Designed to achieve return through **ongoing payments**	Achieve return through **ongoing work**
F2P: Require no **initial commitment/payment**	Cannot ask for **initial commitment**
F2P: Have very **uneven player contribution** receiving the majority of their revenue from a very small minority of their players (known as whales)	Receive the majority of their work from a very small minority of their players
SNG: Built for **inclusive play** targeting **casual gamers** on social networks	Need to attract as broad an audience as possible so **inclusive play** is highly beneficial
Incremental: Are often Indie Games: **inexpensive** and small teams	Need to be **inexpensive** to be preferable to crowdsourcing/expert annotation
Incremental: Are designed for long term or **infinite gameplay**	Need long term or **infinite gameplay** available to allow for annotation of large resources
Incremental: Use a design with a **high ludic efficiency**	Require **high ludic efficiency** to operate effectively as tools

Table 10.1: Incremental/F2P vs. GWAP/Annotation and at what level the commonality occurs

10.2.2 Uneven Player Contribution

Both games face challenges with player bases that contribute in terms of either revenue or work, very unevenly. Phrase Detectives reports that, in the Facebook version of their game, 1.6% of its players made 89% of its annotations [228]. They describe their overall contribution distribution which they suggest is Zipfian in its nature (one of the power law distributions). Similarly, in a survey of ZombiLingo players, it is reported that of 986 registered players, there were 20 that

they considered "heavy players" [162], approximately 2.03%.

There is a similar situation in F2P games, where it is 5% of all players is estimated to be a good conversion rate from non-paying to paying customers [37] and that, as with Phrase Detectives, this distribution of, in this case revenue, matches a power law distribution[37]. In F2P gaming industry parlance the heavy players players are known as "whales" [37]. Whales typically represent less than 1% of the player base but can account for 50% or more of the revenue [229, 230]. This places greater importance on retaining converted users, maximising their contribution for their lifetime of play and attracting as broad an audience as possible.

10.2.3 Inclusive Play

'Ville style games, or more broadly speaking SNGs, were designed for inclusive play [213]. Created to target social network users who may not have been gamers before, they were designed to be picked up and put down by a casual gamer at the gamer's convenience rather than being played for long periods of time, and on a variety of devices [215].

10.2.4 Inexpensive

There are few comprehensive cost analysis to draw upon from either type of game, but it is known that modern games typically have large budgets. Since the 1980's companies have been spending millions of dollars in the marketing of their games alone [143]. More recently costs often run into the tens of millions of dollars for development [144–146] and tens if not hundreds of millions for marketing [144, 147]. SNGs however, cost comparatively less than their more conventional counterparts [231], else they risk never recuperating their investment. In GWAPs, one must keep the initial cost of game development low and development time fast, ruling out starting with a large project. If the project is overly expensive or takes a long time it may be faster and cheaper to use alternate methods (e.g. crowdsourcing). For example, after the first two years, the cost of annotation with *PhraseDetectives* was equal to the cost using microtask crowdsourcing, but the final projected cost for completed annotation of the corpora is 50% of the estimated cost of using crowdsourcing [19]. Whilst *PhraseDetectives* has evidently struck a good balance, had the creators invested much more in game development it may have been more cost-effective to use alternate methods. Serious games face an almost identical challenge in which they must return a better educational value than that offered by similarly priced educational materials [232]. "Clicker games" can be, and most often are [226], created by very small teams or individuals [225, 233], in very little time [233], inexpensively and without much expertise.

10.2.5 High Ludic Efficiency

Furthermore, there is the juxtaposition of a text labelling tool being a game/toy. Challenge in a game is artificially introduced in the form of internal goals, for the sake of entertainment. Tools

however, are designed to reduce the challenge of achieving an external goal [141]. This dissonance in design is nicely summarised by the idea that if games were good tools/applications, they would simply be a “Win game” button [142]. In game design terms, this could be referred to as having a high “ludic efficiency”, but also clearly not a “fun” interaction [181] by the definition of Flow [149]. In the vast majority of game design paradigms achieving a ludically optimal experience means introducing additional artificial challenge in line with the players skill level. This is counter-intuitive when designing an annotation tool. Introducing additional artificial challenge could constrain annotators contribution to the bounds of their ability to complete the artificial challenge, rather than the primary external annotation goal. For example, whilst many find shooting games enjoyable, a designer of a GWAP would not want to sacrifice annotation quality or quantity by introducing such a mechanic if it hindered good annotators ability to provide labels because they had a poor aim. Introduction of a game mechanic into GWAPs that may disrupt the primary task has been termed “orthogonal game mechanics” [48]. Clicker games give an illusion of challenge whilst having a high ludic efficiency, making them suitable for adaptation into tool like systems.

10.2.6 Metrics

A final, but key similarity relates the way that F2P games and annotation games are evaluated. A set of Key Performance Indicators have been proposed for F2P games, to track the effectiveness of their current design and marketing strategy throughout the player lifetime [172, 174]. Already sharing statistics such as ALP and throughput [170], we argue (see Chapter 5) that these metrics are a very effective way of evaluating GWAPs as well. We apply these metrics testing *WordClicker*.

10.3 Training through playing

Linguistic competence can be characterized in terms of layers: phonetics, phonology, morphology, lexical knowledge, syntax, semantics, and discourse. The ability to use language at one layer requires competence at the lower layers: For example, understanding the concept of a noun phrase requires understanding what a noun is. But these competencies are typically seen as distinct. In the context of game design, it has been said: “New data is all it needs to flesh out a pattern. A new experience might force a whole new system on the brain, and often the brain does not like that. It’s disruptive.” [140]. We propose therefore that the training required to understand linguistic concepts at a lower layer is best achieved through a separate game, so as not to “force multiple patterns on the brain”.

Learning using drills, flashcards, or generally learning by rote, remains the preferred method in modern Mobile Assisted Language Learning (MALL) with the most popular apps [234] such as Duolingo [235], Busuu and Memrise [236] using spaced repetition algorithms that calculate the optimal interval with which to test a player to ensure long-term memory. We believe that

the compelling yet highly repetitive nature of clicker games is an engaging approach to the often otherwise tedious process of learning by rote [237].

Use of a derivative game mechanic can be used to construct an environment for formative assessment. A one-to-one mapping between concepts and resources results in the player not being awarded resources for concepts they fail to comprehend. The game will effectively be encouraging them to invest more time practising the unfamiliar concepts. For example, in the case of *WordClicker*, players that only understood nouns, would be consistently lacking the pronoun and proper noun resources/ingredients. This is quite gently enforced in *WordClicker*, but one could modify this depending on the training requirements. For example, making combinations of ingredients a requirement to progress (e.g. players must acquire nouns and pronouns).

10.4 Design

In this section we will discuss the design of *WordClicker*, our annotation game and our adaptation of clicker game mechanics.

WordClicker is a web-based [1], desktop and mobile friendly, one-player game in which a player learns the classes of words by playing a baker that gets her/his ingredients by clicking on words associated with those ingredients. The core game mechanics is simply classifying individual words into classes (associated with ingredient jars) by clicking on them, a mechanic that should be transferable to the majority of word-labelling tasks. If the player is correct, after clicking they get ingredients, that are used to make the cakes. The game is very simple, taking approximately two weeks for one person to develop.

10.4.1 Story

The story takes inspiration from the game "Cookie Clicker", in which the player plays the role of someone with a cookie business. In *WordClicker* the player plays the role of a cake shop owner. Their job is to produce cakes by discovering the relevant ingredients, and the business by choosing when and how to reinvest their profits in expanding their business (buying bakeries), improving production (ovens) or increasing efficiency (improved equipment/the amount of ingredients found). Unlike a normal cake shop, the owner/player is responsible for finding individual ingredients. The player must identify words that match an ingredient or part of speech to collect them.

10.4.2 Art

From the outset, the game is styled like a cake shop front with a red and white awning, and a noticeboard that introduces the game (Figure 10.5). This is theme is continued into the game itself.

[1] https://wordclicker.com/

Ingredients and part of speech tags are colour coded to communicate the link between labels/ingredients and game elements with players. Each ingredient has a colour, this is used in the ingredients jar and on correctly labelled tokens. When an ingredients jar is selected, the background colour of the game changes to that colour to reflect the current selection.

The cake pictures, both the cake in the foreground and those falling in the background, reflect the players currently available ingredients. All the combinations of cakes available, depending on the availability of ingredients, are shown in Table 10.2.

The game is predominately 2D, making use of a minimalistic, vector drawn, cel shading graphics style reminiscent of early Social Network Games. However, there are small 3D elements to add emphasis. For example, the instructions are shown in the form of a 3D cookbook that opens with an animation (Figure 10.7). The cakes falling in the background are displayed at various sizes and fall at various speeds to give the illusion of depth.

Three fonts have been selected. Two are heavily stylised serif fonts. The first gives the appearance of a very old cash register, and is used to display the virtual currency. The second has a restaurant menu feel to it, and is used throughout the in game displays. The third is a sans-serif font that is used only in instructional or help settings for enhanced readability.

Animations are used to communicate the relationship between resources and the virtual currency. A correctly identified token shows an animation of a colour coded square travelling from the token to its ingredients jar, and from the ingredients jar to the cake.

PN	N	P	A	V	cake
✗	✗	✗	✗	✗	
✓	✗	✗	✗	✗	
✗	✓	✗	✗	✗	
✗	✗	✓	✗	✗	
✗	✗	✗	✓	✗	
✗	✗	✗	✗	✓	
✗	✓	✗	✓	✗	
✗	✓	✗	✓	✓	
✗	✓	✗	✗	✓	
✗	✗	✗	✓	✗	

PN	N	P	A	V	cake
✗	✗	✓	✓	✗	
✗	✓	✓	✓	✗	
✗	✓	✓	✗	✗	
✗	✗	✓	✗	✓	
✗	✗	✓	✓	✓	
✗	✓	✓	✓	✓	
✗	✓	✓	✓	✓	
✓	✗	✗	✓	✗	
✓	✓	✗	✓	✗	
✓	✓	✗	✗	✓	

PN	N	P	A	V	cake
✓	✓	✗	✗	✓	
✓	✓	✗	✗	✗	
✓	✗	✗	✗	✓	
✓	✗	✗	✓	✓	
✓	✗	✓	✗	✗	
✓	✓	✓	✗	✗	
✓	✗	✓	✓	✗	
✓	✓	✓	✓	✗	
✓	✗	✓	✗	✓	
✓	✗	✓	✓	✓	
✓	✓	✓	✗	✓	

Pickup	
Sieve	
Whisk	
Baking Tray	
Mixer	
Mixing Bowl	
Oven	
Piping Tool	
Scales	
Bakery	

Table 10.2: Cakes For Parts-of-Speech/Ingredients
(Proper-noun: PN; Noun: N; Pronoun: P; Adjective: A; Verb: V) and Examples Of Pickups

We give the illusion of a three dimensional background by having various size cakes falling at different speeds. The ingredients present on the cakes changes depending on the ingredients the player has available.

10.4.3 User Interface and Game Controls

A panel like interface is used to easily support the responsive design. The interface operates at a variety of screen sizes, including mobile. When viewed on a mobile, the primary interface panels collapse into a vertical view. The secondary panels are available as modal interfaces via buttons that then show in the navigation bar. All interactive elements are large buttons designed to be suitable for touch screen or mouse use.

