Phrase Detectives Corpus 1.0
Crowdsourced Anaphoric Coreference

Jon Chamberlain, Massimo Poesio, Udo Kruschwitz
School of Computer Science and Electronic Engineering
University of Essex
{jchamb,poesio,udo}@essex.ac.uk

Abstract

Natural Language Engineering tasks require large and complex annotated datasets to build more advanced models of language. Corpora are typically annotated by several experts to create a gold standard; however, there are now compelling reasons to use a non-expert crowd to annotate text, driven by cost, speed and scalability. Phrase Detectives Corpus 1.0 is an anaphorically-annotated corpus of encyclopedic and narrative text that contains a gold standard created by multiple experts, as well as a set of annotations created by a large non-expert crowd. Analysis shows very good inter-expert agreement ($\kappa = .88 - .93$) but a more variable baseline crowd agreement ($\kappa = .52 - .96$). Encyclopedic texts show less agreement (and by implication are harder to annotate) than narrative texts. The release of this corpus is intended to encourage research into the use of crowds for text annotation and the development of more advanced, probabilistic language models, in particular for anaphoric coreference.

Keywords: anaphoric coreference, anaphora, annotation, crowdsourcing, gwap, games-with-a-purpose, phrase detectives, corpora

1. Introduction

A revolution in the way tasks can be completed occurred when it was proposed to take a job traditionally performed by a designated employee and outsource it to an undefined large group of Internet users. This approach, called **crowdsourcing** (Howe, 2008), changed traditional thinking behind methods for language resource creation and made new tasks possible that were previously inconceivable due to cost or labour limitations. For example, **microworking** is now a common crowdsourcing approach to create small to medium-sized language resources by engaging a crowd using small payments (Snow et al., 2008). An alternative approach is to use a **game-with-a-purpose (GWAP)** to aggregate data from non-expert players, who are motivated by entertainment, to create collective decisions similar to those from an expert (von Ahn, 2006).

Phrase Detectives, an interactive online game for creating anaphorically-annotated corpora, is an illustration of the GWAP approach for creating large-scale resources. The Phrase Detectives corpus differs from existing corpora for anaphora in two key respects: (i) it covers genres for which no other data are available, including encyclopedic and narrative text; and (ii) multiple solutions (or interpretations) are collected per task. This paper briefly describes the game and annotation scheme, before describing in more detail the measures of quality used and an analysis of a subset of the corpus which has been made available to the language resource community.

2. Related Work

Generally speaking, a game-based crowdsourcing approach uses entertainment rather than financial payment to motivate participation. GWAPs come in many forms; they tend to be graphically rich, with simple interfaces, and give the player an experience of progression by scoring points, being assigned levels and recognising their effort. Systems are required to control the behaviour of players: to encourage them to concentrate on the tasks and to discourage them from malicious behaviour.

1001 Paraphrases (Chklovski, 2005), one of the first GWAPs with the aim of collecting corpora, was developed to collect training data for a machine translation system. The *Open Mind Common Sense* project also led to the development of a game for collecting commonsense knowledge, called *LEARNER* (Chklovski and Gil, 2005).

Perhaps the most successful GWAP that brought the approach into the mainstream was *The ESP Game* which attracted over 200,000 players who produced over 50 million labels for images (von Ahn, 2006). Since then GWAPs have been developed for numerous tasks, including image and video annotation, natural language processing, biomedical research and search refinement (Chamberlain et al., 2013).

Several GWAPs have attempted anaphoric coreference including *PlayCoref*, a two-player game in which players mark coreferential pairs between words in a text (Hladká et al., 2009), and *PhraTris*, a GWAP for syntactic annotation using a general-purpose development platform called GALOAP (Attardi and the Galoap Team, 2010).[1] *PackPlay* was another attempt to build semantically-rich annotated corpora (Green et al., 2010). The two game variants *Entity Discovery* and *Name That Entity* use slightly different approaches in multi-player games to elicit annotations from players. A more unified attempt at creating a gaming platform, named *Wordrobe*[2], targeted different linguistic tasks including part-of-speech tagging, named entity tagging, coreference resolution, word sense disambiguation and compound relations (Venhuizen et al., 2013).