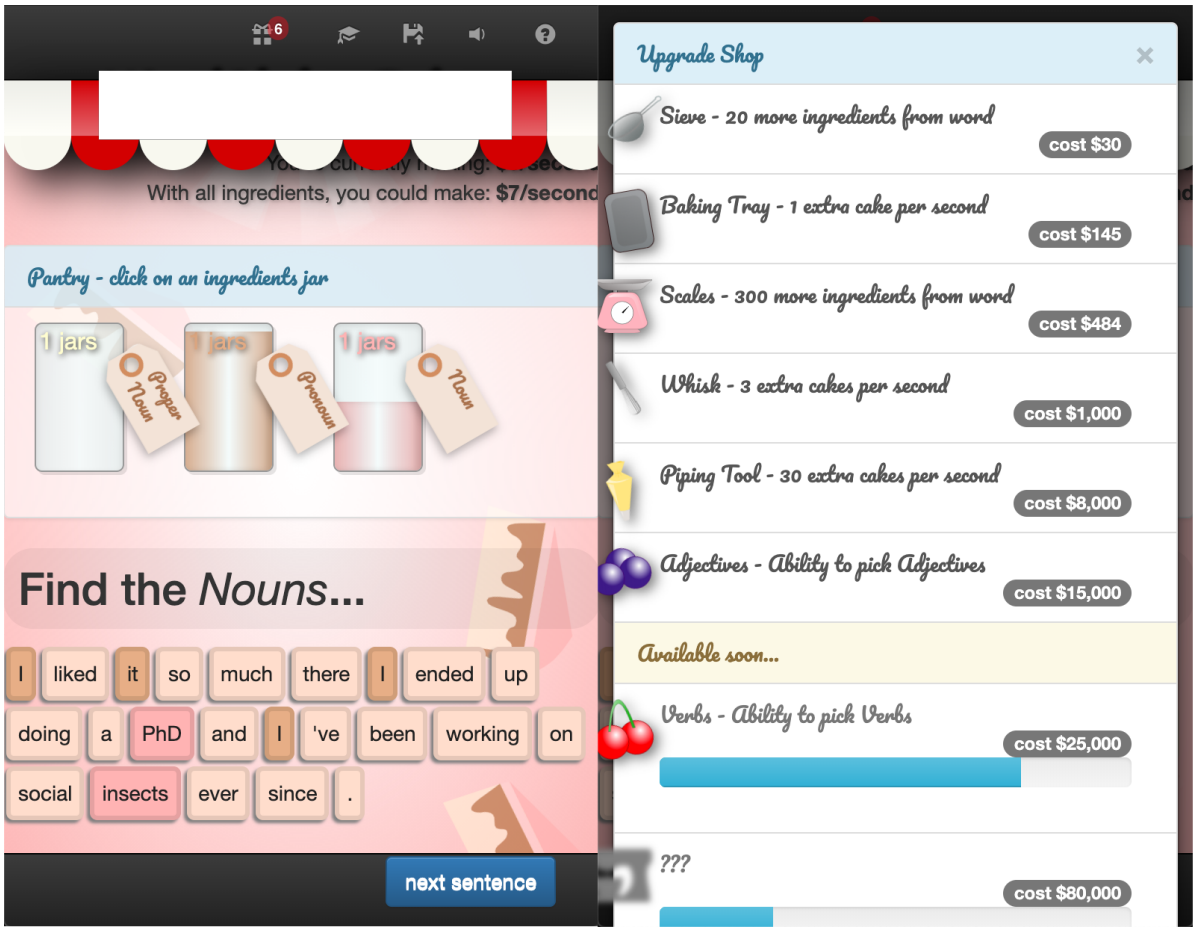

Figure 10.4: *WordClicker*- Responsive Interface (Game and shop side by side)

The game is styled like a cake shop front with a red and white awning. The help system has a three dimensional menu theme (Figure 10.7).

Each label is associated with an ingredient. Cakes are displayed both in the background (falling) and in the foreground. These show the player which cake is currently being made from their available ingredients. All the combinations of cakes available are shown in Table 10.2. In addition, colours are associated with the ingredients to show the player which ingredient jar they currently have selected. Selecting a jar changes the background colour of the screen to remind

the player.

The ingredients displayed on the cake change depending on which ingredients the player has available at that time, as do the ingredients on the cakes falling in the background.

10.4.4 Sound and Music

A dissonant two note sound effect descending in pitch is given to feedback an incorrect action to the player. A harmonious sound increasing in pitch is used to feedback a correct action.

10.4.5 Gameplay

To begin with, the player is shown details of the task they will be performing with a short explanation (shown in Figure 10.5). When they press play they are presented with an interactive

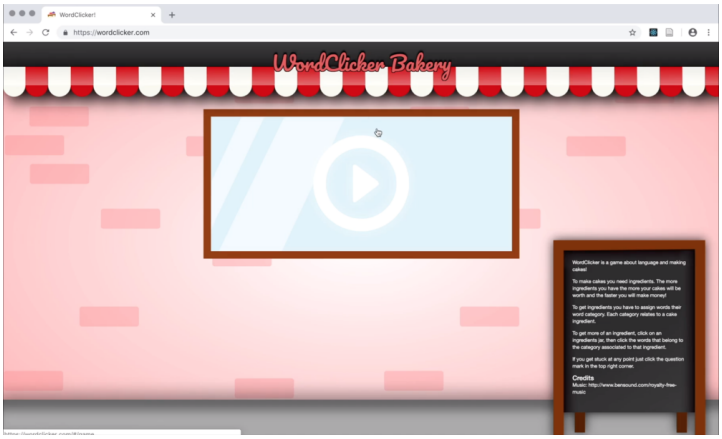

Figure 10.5: *WordClicker*- Introduction

tutorial that takes them through basics of the game (shown in Figure 10.6). They can repeat

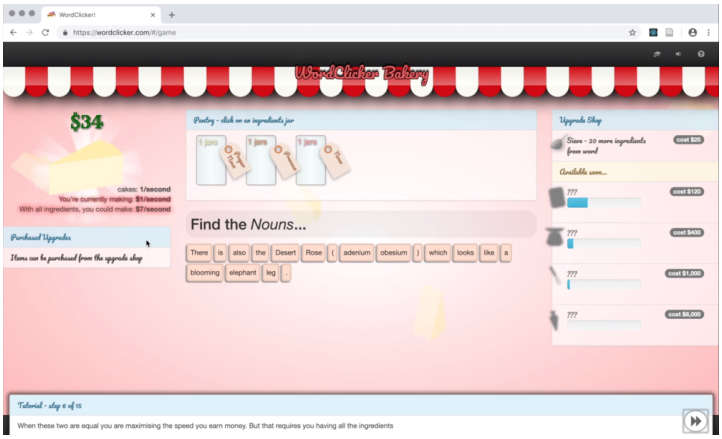

Figure 10.6: *WordClicker*- Tutorial

this tutorial and view additional instructions regarding the classes at any point (shown in Figure 10.7). During gameplay, the player is shown a single sentence at a time (see Figure 10.9). They

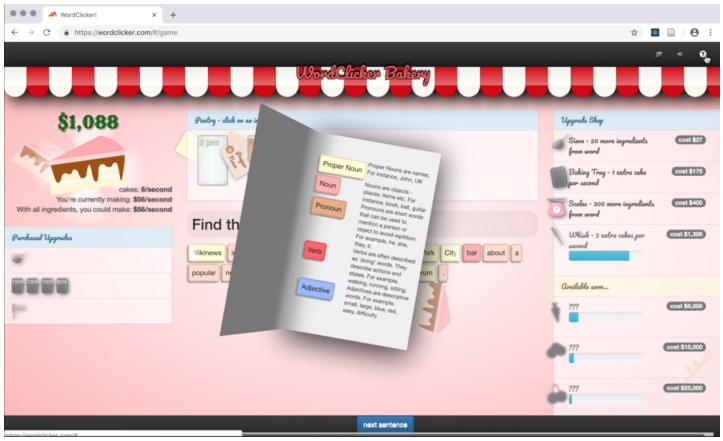

Figure 10.7: *WordClicker*- Instructions

can advance to the next sentence by using the "Next sentence" button. Once players have earned a sufficient amount of in-game credits they unlock and are offered the opportunity to progress onto a language resourcing game (see Figure 10.8).

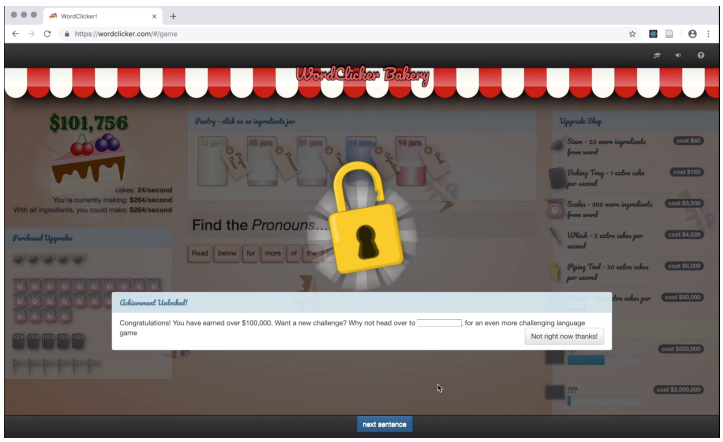

Figure 10.8: *WordClicker*- Progression

10.4.6 Mechanics

In this section we zoom in at a more granular level on the gameplay focussing on the core game mechanics, organised by their place in the incremental game design loop.

10.4.6.1 Action step

In the action step of the game loop, the player collects the resources (ingredients). With the goal of **high ludic efficiency**, the user interaction involves simply selecting the appropriate ingredients jar (category), then selecting one or more words in the sentence that are of that ingredient (category). The incremental game design choices negates the need for adding orthogonal mechanics (e.g. shooting the appropriate tokens).

Accumulating ingredients are shown in their respective jars. An animation is used to show the ingredient moving from the token to the jar. The correctly marked token is then shown with a shimmering effect.

10.4.6.2 Wait Step

In the wait step of the loop, cakes are automatically produced and sold in the quantity specified by the current multiplier (in the generator) giving the player a reward. Resources that the player has gathered (cake ingredients), are consumed synchronously and added to the cakes when available. This relationship is illustrated to the user through an animation that shows ingredients leaving the jars and moving to the cake and the ingredients being shown on the cake itself. The more ingredients a cake has, the more it is worth. The player is shown the cakes potential worth and their current worth in the game. This is designed to encourage the player to explore all of the labelling categories currently available to them, to maximise their potential gain by leveraging the notion that players do not want to waste their purchases, known as the *sunk-cost fallacy* [216, 238]. Here we are using a *avoidance fixed interval schedule* with *fixed avoidance schedule* (known to be suitable for a slow but steady response) [216] underneath to soften it. That is, the players receive a reward based on their investments regardless, but they receive far less reward unless they manage to continue to steadily find ingredients. Here we are directing players towards marking labels.

In some games, the wait part of the loop would block, preventing the user from taking another action. Here, the wait is non-blocking to encourage players to continue to annotate whilst the ingredients are being consumed. This short non-blocking game-loop is also designed to appeal to **inclusive-play**, allowing the game to be easily picked up and put down without a long time commitment.

10.4.6.3 Reward

We require no action (e.g. harvesting) on the part of the player to receive their reward. However, there is a deliberate disconnect between the resource that is gathered and the virtual currency (known as a derivative game [226]) to add an additional opportunity for control that is utilised, as described in the wait step, to motivate the player to label all the categories.

10.4.6.4 Upgrade

Upgrades are purchased from the shop by the player investing their primary reward and affect the game in two ways. They can either, increase the generator multiplier or increase the quantity of resources produced by correctly applying a label. These purchases effectively either increase gameplay interaction speed, or slow down the interaction game whilst preserving the reward. This gives the player an added choice. The faster player may opt to invest more in upgrades that increase the multiplier and ingredients consumption. The slower player may favour investing more in upgrades that produce more ingredients, so they do not regularly find themselves in the situation where all of their ingredients have been consumed. The cost of each upgrade increases infinitely, providing potentially **infinite gameplay**, and exponentially, with each purchase (in line with typical idle game formulas [239]).

Leveraging the *goal-gradient hypothesis* that players exert more effort when approaching a reward [216], upgrades are obscured in the store until the player has almost sufficient funds, and a progress bar shows how close the player is to being able to purchase that reward (Figure 10.9). Here again, we direct players towards marking labels.

As the game progresses the player also has the opportunity to purchase additional labelling categories. This allows for a configurable, self-paced player progression.

10.4.6.5 Penalizing incorrect responses

When the player labels incorrectly one of their purchases, if available, becomes damaged. This negative reinforcement leverages the players loss aversion to encourage considered annotation. This also has the natural effect of only penalizing players after they have been playing for a while and have likely understood interacting with the user interface. At the beginning, up until the point they start making purchases, they are just given feedback. This is a very unusual approach in an incremental game, that otherwise usually only offer positive rewards of various sizes. Our motivation for using a negative reinforcement is that should this design be transferred to a GWAP, the absence of any penalty would most likely encourage a gameplay strategy in which the player clicked all tokens quickly in search of the correct label, leaving the annotation process to chance. This, in turn, would result in an imbalance of high recall and very low precision.

A more descriptive feedback is given in the form of a text notification message that appears in the bottom left hand corner and a flashing red outline on the token (shown in Figure 10.9).

10.5 Ethical Considerations Affecting the Design

In this Section we discuss the aspects of the game design more directly motivated by ethical considerations. In particular, we discuss aspects of the design related to so-called **dark patterns** in 'Ville games, classified into three groups: temporal, monetary and social-capital based [240].

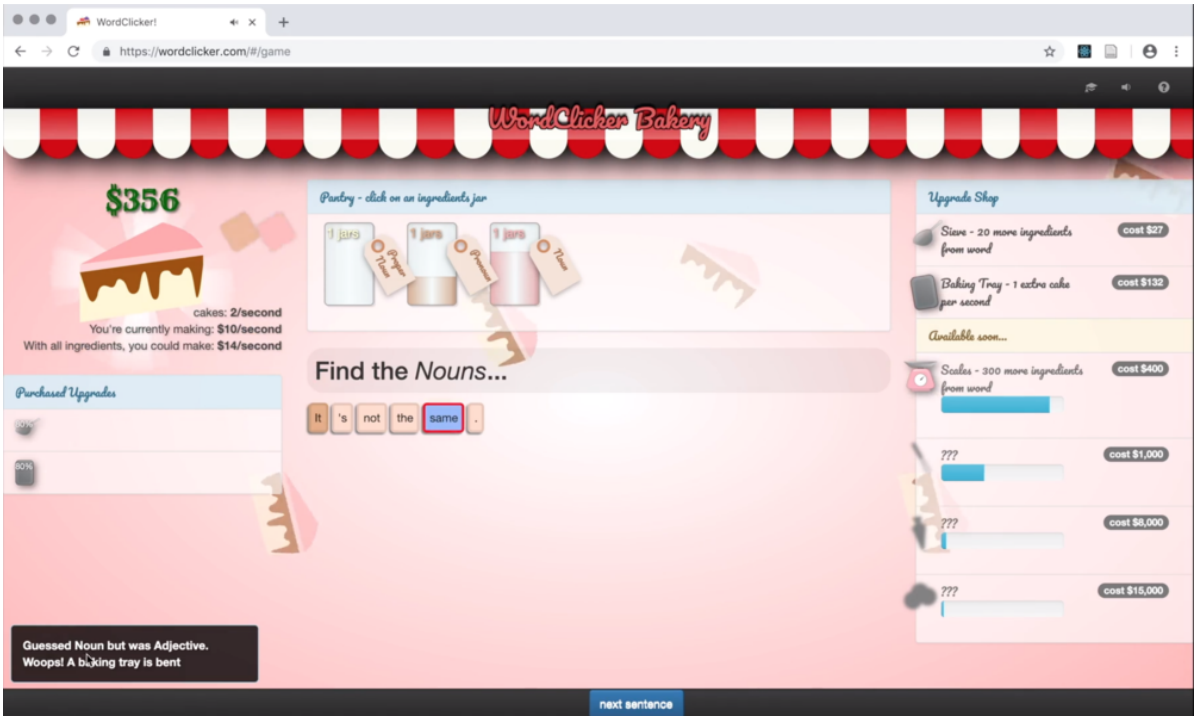

Figure 10.9: *WordClicker*- Gameplay with errors and feedback

10.5.1 Temporal

10.5.1.1 Grinding

Grinding is a mechanic that sees the player performing a repeated activity over and over, investing time as opposed to skill. This allows the game designer to pad out the game out to extend its duration without introducing new features [240]. This is very common in massively multiplayer online role playing games (MMORPG) and social network games [37]. The criticism is that players may not be able to judge how much time is likely to be demanded of them [240]. However, many take the view that grinding is a necessary part of any game with infinite gameplay [241, 242]. Both incremental games and GWAPs unavoidably feature grinding as a key part of their design. *WordClicker*, like most GWAP designs and incremental games, does feature grinding heavily.

10.5.1.2 Playing By Appointment

Playing By Appointment is the practice of applying the sunk cost fallacy, which punishes the player by destroying the resources they have spent time gathering or invested in, unless they return to the game to collect them at a certain time [216]. The strong effect of loss aversion trains the player to return regularly to ensure long term retention [37]. This normally appears in the form of harvesting/withering. This pattern is criticised for forcing players to orientate their

real world schedule around their gameplay as opposed to fulfilling their personal desires to play [240]. Aside from increased long term retention, the potential benefits for application in a GWAP would be a means of offering delayed reward. However, this is potentially harmful, therefore *WordClicker* does not make use of a harvesting mechanic.

10.5.2 Monetary

It would seem that the biggest ethical questions and problems occur in monetization, particularly when designers conflate the notion of in-game time with real world money [242, 243]. No financial transactions take place in *WordClicker*; however it is still important to consider potential problems in this area. GWAPs extract value from players in terms of work rather than money. GWAPs and gamification have also been the subject of extensive ethical discussion [240, 244]. Regardless of how value is extracted from players, it is important that it is done so responsibly and fairly.

Examples of potentially harmful strategies in this area include **pay-to-skip** and **pay-to-win**. Pay-to-skip encourages players to pay to skip grinding or, in some cases, slowly increasing the difficulty of the game until they are forced to pay to continue. Pay-to-win encourages players to purchase items that give them an advantage over other players or some status [240]. Many game studios [245, 246] have now dropped such strategies over ethical concerns. *WordClicker* does not have any sudden changes in pace or progression that demand heavy workloads from players to allow them to progress or enhance their status. In fact, whilst our results show that players have an inclination to continue to work throughout the game, they can progress without working at all.

Another questionable, more recent practice in the area of monetization, is that of **loot boxes** [247], which bear such resemblance to gambling [247] that they are the subject of new gambling legislation in many countries [248]. These are indicative of an ongoing convergence between games and various gambling like themes [249] that we also avoid in this work. This raises another concern that there is a potential for such games to lead players into gambling, although studies currently seem to indicate that there is no such link [250].

10.5.3 Social-Capital Based

WordClicker does not make use of any social features so we look only very briefly at the potential issues regarding their application in this work.

The most common approach here is the application of social obligation. A game requires a player to recruit their friends, and for those friends to play the game for the player to advance, who are in turn required to invite more friends, in what has been described as a “Social Pyramid Scheme” [240].

10.5.4 Our Deployment

These game design and wider genres (e.g. 'Ville games and MMORPG's) are experiences people enjoy. The general justification for application of the more aggressive design choices seems to be to assume cognizance on the part of the user [37, 246]. But many players fall victim to certain strategies, contributing disproportionately to their peers, which can have severe reprocussions on their lives [251]. There is still much work to be done to understand the ethical design and implications of modern games. We hope that none of the choices we have made in our design can have a negative impact. To try and mitigate any issues we have done our utmost to be transparent about the nature of *WordClicker* and the experience players should expect.

True to the satirical style of clicker games, and unlike games in the parent genres, we took a sterilised version of the mechanics, free of social integration, payment or sudden requirements to progress (pay-to-win), and lay them out in the open [252] to deliver a very obvious version of the compulsion loop. To further ensure this awareness, we selected popular portals with means of marking games (tagging/categorisation) as "Incremental Games" (or similar). The nature of the game is also reflected in its name. We would therefore expect that the player is conscious of the nature of the experience they can expect, and that this is something they have sought out, possibly as an alternative to a F2P equivalent.

It would be naive to image that just because *WordClicker* is transparent about its nature, does not apply social features or take money from players that there are no ethical implications, but we have approached our implementation as cautiously as possible. There are many opportunities to extend this work that have been deliberately avoided pending further investigation of ethical considerations. Despite the numerous benefits and opportunities, we urge objective and judicious caution on behalf of future designers applying and extending this work.

CHAPTER 11

ENGAGEMENT

In this chapter we carry out an experiment to evaluate the *WordClicker* game-design in relation to RQ2, by testing the game with an audience of a particularly discerning community of game creators and players via three indie games portals. We evaluate the results of this using the aforementioned F2P metrics. We also measure how players' performance improves over game rounds, to test the viability of progression to latter more complex games, as proposed in the previous chapter.

11.1 The Experiment

WordClicker was publicised via integration with three popular indie games portals: NewGrounds [1], Kongregate [2] and itch.io [3] and measurements taken over a 70 day period. We evaluate these results using the aforementioned F2P metrics. As discussed, there are few comparison points, so whilst both games are very different, we compare against Verbosity as one of the few games that uses a subset of metrics used here [18]. *WordClicker* requires a corpora, or large body of annotated data. There were several key criteria when selecting the corpora. There needs to be a sufficiently permissive copyright licence that we could present the texts in the context of a game; the part-of-speech tags needed to be of particularly high quality and whilst not a requirement, ideally the texts would be interesting to read. The GUM corpora [253] was selected as corpora matching these criteria. The part of speech tags were labelled from scratch and annotations were verified by experts in the case of disagreement.

[1] https://www.newgrounds.com/
[2] https://www.kongregate.com/
[3] https://itch.io/

11.2 Results

Here we report the results using adapted F2P metrics (Section 5). Metrics have been selected specifically to show player contribution (i.e. average judgements per player and lifetime judgements). Many of the F2P metrics evaluate success in the context of an advertising campaign and the costs associated with that campaign (e.g. Cost per Judgement), or virality. As this is not a focus of this study or current design (pending further ethical considerations), these metrics are omitted.

Graphs are shown with a logarithmic axis as, as expected, the data loosely conforms to a power law distribution.

11.2.1 Average Judgements (Tokens) Per Player: 32.17

This is the average number of **tokens** marked with a part-of-speech tag marked per player, per gaming session. The maximum number of tokens marked in a single session was 763.

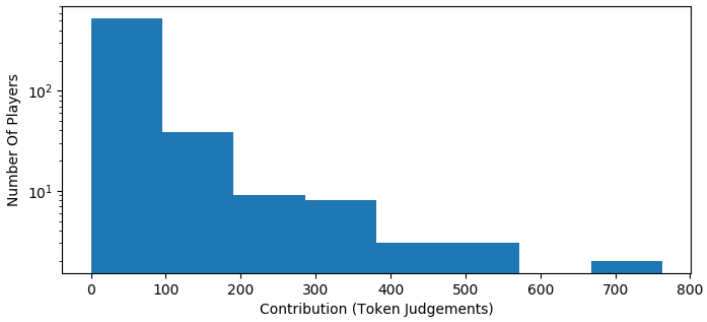

Figure 11.1: Average Judgements (Tokens) Per Player

11.2.2 Average Judgements (Sentences) Per Player: 8.31

This is the average number of **sentences** viewed per player, per gaming session. All players view at least one sentence. The maximum number of sentences viewed in a single session was 212.

11.2.3 Lifetime Judgements (Tokens): 45.85

As previously mentioned, many players return to play in more than one session. This is similar to statistic reported above, but for all of a players sessions (their lifetime). The maximum number of tokens annotated by a player over their lifetime of play was 790. In comparison, players of *Verbosity*, one of the original GWAPs, designed to collecting "common sense facts" are said to have provided on average 29.47 judgements per player [18].

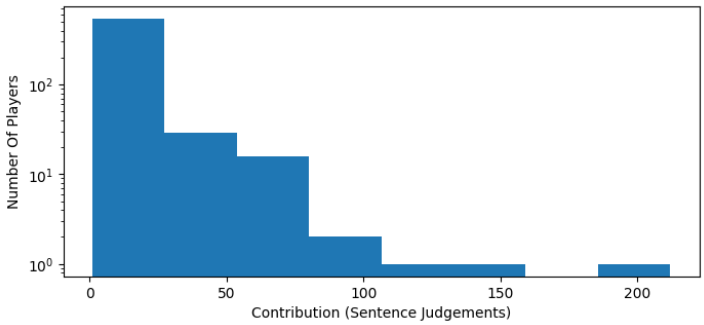

Figure 11.2: Average Judgements (Sentences) Per Player

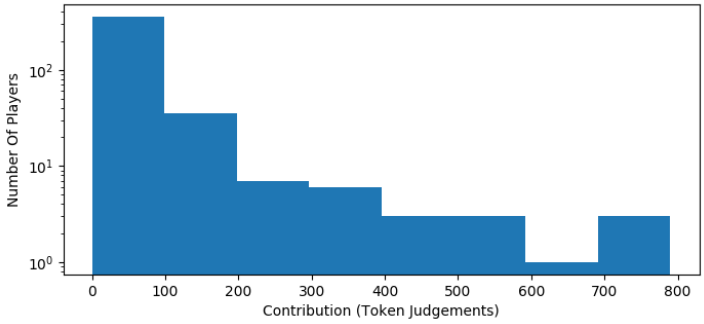

Figure 11.3: Lifetime Judgements (Tokens)

11.2.4 Lifetime Judgements (Sentences): 11.85

As above for tokens, but for sentences viewed. The maximum number of sentences viewed by a player over their lifetime of play was 212.