More recently, GWAPs integrated into social networking sites such as *Sentiment Quiz* (Rafelsberger and Scharl, 2009) and *TypeAttack* (Jovian and Amprimo, 2011) on Facebook show that social interaction within a game environment also motivates players to participate in corpus an-

[1] http://galoap.codeplex.com
[2] http://www.wordrobe.org

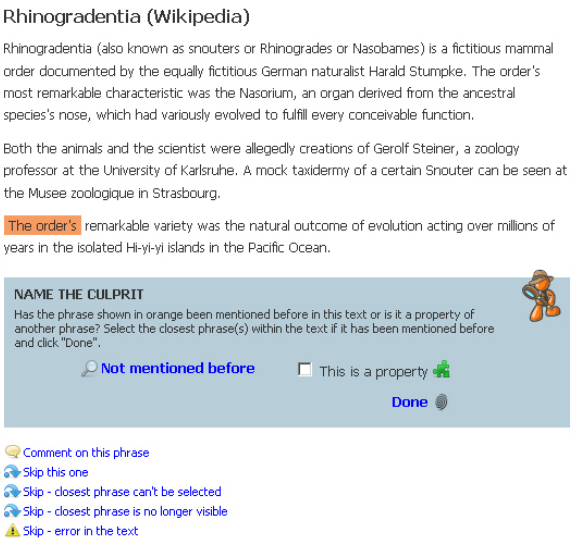

Figure 1: A task presented in Annotation Mode.

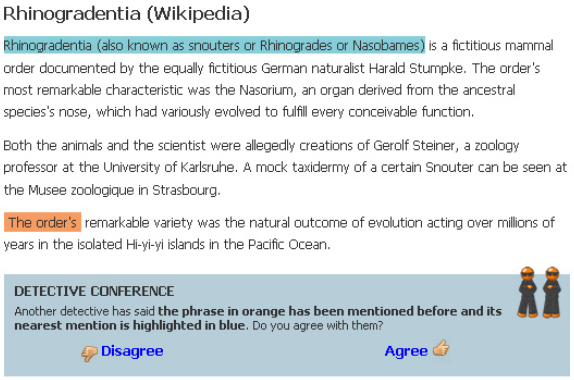

Figure 2: A task presented in Validation Mode.

notation. DigiTalkoot's games *Mole Hunt* and *Mole Bridge*, released on Facebook by the National Library of Finland and Microtask to help digitise old Finnish documents, attracted 110,000 participants who completed over 8 million word fixing tasks in 22 months[3], highlighting the potential for large-scale annotation efforts using a GWAP approach.

3. The Phrase Detectives Game

Phrase Detectives[4] is primarily a GWAP designed to collect data about English (and subsequently Italian) anaphoric coreference (Poesio et al., 2013; Chamberlain et al., 2008).[5] A Facebook version of the game[6] maintained the overall architecture whilst incorporating a number of new features developed specifically for the social network platform.

The game uses two styles of text annotation for players to complete a linguistic task. Initially text is presented in

[3]http://www.digitalkoot.fi

[4]http://www.phrasedetectives.com

[5]Anaphoric coreference is a type of linguistic reference where one expression depends on another referential element. An example would be the relation between the entity *'Jon'* and the pronoun *'his'* in the text *'Jon rode his bike to school.'*

[6]https://apps.facebook.com/phrasedetectives

ID	Gold Standard	*D*	*W*	*M*
GN	Consensus+1	5	874	274
W2	2 experts	5	495	185
G2	2 experts	1	180	69
W1	1 expert	30	12,106	3,953
G1	1 expert	4	6,231	1,971
		45	19,886	6,452

Table 1: Summary of corpora from `Phrase Detectives Corpus 1.0` showing total documents (*D*), total words (*W*) and total markables (*M*).

Annotation Mode (called Name the Culprit in the game, see Figure 1). This is a straightforward annotation mode in which the player makes an **interpretation** (annotation decision) about a highlighted **markable** (section of text). If different players enter different interpretations for a markable then each interpretation is presented to more players in **Validation Mode** (called Detectives Conference in the game, see Figure 2). The players in Validation Mode have to agree or disagree with the interpretation. Players may also make comments about the task and/or skip the task if they do not want to provide an interpretation.