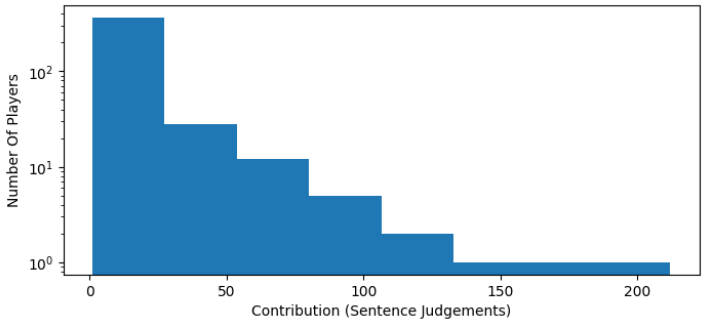

Figure 11.4: Lifetime Judgements (Sentences)

11.2.5 Average Session Length: 25mins 17.3secs

This is the average length of a play session. In comparison, the average session length for *Verbosity* players was 23.58 minutes [18].

11.3 Training

To test whether *WordClicker* improves players' understanding of the task as they played, we calculated the accuracy of their labelling on each consecutive sentence they labelled against the selected corpora discussed earlier. Figure 11.5 shows the cumulative moving average of all player accuracies for each round and the number of players that were playing at that round for the first 20 rounds of play. We do not count rounds/sentences in which the player did not label any tokens.

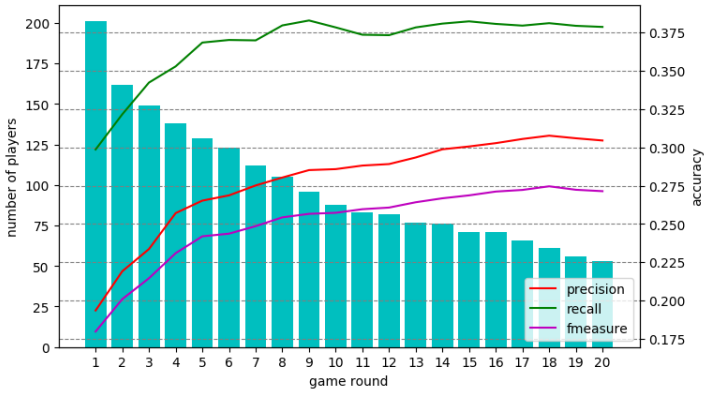

Figure 11.5: Average player accuracy over multiple rounds

To account for the possibility that players that continue to play were of a high skill level and disproportionately effect the learning rate, Figure 11.6 shows the learning rate only for the players that completed at least 20 rounds.

11.4 Discussion

11.4.1 Enjoyability

GWAPs are typically evaluated with regards to the amount of data they produce rather than their enjoyability, so there are not many previous results we can compare to, but for those few cases when such figures are available (Lifetime Judgements, Average Session Length) the results reported in the previous Section are very encouraging and appear to suggest that we succeeded in designing an annotation game for text as enjoyable as the original games developed by von Ahn.

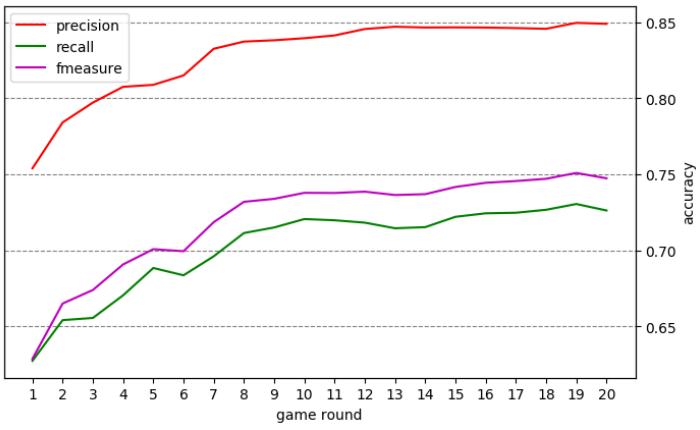

Figure 11.6: Average player accuracy over multiple rounds for only players that have played 20 rounds

11.4.2 Using *WordClicker* to annotate data

In this experiment *WordClicker* was only evaluated in a training/assessment mode, but actual GWAPs alternate between this mode and annotation mode, in which data are collected. (E.g., *Phrase Detectives* alternates between feeding players gold data and data for which no annotation already exists.) In this Section we briefly discuss the steps that we (or somebody else wanting to use the *WordClicker* for this purpose) would need to take to introduce an annotation mode.

The one difficulty to be addressed when using the game for data collection is how to award points when the answer is unknown. A number of mechanisms for doing this have been proposed in the literature. In particular, Von Ahn proposed a number of methods for games involving two players [170]. Clicker games are single player, but the methods proposed by von Ahn could be adapted for this context. For instance, we could supplement the output-agreement method. The original output-agreement game was capable of supporting a single player game, by replaying a previous player's labels on the same content [132]. We could adopt this idea.

A third option is to reward judgements retrospectively once, subject to aggregation of player labels, a reasonable level of certainty has been attained. Delayed rewards are an easy fit into F2P games where they commonly take the form of **return triggers** [37]. These return triggers have the added benefit of having a positive effect on player retention. This is the main strategy adopted in *Phrase Detectives*, where most points are awarded through validation.

11.5 Conclusion

This chapter has focused on RQ2, of whether it is possible to create truly entertaining games for NLP. We present an adaptation and application of clicker game mechanics to address the challenging design space of game-like text annotation. To be considered a truly effective approach

and useful for integration into a GWAP, a text labelling game and its constituent mechanics should be capable of acquiring players alongside traditional games. We test the effectiveness of this approach by making the game available to discerning game players/creators on indie games portals and evaluating with adapted F2P metrics. There are very promising results. In some cases, where available and comparable, we show an improvement over statistics presented for one of the original GWAPs.

In addition, continuing with RQ4 we propose the idea of using a separate game for training players before they proceed to a GWAP. Here we show a steady improvement on average player ability over time.

Future research may adopt one or more of the methods in the literature discussed, to test *WordClicker* directly as a GWAP.

Part VI

Conclusions

Prior work has demonstrated the potential benefits of GWAPs compared with other human computation methods. In particular, the application of more game-like approaches in the context of complex skill based tasks, such has language resourcing. Acknowledging these benefits, more game-like GWAPs for NLP related tasks has been the goal of multiple works. However, despite many interesting ideas being proposed, many of these games have failed to achieve the quantity or quality of annotations comparable to GWAPs in other domains when applying this approach. This work has examined these past approaches, their results and identified some of the challenges in accessing the opportunities that the GWAP approach offers. These challenges have been formulated as four research questions.

The rest of this section will look at the detailed conclusions in the context of the respective research questions and discuss potential future work.

RQ1: Is it possible to develop enjoyable games for NLP that produce quality output? The overarching question for this thesis was whether it was possible to develop enjoyable games for NLP that product quality output. Before we could answer that question there were multiple other questions that had to be answered around game design, methods of measuring game design effects and NLP.

This work started by taking a comprehensive look at the challenges and opportunities in GWAPs for NLP (Section 3.3). Following that study several key challenges emerged.

RQ2: Can we develop truly entertaining games for NLP? Despite multiple design approaches being proposed, engagement and play focused data on GWAPs is largely under reported, with most focusing on accuracy. This made it difficult to determine which approach is best. We tackled this with a new set of proposed metrics that we use throughout our evaluation of the games presented in this work.

One of the original motivations for creating new GWAPs for NLP was information bottlenecks in the highly successfully Phrase Detectives. To address this, we looked at gamifying the entire pipeline of tasks, with three games, targeting some of the most common tasks in NLP. These feed information into each other. In addition, we discuss progression between these games to enhance non-expert player understanding.

Whilst it was clear there were multiple benefits to producing a more game-like game, it also became clear that text annotation was not an easy fit into games. We found a very constrained design space. To address the question of whether it is possible to develop entertaining and game-like GWAPs for text annotation, we first conducted an in depth study of the design challenges, this revealed a very constrained design space. This work zones in on the mechanic of text annotation itself. We decided that to be considered a truly effective approach, a GWAP should be capable of acquiring players alongside traditional games. The design was tested by making an implementation, named *WordClicker*, available to discerning game players/creators on indie

games portals and evaluating with our proposed metrics. Results showed an improvement over statistics presented for one of the original GWAPs.

RQ3: Can we improve on the performance of an automated pipeline using a GWAP to correct its output? Both when performing expert and non-expert annotation, the preferred method in the literature was to improve upon an existing pipeline where available. We needed a game-like method of conveying our uncertainty about this data source that didn't involve discussion of automated pipelines. The pipeline was used as a second player in our game design. To address this challenge, we looked at the complex nested sequence labelling task, mention detection. Our methodology combined state-of-the-art automatic mention detectors with a gamified, two-player interface to correct markable judgements that were then clustered and aggregated using probabilistic methods. We showed that this method allows us to derive value from existing pipelines out of their original domain in combination with non-expert annotation.

RQ4: How can we perform annotation with non-expert players? The final challenge was utilising the training and progression benefits of games to raise our accuracy with non-expert annotators. Our use of a tutorial, probabilistic aggregation, use of multiple agents and aggregations as opponents helped train our players. However, we then took this a step further with a novel method of assessing players performance and progressing players through increasingly difficult tasks. Crucially, for annotation purposes, our approach works in the absence of a gold standard or detailed difficulty labels. This raised both player performance and final accuracy greatly for more the more complex annotations.

In summary, this work has identified a range of challenges that exist in GWAPs for NLP, from how to recruit and train players, to performing large scale high accuracy annotation with non-experts. For the purpose of controlled experimentation we examine these concepts largely in isolation. We hope future can combine these for non-expert large scale, inexpensive, complex annotation, with high accuracy.

Future Work This work has sought to identify and address the core challenges in utilising the GWAP approach for language resourcing. Broken down into three separate research questions, there is strong evidence to suggest that, individually, this work has been successful in answering or at least offering a contribution to these challenges. Naturally, each challenge has been addressed in isolation to remove confounding variables.

In some places this work has made suggestions over how these methods could be combined to a greater collective effect (e.g. patterns for adapting WordClicker into a full GWAP in Section 11.4.2). However, there is the opportunity to ask further questions in relation to the interplay between the solutions proposed in this work when applying them to form a complete game-like GWAP. For example, whilst we have tested and demonstrated benefits towards accuracy when applying a progression approach (see Section IV), when applying progression in conjunction with

the game design approach given (see Section V) one may wish to investigate in detail the impact of progression in that game design setting with regards to both accuracy and entertainment.

This can be said of virtually any of the contributions in this thesis. While they have been investigated in isolation, one might hypothesise that combined, annotation, progression and the game designs proposed could be combined to great effect and merit significant further investigation at their intersection. Furthermore, many of the lessons learned may well be applicable outside the domain of text annotation which, as discussed, presents one of the more challenging domains for GWAPs. These include the largely domain agnostic metrics (Section 5), progression method (Section IV) and game design approach (Section V).

BIBLIOGRAPHY

[1] C. Madge, J. Chamberlain, U. Kruschwitz, and M. Poesio, "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*, (Amsterdam, The Netherlands), pp. 397–404, ACM Press, 2017.

[2] C. Madge, U. Kruschwitz, J. Chamberlain, R. Bartle, and M. Poesio, "Testing game mechanics in games with a purpose for NLP applications," in *In Proceedings of Games4NLP: Using Games and Gamification for Natural Language Processing.*, (Valencia, Spain), p. 2, 2017.

[3] C. Madge, J. Yu, J. Chamberlain, U. Kruschwitz, S. Paun, and M. Poesio, "Crowdsourcing and Aggregating Nested Markable Annotations," in *Proceedings of the 57th Conference of the Association for Computational Linguistics*, (Florence, Italy), pp. 797–807, Association for Computational Linguistics, July 2019.

[4] C. Madge, J. Chamberlain, U. Kruschwitz, and M. Poesio, "Progression In A Language Annotation Game With A Purpose," in *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2019, Washington, USA, October 28-30, 2019.*, (Washington, USA), AAAI, 2019.

[5] C. Madge, R. Bartle, J. Chamberlain, U. Kruschwitz, and M. Poesio, "Incremental Game Mechanics Applied to Text Annotation," in *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '19, (Barcelona, Spain), pp. 545–558, ACM, 2019.

[6] C. Madge, R. Bartle, J. Chamberlain, U. Kruschwitz, and M. Poesio, "Making Text Annotation Fun with a Clicker Game," in *Proceedings of the 14th International Conference on the Foundations of Digital Games*, FDG '19, (San Luis Obispo, California), pp. 77:1–77:6, ACM, 2019.

[7] C. Madge, R. Bartle, J. Chamberlain, U. Kruschwitz, and M. Poesio, "The Design Of A Clicker Game for Text Labelling," (London), IEEE, 2019.

[8] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: an Architecture for Development of Robust HLT applications," in *Proceedings of the 40th Annual Meeting*

of the Association for Computational Linguistics, (Philadelphia, Pennsylvania, USA), pp. 168–175, Association for Computational Linguistics, July 2002.

[9] C. Müller and M. Strube, "Multi-level annotation of linguistic data with MMAX 2," 2006.

[10] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii, "brat: a Web-based Tool for NLP-Assisted Text Annotation," in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, (Avignon, France), pp. 102–107, Association for Computational Linguistics, Apr. 2012.

[11] S. M. Yimam, I. Gurevych, R. Eckart de Castilho, and C. Biemann, "WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (Sofia, Bulgaria), pp. 1–6, Association for Computational Linguistics, Aug. 2013.

[12] K. Bontcheva, L. Derczynski, and I. Roberts, "Crowdsourcing Named Entity Recognition and Entity Linking Corpora," in *Handbook of Linguistic Annotation* (N. Ide and J. Pustejovsky, eds.), pp. 875–892, Dordrecht: Springer Netherlands, 2017.

[13] N. Lawson, K. Eustice, M. Perkowitz, and M. Yetisgen-Yildiz, "Annotating large email datasets for named entity recognition with Mechanical Turk," p. 9, 2010.

[14] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, "Annotating Named Entities in Twitter Data with Crowdsourcing," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, (Los Angeles), pp. 80–88, Association for Computational Linguistics, June 2010.

[15] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg, "Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey," *Monthly Notices of the Royal Astronomical Society*, vol. 389, pp. 1179–1189, Sept. 2008.

[16] A. Kawrykow, G. Roumanis, A. Kam, D. Kwak, C. Leung, C. Wu, E. Zarour, P. Players, L. Sarmenta, M. Blanchette, and J. Waldispühl, "Phylo: A Citizen Science Approach for Improving Multiple Sequence Alignment," *PLOS ONE*, vol. 7, p. e31362, Mar. 2012.

[17] T. Chklovski, "1001 Paraphrases: Incenting Responsible Contributions in Collecting Paraphrases from Volunteers," p. 5, 2005.

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

[19] M. Poesio, J. Chamberlain, U. Kruschwitz, L. Robaldo, and L. Ducceschi, "Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation," *ACM TiiS*, vol. 3, no. 1, p. 3, 2013.

[20] B. Hladká, J. Mírovský, and J. Kohout, "An attractive game with the document: (im)possible?," *The Prague Bulletin of Mathematical Linguistics*, vol. 96, Jan. 2011.

[21] N. Green, P. Breimyer, V. Kumar, and N. Samatova, "PackPlay: Mining Semantic Data in Collaborative Games," in *Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010,*, (Uppsala, Sweden), pp. 227–234, Association for Computational Linguistics, 2010.

[22] N. Seemakurty, J. Chu, L. Von Ahn, and A. Tomasic, "Word sense disambiguation via human computation," in *Proceedings of the acm sigkdd workshop on human computation*, pp. 60–63, ACM, 2010.

[23] O. Chrons and S. Sundell, "Digitalkoot: Making Old Archives Accessible Using Crowdsourcing," in *Human Computation: Papers from the 2011 AAAI Workshop*, p. 6, Association for the Advancement of Artificial Intelligence, 2011.

[24] N. Venhuizen, V. Basile, K. Evang, and J. Bos, "Gamification for word sense labeling," in *Proc. 10th International Conference on Computational Semantics (IWCS-2013)*, pp. 397–403, 2013.

[25] A. Dumitrache, L. Aroyo, C. Welty, R.-J. Sips, and A. Levas, "'Dr. Detective": combining gamification techniques and crowdsourcing to create a gold standard in medical text," p. 16, 2013.

[26] D. Vannella, D. Jurgens, D. Scarfini, D. Toscani, and R. Navigli, "Validating and extending semantic knowledge bases using video games with a purpose.," in *ACL (1)*, pp. 1294–1304, 2014.

[27] D. Jurgens and R. Navigli, "It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation," *TACL*, vol. 2, pp. 449–464, 2014.

[28] A. Guha, M. Iyyer, D. Bouman, and J. Boyd-Graber, "Removing the Training Wheels: A Coreference Dataset that Entertains Humans and Challenges Computers," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Denver, Colorado), pp. 1108–1118, Association for Computational Linguistics, 2015.

[29] A. Scharl and M. Fols, "uComp Language Quiz - A Game with a Purpose for Multilingual Language Resource Acquisition," in *In Proceedings of Games4NLP: Using Games and Gamification for Natural Language Processing.*, (Valencia, Spain), p. 2, 2017.

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

[31] I.-E. Parasca, A. L. Rauter, J. Roper, A. Rusinov, G. Bouchard, S. Riedel, and P. Stenetorp, "Defining Words with Words: Beyond the Distributional Hypothesis," in *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, (Berlin, Germany), pp. 122–126, Association for Computational Linguistics, Aug. 2016.

[32] G. Han, M. Menezes, and L. Halaseh, "Towards a Word Sheriff 2.0: Lessons learnt and the road ahead," in *In Proceedings of Games4NLP: Using Games and Gamification for Natural Language Processing.*, (Valencia, Spain), p. 2, 2017.

[33] I. Habernal, R. Hannemann, C. Pollak, C. Klamm, P. Pauli, and I. Gurevych, "Argotario: Computational Argumentation Meets Serious Games," *arXiv:1707.06002 [cs]*, July 2017.

[34] K. Fort, B. Guillaume, M. Constant, N. Lefèbvre, and Y.-A. Pilatte, ""Fingers in the Nose": Evaluating Speakers' Identification of Multi-Word Expressions Using a Slightly Gamified Crowdsourcing Platform," in *LAW-MWE-CxG 2018 - COLING 2018 Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions*, Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), (Santa Fe, United States), pp. 207 – 213, Aug. 2018.

[35] S. Paun, B. Carpenter, J. Chamberlain, D. Hovy, U. Kruschwitz, and M. Poesio, "Comparing Bayesian Models of Annotation," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 571–585, Dec. 2018.

[36] J. Schell, *The Art of Game Design: A book of lenses*.
CRC Press, 2014.

[37] W. Luton, *Free-to-play: Making money from games you give away*.
New Riders, 2013.

[38] D. Games, "A Dark Room." http://adarkroom.doublespeakgames.com/, 2013.

[39] M. Palmer, D. Gildea, and P. Kingsbury, "The proposition bank: An annotated corpus of semantic roles," *Computational linguistics*, vol. 31, no. 1, pp. 71–106, 2005.

[40] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "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*, pp. 254–263, Association for Computational Linguistics, 2008.

[41] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: defining gamification," in *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, pp. 9–15, ACM, 2011.

[42] T. Y. Lee, C. Dugan, W. Geyer, T. Ratchford, J. C. Rasmussen, N. S. Shami, and S. Lupushor, "Experiments on motivational feedback for crowdsourced workers.," in *ICWSM*, 2013.

[43] L. von Ahn and L. Dabbish, "Designing games with a purpose," *Communications of the ACM*, vol. 51, p. 57, Aug. 2008.

[44] ustwogames, "Monument valley in numbers."

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

[46] L. Von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 319–326, ACM, 2004.

[47] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, *et al.*, "Predicting protein structures with a multiplayer online game," *Nature*, vol. 466, no. 7307, pp. 756–760, 2010.

[48] K. Tuite, "GWAPs: Games with a Problem," in *In FDG '14*, (Ft. Lauderdale, FL, USA), p. 7, 2014.

[49] W. Mason and D. J. Watts, "Financial incentives and the performance of crowds," *ACM SigKDD Explorations Newsletter*, vol. 11, no. 2, pp. 100–108, 2010.

[50] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (Baltimore, Maryland), pp. 55–60, Association for Computational Linguistics, June 2014.

[51] K. Lee, L. He, M. Lewis, and L. Zettlemoyer, "End-to-end Neural Coreference Resolution," *arXiv:1707.07045 [cs]*, July 2017.

[52] D. Jurafsky, *Speech & language processing*. Pearson Education India, 2000.

[53] L. Karttunen, J.-P. Chanod, G. Grefenstette, and A. Schille, "Regular expressions for language engineering," *Natural Language Engineering*, vol. 2, pp. 305–328, Dec. 1996.

[54] K. R. Beesley and L. Karttunen, "Finite-State Morphology: Xerox Tools and Techniques," *CSLI, Stanford*, 2002.

[55] E. Charniak, G. Carroll, J. Adcock, A. Cassandra, Y. Gotoh, J. Katz, M. Littman, and J. McCann, "Taggers for parsers," *Artificial Intelligence*, vol. 85, no. 1-2, pp. 45–57, 1996.

[56] R. Watson, "Part-of-speech tagging models for parsing," in *Proc. of CLUK*, vol. 2006, 2006.

[57] E. Brill, "Some Advances in Transformation-Based Part of Speech Tagging," *arXiv:cmp-lg/9406010*, June 1994.

[58] E. Brill, "A Simple Rule-Based Part of Speech Tagger," in *Proceedings of the third conference on Applied natural language processing.*, (Trento, Italy), p. 6, Association for Computational Linguistics, 1992.

[59] F. Karlsson, A. Voutilainen, J. Heikkilae, and A. Anttila, *Constraint Grammar: a language-independent system for parsing unrestricted text*, vol. 4.
USA: Walter de Gruyter & Co., 1995.

[60] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, (Williamstown, MA, USA), pp. 282–289, ACM, 2001.

[61] A. McCallum, D. Freitag, and F. Pereira, "Maximum Entropy Markov Models for Information Extraction and Segmentation," in *Proceedings of the Seventeenth International Conference on Machine Learning*, (CA, USA), pp. 591–598, ACM, 2000.

[62] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "A Practical Part-of-Speech Tagger," in *Third Conference on Applied Natural Language Processing*, (Trento, Italy), pp. 133–140, Association for Computational Linguistics, Mar. 1992.

[63] H. Schmid, "Part-of-Speech Tagging with Neural Networks," *arXiv:cmp-lg/9410018*, Oct. 1994.

[64] P. Wang, Y. Qian, F. K. Soong, L. He, and H. Zhao, "Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Recurrent Neural Network," *arXiv:1510.06168 [cs]*, Oct. 2015.

[65] M. Poesio, S. Pradhan, M. Recasens, K. Rodriguez, and Y. Versley, "Annotated corpora and annotation tools," in *Anaphora Resolution: Algorithms, Resources and Applications* (M. Poesio, R. Stuckardt, and Y. Versley, eds.), ch. 4, Springer, 2016.

[66] O. Uryupina and R. Zanoli, "Preprocessing," in *Anaphora Resolution: Algorithms, Resources and Applications* (M. Poesio, R. Stuckardt, and Y. Versley, eds.), ch. 6, pp. 209–236, Springer, 2016.

[67] M. Poesio, "The MATE/GNOME Scheme for Anaphoric Annotation, Revisited," Jan. 2004.

[68] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang, "CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes," in *Joint Conference on EMNLP and CoNLL - Shared Task*, (Jeju Island, Korea), pp. 1–40, Association for Computational Linguistics, July 2012.

[69] O. Uryupina, R. Artstein, A. Bristot, F. Cavicchio, F. Delogu, K. J. Rodriguez, and M. Poesio, "Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU Corpus," *Natural Language Engineering*, p. 52, 2019.

[70] J. Chamberlain, M. Poesio, and U. Kruschwitz, "Phrase Detectives Corpus 1.0 Crowdsourced Anaphoric Coreference," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 2039–2046, ACL, 2016.

[71] V. Stoyanov, N. Gilbert, C. Cardie, and E. Riloff, "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*, pp. 656–664, Association for Computational Linguistics, 2009.

[72] K. Hacioglu, B. Douglas, and Y. Chen, "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*, pp. 379–386, Association for Computational Linguistics, 2005.

[73] D. Zhekova and S. Kübler, "Ubiu: A language-independent system for coreference resolution," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 96–99, Association for Computational Linguistics, 2010.

[74] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The stanford CoreNLP natural language processing toolkit," in *Proc. of the 52nd ACL (system demos)*, pp. 55–60, 2014.

[75] H. Peng, K.-W. Chang, and D. Roth, "A joint framework for coreference resolution and mention head detection.," in *CoNLL*, vol. 51, p. 12, 2015.

[76] K. Lee, L. He, and L. Zettlemoyer, "Higher-order Coreference Resolution with Coarse-to-fine Inference," *arXiv:1804.05392 [cs]*, Apr. 2018.

[77] Y. Kim, E. Riloff, and N. Gilbert, "The taming of reconcile as a biomedical coreference resolver," in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 89–93, Association for Computational Linguistics, 2011.

[78] A. Soraluze, O. Arregi, X. Arregi, K. Ceberio, and A. D. De Ilarraza, "Mention detection: First steps in the development of a basque coreference resolution system.," in *KONVENS*, pp. 128–136, 2012.

[79] J. K. Kummerfeld, M. Bansal, D. Burkett, and D. Klein, "Mention detection: heuristics for the ontonotes annotations," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 102–106, Association for Computational Linguistics, 2011.

[80] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky, "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*, pp. 28–34, Association for Computational Linguistics, 2011.

[81] P. Pakray, S. Neogi, P. Bhaskar, S. Poria, S. Bandyopadhyay, and A. Gelbukh, "A textual entailment system using anaphora resolution," in *System Report. Text analysis conference recognizing textual entailment track notebook*, 2011.

[82] J. Steinberger, M. Poesio, M. A. Kabadjov, and K. Ježek, "Two uses of anaphora resolution in summarization," *Information Processing & Management*, vol. 43, no. 6, pp. 1663–1680, 2007.

[83] R. Mitkov, R. Evans, C. Orăsan, V. Pekar, *et al.*, "Anaphora resolution: To what extent does it help nlp applications?," in *Discourse Anaphora and Anaphor Resolution Colloquium*, pp. 179–190, Springer, 2007.

[84] N. Nicolov, F. Salvetti, and S. Ivanova, "Sentiment analysis: Does coreference matter," in *AISB 2008 Convention Communication, Interaction and Social Intelligence*, vol. 1, p. 37, 2008.

[85] J. H. Martin and D. Jurafsky, "Speech and language processing," *International Edition*, 2000.

[86] V. Ng, "Advanced machine learning models for coreference resolution," in *Anaphora Resolution*, pp. 283–313, Springer, 2016.

[87] H. H. Clark, "Bridging," in *Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing*, TINLAP '75, (Stroudsburg, PA, USA), pp. 169–174, Association for Computational Linguistics, 1975.

[88] R. Gaizauskas, H. Cunningham, Y. Wilks, P. Rodgers, and K. Humphreys, "GATE: an environment to support research and development in natural language engineering," in *Proceedings Eighth IEEE International Conference on Tools with Artificial Intelligence*, pp. 58–66, IEEE, 1996.

[89] C. Mueller and M. Strube, "MMAX: A Tool for the Annotation of Multi-modal Corpora," in *In Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pp. 45–50, 2001.

[90] K. Bontcheva, H. Cunningham, I. Roberts, A. Roberts, V. Tablan, N. Aswani, and G. Gorrell, "GATE Teamware: a web-based, collaborative text annotation framework," *Language Resources and Evaluation*, vol. 47, pp. 1007–1029, Dec. 2013.

[91] K. Bontcheva, I. Roberts, L. Derczynski, and D. Rout, "The GATE Crowdsourcing Plugin: Crowdsourcing Annotated Corpora Made Easy," in *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, (Gothenburg, Sweden), pp. 97–100, Association for Computational Linguistics, Apr. 2014.

[92] H. Cunningham, "GATE, a General Architecture for Text Engineering," *Computers and the Humanities*, vol. 36, no. 2, pp. 223–254, 2002.

[93] C. Müller, "Representing and Accessing Multi-Level Annotations in MMAX2," in *Proceedings of the 5th Workshop on NLP and XML (NLPXML-2006): Multi-Dimensional Markup in Natural Language Processing*, 2006.

[94] P. Berck and A. Russel, "ANNEX - a web-based Framework for Exploiting Annotated Media Resources," in *LREC*, 2006.

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

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

[97] V. Basile, J. Bos, K. Evang, and N. Venhuizen, "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*, pp. 92–96, Association for Computational Linguistics, 2012.

[98] E. Law and L. V. Ahn, "Input-agreement: A New Mechanism for Collecting Data Using Human Computation Games," in *Proc. of CHI*, ACM Press, 2009.

[99] A. J. Quinn and B. B. Bederson, "Human computation: a survey and taxonomy of a growing field," in *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, (Vancouver, BC, Canada), p. 1403, ACM Press, 2011.

[100] Howe, "Crowdsourcing: A definition." [Online] http://www.crowdsourcing.com/cs/2006/06/crowdsourcing_a.html, 2006.

[101] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with Mechanical Turk," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 453–456, ACM, 2008.

[102] M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl, "Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, p. 9, 2014.

[103] W. Mason and D. J. Watts, "Financial Incentives and the "Performance of Crowds"," in *Proceedings of the ACM SIGKDD workshop on human computation*, (Paris, France), p. 9, 2009.

[104] A. Aker, M. El-haj, M.-d. Albakour, and U. Kruschwitz, "Assessing crowdsourcing quality through objective tasks," in *In Proceedings of LREC'12*, 2012.

[105] F. Laws, C. Scheible, and H. Schütze, "Active Learning with Amazon Mechanical Turk," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, (Edinburgh, Scotland, UK.), pp. 1546–1556, Association for Computational Linguistics, July 2011.

[106] T. Gillier, C. Chaffois, M. Belkhouja, Y. Roth, and B. L. Bayus, "The effects of task instructions in crowdsourcing innovative ideas," *Technological Forecasting and Social Change*, vol. 134, pp. 35–44, 2018.
Publisher: Elsevier.

[107] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy, "Learning Whom to Trust with MACE," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Atlanta, Georgia), pp. 1120–1130, Association for Computational Linguistics, June 2013.

[108] A. P. Dawid and A. M. Skene, "Maximum Likelihood Estimation of observer error-rates using the EM algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 20–28, 1979.

[109] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, vol. 11, no. Apr, pp. 1297–1322, 2010.

[110] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng, "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*, (Honolulu, Hawaii), pp. 254–263, Association for Computational Linguistics, Oct. 2008.

[111] F. Rodrigues, F. Pereira, and B. Ribeiro, "Sequence labeling with multiple annotators," *Machine Learning*, vol. 95, pp. 165–181, May 2014.

[112] A. T. Nguyen, B. C. Wallace, J. J. Li, A. Nenkova, and M. Lease, "Aggregating and Predicting Sequence Labels from Crowd Annotations," *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2017, pp. 299–309, 2017.

[113] R. C. Jordan, S. A. Gray, D. V. Howe, W. R. Brooks, and J. G. Ehrenfeld, "Knowledge Gain and Behavioral Change in Citizen-Science Programs: Citizen-Scientist Knowledge Gain," *Conservation Biology*, vol. 25, pp. 1148–1154, Dec. 2011.

[114] A. Eveleigh, C. Jennett, S. Lynn, and A. L. Cox, ""I want to be a Captain! I want to be a Captain!": Gamification in the Old Weather Citizen Science Project," in *In Proceedings of Gamification '13*, (Ontario, Canada), p. 4, ACM, 2013.

[115] R. Tinati, M. Luczak-Roesch, E. Simperl, and W. Hall, "An investigation of player motivations in Eyewire, a gamified citizen science project," *Computers in Human Behavior*, vol. 73, pp. 527–540, Aug. 2017.

[116] V. Curtis, "Motivation to Participate in an Online Citizen Science Game: A Study of Foldit," *Science Communication*, vol. 37, pp. 723–746, Dec. 2015.

[117] I. Iacovides, C. Jennett, C. Cornish-Trestrail, and A. L. Cox, "Do games attract or sustain engagement in citizen science?: a study of volunteer motivations," in *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13*, (Paris, France), p. 1101, ACM Press, 2013.

[118] C. Jennett, D. J. Furniss, I. Iacovides, S. Wiseman, S. J. J. Gould, and A. L. Cox, "Exploring Citizen Psych-Science and the Motivations of Errordiary Volunteers," *Human Computation*, vol. 1, Dec. 2014.

[119] M. J. Raddick, G. Bracey, P. L. Gay, C. J. Lintott, P. Murray, K. Schawinski, A. S. Szalay, and J. Vandenberg, "Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers," *Astronomy Education Review*, vol. 9, Dec. 2010.
arXiv: 0909.2925.

[120] M. J. Raddick, G. Bracey, P. L. Gay, C. J. Lintott, C. Cardamone, P. Murray, K. Schawinski, A. S. Szalay, and J. Vandenberg, "Galaxy Zoo: Motivations of Citizen Scientists," *arXiv:1303.6886 [astro-ph, physics:physics]*, Mar. 2013.

arXiv: 1303.6886.

[121] K. Crowston and N. R. Prestopnik, "Motivation and Data Quality in a Citizen Science Game: A Design Science Evaluation," in *2013 46th Hawaii International Conference on System Sciences*, (Wailea, HI, USA), pp. 450–459, IEEE, Jan. 2013.

[122] J. Cox, E. Y. Oh, B. Simmons, C. Lintott, K. Masters, A. Greenhill, G. Graham, and K. Holmes, "Defining and Measuring Success in Online Citizen Science: A Case Study of Zooniverse Projects," *Computing in Science Engineering*, vol. 17, pp. 28–41, July 2015.

[123] G. Tsueng, S. M. Nanis, J. Fouquier, B. M. Good, and A. I. Su, "Citizen Science for Mining the Biomedical Literature," *Citizen science : theory and practice*, vol. 1, no. 2, 2016.

[124] R. Simpson, K. R. Page, and D. De Roure, "Zooniverse: observing the world's largest citizen science platform," in *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*, (Seoul, Korea), pp. 1049–1054, ACM Press, 2014.

[125] C. Lintott, K. Schawinski, S. Bamford, A. Slosar, K. Land, D. Thomas, E. Edmondson, K. Masters, R. C. Nichol, M. J. Raddick, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg, "Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies: Galaxy Zoo," *Monthly Notices of the Royal Astronomical Society*, vol. 410, pp. 166–178, Jan. 2011.

[126] K. W. Willett, C. J. Lintott, S. P. Bamford, K. L. Masters, B. D. Simmons, K. R. V. Casteels, E. M. Edmondson, L. F. Fortson, S. Kaviraj, W. C. Keel, T. Melvin, R. C. Nichol, M. J. Raddick, K. Schawinski, R. J. Simpson, R. A. Skibba, A. M. Smith, and D. Thomas, "Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey," *Monthly Notices of the Royal Astronomical Society*, vol. 435, pp. 2835–2860, Nov. 2013.

[127] "Old Weather." https://www.oldweather.org/, 2010.

[128] Philip, "The four million." https://blog.oldweather.org/2016/12/11/the-four-million/, Dec. 2016.

[129] Philip, "There's a green one and a pink one and a blue one and a yellow one." https://blog.oldweather.org/2012/09/05/theres-a-green-one-and-a-pink-one-and-a-blue-one-and-a-yellow-one/, Sept. 2012.

[130] R. I. Doğan, R. Leaman, and Z. Lu, "NCBI disease corpus: a resource for disease name recognition and concept normalization," *Journal of biomedical informatics*, vol. 47, pp. 1–10, 2014.
Publisher: Elsevier.

[131] L. von Ahn, "Games with a Purpose," *Computer*, vol. 39, pp. 92–94, June 2006.

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

[133] L. von Ahn, R. Liu, and M. Blum, "Peekaboom: a game for locating objects in images," in *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*, (Montreal, Quebec, Canada), p. 55, ACM Press, 2006.

[134] L. von Ahn, S. Ginosar, M. Kedia, and M. Blum, "Improving Image Search with PHETCH," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, (Honolulu, HI), pp. IV–1209–IV–1212, IEEE, Apr. 2007.

[135] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, and F. players, "Predicting protein structures with a multiplayer online game," *Nature*, vol. 466, pp. 756–760, Aug. 2010.

[136] J. Lee, W. Kladwang, M. Lee, D. Cantu, M. Azizyan, H. Kim, A. Limpaecher, S. Gaikwad, S. Yoon, A. Treuille, R. Das, and E. Participants, "RNA design rules from a massive open laboratory," *Proceedings of the National Academy of Sciences*, vol. 111, pp. 2122–2127, Feb. 2014.

[137] M. Lafourcade, "Making people play for Lexical Acquisition with the JeuxDeMots prototype," in *In Proceedings SNLP'07: 7th International Symposium on Natural Language Processing*, (Bangok, Thailand), p. 8, 2007.

[138] U. Gadiraju, B. Fetahu, and R. Kawase, "Training Workers for Improving Performance in Crowdsourcing Microtasks," in *Design for Teaching and Learning in a Networked World* (G. Conole, T. Klobučar, C. Rensing, J. Konert, and E. Lavoué, eds.), vol. 9307, pp. 100–114, Cham: Springer International Publishing, 2015.

[139] J. Hamari, D. J. Shernoff, E. Rowe, B. Coller, J. Asbell-Clarke, and T. Edwards, "Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning," *Computers in Human Behavior*, vol. 54, pp. 170–179, 2016.

[140] R. Koster, *Theory of fun for game design*.
" O'Reilly Media, Inc.", 2013.

[141] T. W. Malone, "Toward a theory of intrinsically motivating instruction," *Cognitive science*, vol. 5, no. 4, pp. 333–369, 1981.

[142] D. Cook, "The Princess Rescuing Application: Slides." http://www.lostgarden.com/2008/10/princess-rescuing-application-slides.html, 2008.

[143] A. Harmetz, "Sigh of Relief on Video Games," *The New York Times*, Jan. 1984.

[144] J. Funk, "How Much Did Modern Warfare 2 Cost to Make? | The Escapist." http://www.escapistmagazine.com/news/view/96227-How-Much-Did-Modern-Warfare-2-Cost-to-Make, 2009.

[145] E. Makuch, "Dead Space 2 Dev Says Game Cost $60M To Make And Sold Millions, But Failed Commercially." https://www.gamespot.com/articles/dead-space-2-dev-says-game-cost-60m-to-make-and-so/1100-6454150/, 2017.

[146] C. Yerli, "GC 2008: Crysis Cost 22 Million to Make." https://www.ign.com/articles/2008/08/19/gc-2008-crysis-cost-22-million-to-make, 2008.

[147] R. Grover and M. Nayak, "Activision plans $500 million date with 'Destiny'," *Reuters*, May 2014.

[148] L. Von Ahn, R. Liu, and M. Blum, "Peekaboom: a game for locating objects in images," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pp. 55–64, ACM, 2006.

[149] M. Csikszentmihalyi, "Flow: The psychology of optimal experience," *New York: Harper Collins*, 1990.

[150] F. Khatib, F. DiMaio, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popović, *et al.*, "Crystal structure of a monomeric retroviral protease solved by protein folding game players," *Nature structural & molecular biology*, vol. 18, no. 10, pp. 1175–1177, 2011.

[151] M. Lafourcade, "Making people play for lexical acquisition with the jeuxdemots prototype," in *SNLP'07: 7th international symposium on natural language processing*, p. 7, 2007.

[152] M. Lafourcade, A. Joubert, and N. Le Brun, *Games With A Purpose (GWAPs)*. Wiley, 2015.

[153] B. Hladká, J. Mírovskỳ, and P. Schlesinger, "Play the language: Play coreference," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 209–212, Association for Computational Linguistics, 2009.

[154] B. Hladká, J. Mírovský, and P. Schlesinger, "Designing a Language Game for Collecting Coreference Annotation," in *Proceedings of the Third Linguistic Annotation Workshop*, (Suntec, Singapore), pp. 52–55, Association for Computational Linguistics, Aug. 2009.

[155] D. Dziedzic, "Use of the free to play model in games with a purpose: the robocorp game case study," *Bio-Algorithms and Med-Systems*, vol. 12, no. 4, pp. 187–197, 2016.

[156] B. Guillaume, K. Fort, and N. Lefèbvre, "Crowdsourcing Complex Language Resources: Playing to Annotate Dependency Syntax," p. 13, 2017.

[157] N. Otani, D. Kawahara, S. Kurohashi, N. Kaji, and M. Sassano, "Large-Scale Acquisition of Commonsense Knowledge via a Quiz Game on a Dialogue System," in *Proceedings of the Open Knowledge Base and Question Answering Workshop (OKBQA 2016)*, (Osaka, Japan), pp. 11–20, The COLING 2016 Organizing Committee, Dec. 2016.

[158] J. Chamberlain, "The annotation-validation (AV) model: rewarding contribution using retrospective agreement," in *Proceedings of the First International Workshop on Gamification for Information Retrieval - GamifIR '14*, (Amsterdam, The Netherlands), pp. 12–16, ACM Press, 2014.

[159] G. H. Mc Laughlin, "SMOG Grading-a New Readability Formula," *Journal of Reading*, vol. 12, no. 8, pp. 639–646, 1969.

[160] Y.-l. Kuo, J.-C. Lee, K.-y. Chiang, R. Wang, E. Shen, C.-w. Chan, and J. Y.-j. Hsu, "Community-based game design: experiments on social games for commonsense data collection," in *Proceedings of the acm sigkdd workshop on human computation*, pp. 15–22, ACM, 2009.

[161] R. Burkett, "An Alternative Framework for Agent Recruitment: From MICE to RASCLS," vol. 57, no. 1, p. 12, 2013.

[162] K. Fort, B. Guillaume, and N. Lefèbvre, "Who wants to play Zombie? A survey of the players on ZOMBILINGO," in *In Proceedings of Games4NLP: Using Games and Gamification for Natural Language Processing.*, (Valencia, Spain), p. 3, 2017.

[163] L. von Ahn, M. Kedia, and M. Blum, "Verbosity: a game for collecting common-sense facts," in *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*, (Montréal, Québec, Canada), p. 75, ACM Press, 2006.

[164] S. Petrov, L. Barrett, R. Thibaux, and D. Klein, "Learning Accurate, Compact, and Interpretable Tree Annotation," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, (Sydney, Australia), pp. 433–440, Association for Computational Linguistics, July 2006.

[165] M. Poesio, J. Chamberlain, and U. Kruschwitz, "Phrase Detectives," in *Handbook of Linguistic Annotation* (N. Ide and J. Pustejovsky, eds.), pp. 1149–1176, Dordrecht: Springer Netherlands, 2017.

[166] E. Loper and S. Bird, "NLTK: The Natural Language Toolkit," *arXiv:cs/0205028*, May 2002.

[167] D. A. Ferrucci and A. Lally, "UIMA: an architectural approach to unstructured information processing in the corporate research environment," *Natural Language Engineering*, vol. 10, pp. 327–348, 2004.

[168] V. Basile, J. Bos, K. Evang, and N. Venhuizen, "Developing a large semantically annotated corpus," in *In Proceedings Eighth International Conference on Language Resources and Evaluation*, (Istanbul, Turkey), p. 6, ELRA, 2012.

[169] N. Venhuizen, K. Evang, V. Basile, and J. Bos, "Gamification for Word Sense Labeling," in *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Short Papers*, p. 7, ACL, 2013.

[170] L. Von Ahn and L. Dabbish, "Designing games with a purpose," *Communications of the ACM*, vol. 51, no. 8, pp. 58–67, 2008.

[171] M. Oltmans, "Which Metrics Should You Measure According To The Pirate Analytics (Mobile/F2P Games)." https://www.gamasutra.com/blogs/MarvinOltmans/20160525/273487/Which_Metrics_Should_You_Measure_According_To_The_Pirate_Analytics_MobileF2P_Games.php, 2016.

[172] D. Xicota, "Free to play and its Key Performance Indicators." http://www.gamasutra.com/blogs/DavidXicota/20140527/218550/Free_to_play_and_its_Key_Performance_Indicators.php, 2014.

[173] R. Weber, "The Future of Mobile User Acquisition and Monetization." http://www.gamasutra.com/blogs/RobertWeber/20131106/204210/The_Future_of_Mobile_User_Acquisition_and_Monetization.php, 2013.

[174] "Glossary of Metrics." https://unity3d.com/learn/tutorials/topics/analytics/glossary-metrics, 2014.

[175] T. McCalmont, "15 Metrics All Game Developers Should Know By Heart." https://www.gamasutra.com/blogs/TrevorMcCalmont/20150709/248118/15_Metrics_All_Game_Developers_Should_Know_By_Heart.php, 2015.

[176] D. C. Schmittlein, D. G. Morrison, and R. Colombo, "Counting Your Customers: Who Are They and What Will They Do Next?," *Management Science*, vol. 33, no. 1, pp. 1–24, 1987.

[177] J.-R. Gauthier, "An Introduction to Predictive Customer Lifetime Value Modeling." https://www.datascience.com/blog/intro-to-predictive-modeling-for-customer-lifetime-value, 2017.

[178] J. Nielsen and T. K. Landauer, "A mathematical model of the finding of usability problems," in *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pp. 206–213, 1993.

[179] J. Brooke *et al.*, "SUS-A quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[180] T. W. Malone, "Toward a Theory of Intrinsically Motivating Instruction*," *Cognitive Science*, vol. 5, pp. 333–369, Oct. 1981.

[181] J. Tanenbaum and J. Bizzocchi, "Rock Band: a case study in the design of embodied interface experience," in *Proceedings of the 2009 ACM SIGGRAPH Symposium on Video Games - Sandbox '09*, (New Orleans, Louisiana), p. 127, ACM Press, 2009.

[182] B. Santorini, "Part-of-speech tagging guidelines for the Penn Treebank Project," *Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania*, 1990.

[183] J. Nivre, M.-C. De Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, and others, "Universal dependencies v1: A multilingual treebank collection," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1659–1666, 2016.

[184] R. J. Passonneau and B. Carpenter, "The benefits of a model of annotation," *Transactions of the ACL*, vol. 2, pp. 311–326, 2014.

[185] S. Paun, B. Carpenter, J. Chamberlain, D. Hovy, U. Kruschwitz, and M. Poesio, "Bayesian annotation methods for NLP: An evaluation.," *Transactions of the ACL*, vol. 6, pp. 571–585, 2018.

[186] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural Architectures for Named Entity Recognition," *arXiv:1603.01360 [cs]*, Mar. 2016.

[187] T. Dozat and C. D. Manning, "Deep Biaffine Attention for Neural Dependency Parsing," *arXiv:1611.01734 [cs]*, Nov. 2016.

[188] B. Bohnet and J. Nivre, "A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, (Stroudsburg, PA, USA), pp. 1455–1465, Association for Computational Linguistics, 2012.

[189] J. Yu, B. Bohnet, and M. Poesio, "Neural Mention Detection," *arXiv:1907.12524 [cs]*, July 2019.

arXiv: 1907.12524.

[190] M. Marcus, "Building a Large Annotated Corpus of English: The Penn Treebank:," tech. rep., Defense Technical Information Center, Fort Belvoir, VA, Apr. 1993.

[191] K. Hara, A. Adams, K. Milland, S. Savage, C. Callison-Burch, and J. P. Bigham, "A data-driven analysis of workers' earnings on Amazon Mechanical Turk," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2018.

[192] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of english: The penn treebank," *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[193] J. Bos, V. Basile, K. Evang, N. Venhuizen, and J. Bjerva, "The groningen meaning bank," in *Handbook of Linguistic Annotation* (N. Ide and J. Pustejovsky, eds.), vol. 2, pp. 463–496, Springer, 2017.

[194] A. Sheshadri and M. Lease, "SQUARE: A Benchmark for Research on Computing Crowd Consensus," in *First AAAI conference on Human Computation and Crowdsourcing*, (California, USA), p. 9, AAAI, 2013.

[195] M. Dredze, P. P. Talukdar, and K. Crammer, "Sequence Learning from Data with Multiple Labels," in *In Proceedings of the 2nd International Workshop on Learning from Multilabel Data (MLD)*, (Haifa, Isreal), p. 10, 2009.

[196] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor, "Are Your Participants Gaming the System? Screening Mechanical Turk Workers," in *Proceedings of the SIGCHI conference on human factors in computing systems*, (Atlanta, USA), p. 4, ACM, 2010.

[197] R. J. Passonneau and B. Carpenter, "The benefits of a model of annotation," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 311–326, 2014.

[198] B. Carpenter, *Multilevel bayesian models of categorical data annotation. Available at http://lingpipe-blog.com/ lingpipe-white-papers*.
2008.

[199] Y. Bachrach, T. Minka, J. Guiver, and T. Graepel, "How To Grade a Test Without Knowing the Answers — A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing," p. 8, 2012.

[200] Sooyoung Lee, Sehwa Park, and Seog Park, "A quality enhancement of crowdsourcing based on quality evaluation and user-level task assignment framework," in *2014 International Conference on Big Data and Smart Computing (BIGCOMP)*, (Bangkok, Thailand), pp. 60–65, IEEE, Jan. 2014.

[201] S. Basu Roy, I. Lykourentzou, S. Thirumuruganathan, S. Amer-Yahia, and G. Das, "Task assignment optimization in knowledge-intensive crowdsourcing," *The VLDB Journal*, vol. 24, pp. 467–491, Aug. 2015.

[202] M. Hossain, "Crowdsourcing: Activities, incentives and users' motivations to participate," in *2012 International Conference on Innovation Management and Technology Research*, (Malacca, Malaysia), pp. 501–506, IEEE, May 2012.

[203] N. Kaufmann, T. Schulze, and D. Veit, "More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk," in *Proceedings of the Seventeenth Americas Conference on Information Systems*, (Detroit, Michigan, USA), p. 12, 2011.

[204] C.-Y. Hung, J. C.-Y. Sun, and P.-T. Yu, "The benefits of a challenge: student motivation and flow experience in tablet-PC-game-based learning," *Interactive Learning Environments*, vol. 23, pp. 172–190, Mar. 2015.

[205] J. M. Carroll and J. M. Thomas, "FUN," *ACM SIGCHI Bulletin*, vol. 19, pp. 21–24, Jan. 1988.

[206] P. Sweetser and P. Wyeth, "Gameflow: a model for evaluating player enjoyment in games," *Computers in Entertainment (CIE)*, vol. 3, no. 3, pp. 3–3, 2005.

[207] A. Kittur, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton, "The future of crowd work," in *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*, (San Antonio, Texas, USA), p. 1301, ACM Press, 2013.

[208] P. G. Ipeirotis and E. Gabrilovich, "Quizz: targeted crowdsourcing with a billion (potential) users," in *Proceedings of the 23rd international conference on World wide web - WWW '14*, (Seoul, Korea), pp. 143–154, ACM Press, 2014.

[209] K. Fort, B. Guillaume, and H. Chastant, "Creating zombilingo, a game with a purpose for dependency syntax annotation," in *Proceedings of the First International Workshop on Gamification for Information Retrieval*, pp. 2–6, ACM, 2014.

[210] D. Boutros, "Difficulty is Difficult: Designing for Hard Modes in Games." https://www.gamasutra.com/view/feature/132181/difficulty_is_difficult_designing_.php, 2008.

[211] E. Adams, "The Designer's Notebook: Difficulty Modes and Dynamic Difficulty Adjustment." https://www.gamasutra.com/view/feature/132061/the_designers_notebook_.php, 2008.

[212] R. Hunicke, "The case for dynamic difficulty adjustment in games," in *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*, pp. 429–433, ACM, 2005.

[213] S. Chen, "The Social Network Game Boom." http://www.gamasutra.com/view/feature/132400/the_social_network_game_boom.php, 2009.

[214] Z. Kleinman, "Free games attract high prices." http://news.bbc.co.uk/1/hi/technology/8376392.stm, Nov. 2009.

[215] B. Kirman, "Emergence and playfulness in social games," in *Proceedings of the 14th International Academic MindTrek Conference on Envisioning Future Media Environments - MindTrek '10*, (Tampere, Finland), p. 71, ACM Press, 2010.

[216] C. Lewis, N. Wardrip-Fruin, and J. Whitehead, "Motivational game design patterns of 'Ville games," in *Proceedings of the International Conference on the Foundations of Digital Games - FDG '12*, (Raleigh, North Carolina), p. 172, ACM Press, 2012.

[217] J. Paavilainen, J. Hamari, J. Stenros, and J. Kinnunen, "Social Network Games: Players' Perspectives," *Simulation & Gaming*, vol. 44, pp. 794–820, Dec. 2013.

[218] B. Purkiss and I. Khaliq, "A study of interaction in idle games & perceptions on the definition of a game," in *2015 IEEE Games Entertainment Media Conference (GEM)*, (Toronto, ON), pp. 1–6, IEEE, Oct. 2015.

[219] K. Alha, E. Koskinen, J. Paavilainen, and J. Hamari, "Free-to-Play Games: Professionals' Perspectives," in *Proceedings of nordic DiGRA 2014*, p. 14, DiGRA, 2014.

[220] P. Luban, "The Design of Free-To-Play Games: Part 1." https://www.gamasutra.com/view/feature/134920/the_design_of_freetoplay_games_.php, 2011.

[221] D. M. Boyd and N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[222] D. Gross, "The Facebook games that millions love (and hate) - CNN.com." http://www.cnn.com/2010/TECH/02/23/facebook.games/index.html, 2010.

[223] D. Helgason, "Thoughts On Browser Plugin Penetration – Unity Blog." https://blogs.unity3d.com/2008/03/31/thoughts-on-browser-plugin-penetration/, 2008.

[224] I. Bogost, "Cow clicker: The making of obsession (2010)." http://bogost.com/writing/blog/cow_clicker_1/, 2010.

[225] J. Tanz, "The Curse of Cow Clicker: How a Cheeky Satire Became a Videogame Hit," *Wired*, vol. 20, Dec. 2011.

[226] S. A. Alharthi, O. Alsaedi, Z. O. Toups, J. Tanenbaum, and J. Hammer, "Playing to Wait: A Taxonomy of Idle Games," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, (Montreal QC, Canada), pp. 1–15, ACM Press, 2018.

[227] D. Dziedzic and W. Włodarczyk, "Making NLP games with a purpose fun to play using Free to Play mechanics: RoboCorp case study," p. 2, 2016.

[228] J. Chamberlain, "Harnessing Collective Intelligence on Social Networks," p. 234, 2014.

[229] I. Team, "How to Identify Whales In Your Game." https://gameanalytics.com/blog/how-to-identify-whales-in-your-game.html, Sept. 2015.

[230] P. Tassi, "Why It's Scary When 0.15% Mobile Gamers Bring In 50% Of The Revenue." https://www.forbes.com/sites/insertcoin/2014/03/01/why-its-scary-when-0-15-mobile-gamers-bring-in-50-of-the-revenue/, 2014.

[231] D.-H. Shin and Y.-J. Shin, "Why do people play social network games?," *Computers in Human Behavior*, vol. 27, pp. 852–861, Mar. 2011.

[232] C. Aldrich, *Learning by doing: A comprehensive guide to simulations, computer games, and pedagogy in e-learning and other educational experiences*.
John Wiley & Sons, 2005.

[233] R. Kiberd, "Cookie Clicker, the Internet's most pointlessly addictive game, is also its most subversive." https://kernelmag.dailydot.com/issue-sections/staff-editorials/15694/cookie-clicker-capitalist-dystopia/, Jan. 2016.

[234] C. R. Heil, J. S. Wu, J. J. Lee, and T. Schmidt, "A Review of Mobile Language Learning Applications: Trends, Challenges, and Opportunities," *The EuroCALL Review*, vol. 24, p. 32, Sept. 2016.

[235] B. Settles and B. Meeder, "A Trainable Spaced Repetition Model for Language Learning," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Berlin, Germany), pp. 1848–1858, Association for Computational Linguistics, 2016.

[236] C. H. Karjo and W. Andreani, "Learning Foreign Languages With Duolingo and Memrise," in *Proceedings of the 2018 International Conference on Distance Education and Learning - ICDEL '18*, (Beijing, China), pp. 109–112, ACM Press, 2018.

[237] R. E. Mayer, "Rote Versus Meaningful Learning," *Theory Into Practice*, vol. 41, pp. 226–232, Nov. 2002.

[238] R. Thaler, "Toward a positive theory of consumer choice," *Journal of Economic Behavior & Organization*, vol. 1, pp. 39–60, Mar. 1980.

[239] A. Pecorella, "The Math of Idle Games, Part I." https://blog.kongregate.com/the-math-of-idle-games-part-i/, Oct. 2016.

[240] J. P. Zagal, S. Björk, and C. Lewis, "Dark Patterns in the Design of Games," in *Foundations of Digital Games 2013*, (Crete, Greece), p. 8, 2013.

[241] B. Morrison, "A Necessary Evil: Grinding in Games." https://www.gamasutra.com/blogs/BriceMorrison/20110211/88931/A_Necessary_Evil_Grinding_in_Games.php, 2011.

[242] G. Costikyan, "Ethical Free-to-Play Game Design (And Why it Matters)." https://www.gamasutra.com/view/feature/207779/ethical_freetoplay_game_design_.php, 2014.

[243] IVC, "Wait Wait... Don't Play Me: The Clicker Game Genre and Configuring Everyday Temporalities – InVisible Culture." http://ivc.lib.rochester.edu/wait-wait-dont-play-me-the-clicker-game-genre-and-configuring-everyday-temporalities/, 2019.

[244] I. Bogost, "Persuasive Games: Exploitationware." https://www.gamasutra.com/view/feature/134735/persuasive_games_exploitationware.php, 2011.

[245] K. Graft, "Wargaming kicks 'pay-to-win' monetization to the curb." /view/news/193520/Wargaming_kicks_paytowin_monetization_to_the_curb.php, 2013.

[246] Fragsworth, "Clicker Heroes 2." http://www.clickerheroes2.com/paytowin.php, 2017.

[247] D. Zendle and P. Cairns, "Video game loot boxes are linked to problem gambling: Results of a large-scale survey," *PLOS ONE*, vol. 13, p. e0206767, Nov. 2018.

[248] T. Hoggins, "Video game loot boxes to be investigated by US after being blamed for rise in young gamblers," *The Telegraph*, Nov. 2018.

[249] S. Gainsbury, D. King, B. Abarbanel, P. Delfabbro, and N. Hing, "Convergence of gambling and gaming in digital media," *Melbourne, VIC: Victorian Responsible Gambling Foundation*, 2015.

[250] J. Macey and J. Hamari, "Investigating relationships between video gaming, spectating esports, and gambling," *Computers in Human Behavior*, vol. 80, pp. 344–353, Mar. 2018.

[251] M. Rose, "Chasing the Whale: Examining the ethics of free-to-play games." https://www.gamasutra.com/view/feature/195806/chasing_the_whale_examining_the_.php, 2013.

[252] A. King, "Numbers Getting Bigger: What Are Incremental Games, and Why Are They Fun?." https://gamedevelopment.tutsplus.com/articles/numbers-getting-bigger-what-are-incremental-games-and-why-are-they-fun--cms-23925, 2015.

[253] A. Zeldes, "The GUM corpus: creating multilayer resources in the classroom," *Language Resources and Evaluation*, vol. 51, pp. 581–612, Sept. 2017.