Training texts show the players whether their decisions agree with the gold standard. Once the player has completed all of the training tasks they are given a user rating (the percentage of correct decisions out of the total number of training tasks). The user rating is recorded with every future annotation or validation decision. Players are given training texts until the rating is sufficiently high enough to be given real text from the corpus.[7]

In the first six years of operation (1 Dec 2008 to 30 Nov 2014) over 38,000 players have registered, 2,746 of which went beyond the initial training phase. 524 documents have been fully annotated, for a total completed corpus of 302,224 words, 25% of the total size of the collection currently uploaded for annotation in the game (1.2M words in 839 documents).

4. Phrase Detectives Corpus 1.0

The documents in the corpus are from collections not subject to copyright restrictions including Wikipedia articles and narrative text from Project Gutenberg.[8] `Phrase Detectives Corpus 1.0`[9] (Chamberlain et al., 2016) contains a subset of documents from the main corpus that have been fully annotated and also have a gold standard:

- The GNOME corpus (GN) which already had a documented gold standard (Poesio, 2004a) and was annotated by an additional expert;
- Documents from the Wikipedia (W2) and Gutenberg corpora (G2) that have a gold standard created by two experts;[10]

[7]A minimum rating threshold of 50% is set for the game.

[8]http://www.gutenberg.org

[9]Available from http://anawiki.essex.ac.uk

[10]W2 was also used for an initial investigation of annotation quality (Chamberlain et al., 2009).

Corpus	*W/S*	*W/M*	*%M del*	*%M edit*	Readability
GN	19.4 sd(7.3) n(45)	3.4 sd(1.0) n(45)	0	0	52.3 sd(10.7) n(5)
W2	16.5 sd(7.2) n(30)	2.9 sd(0.6) n(29)	4.3	4.3	53.6 sd(5.6) n(5)
G2	18.0 sd(8.1) n(10)	3.2 sd(1.1) n(10)	7.2	0	88.2 n(1)
W1	21.3 sd(10.0) n(592)	3.5 sd(1.0) n(586)	3.9	9.1	50.7 sd(10.4) n(30)
G1	25.0 sd(17.9) n(249)	3.5 sd(1.0) n(248)	3.9	4.5	84.3 sd(3.5) n(4)

Table 2: Descriptive analysis of the corpora showing words per sentence (*W/S*), active markables per word (*M/W*), the proportion of markables that were deleted (*%M del*) and edited (*%M edit*) and the readability measured as the average (mean) Flesch Reading Ease Score.

- Documents selected at random from completed documents in the Wikipedia (W1) and Gutenberg (G1) corpora with a gold standard created by one expert.

The `Phrase Detectives Corpus 1.0` contains 19,886 words in 45 documents (see Table 1). The format for this corpus has been previously detailed (Poesio et al., 2012) and a further release of completed documents following these guidelines will be made in the future.

4.1. Preprocessing pipeline

A text preprocessing pipeline to convert text and html documents was developed by combining existing tools with ad-hoc modules for correcting the output in the case of frequent errors:

1. A pre-processing step normalised the input, applied a sentence splitter and ran a tokeniser over each sentence (developed from the *openNLP* toolkit[11]);
2. A custom-developed processing step was carried out to clean systematic errors made by the tokeniser and sentence splitter;
3. Each sentence was then analysed by the Berkeley Parser (Petrov et al., 2006);
4. The parser output was used to identify markables in the sentence. As a result an XML-like representation was created which preserved the syntactic structure of the markables (including nested markables, e.g. noun phrases within a larger noun phrase);
5. A heuristic processor identified additional features associated with markables such as person, case, number, etc. The output format was MAS-XML (see section 5.2.).

Once the text had been processed and uploaded into the game, an administrator could edit the markables (character length and character start position), as well as selecting markables to be deleted.[12]

4.2. Descriptive analysis of the corpora

The corpora were analysed for syntactic and structural differences (see Table 2):

- The number of words per sentence (*W/S*) as calculated by PHP word count (str_word_count) on sentences chunked by the pre-processing;
- The number of words per active (not deleted) markable (*W/M*) where the total number of words in each sentence (as calculated by a PHP word count) is divided by the total active markables per sentence. Sentences with no active markables were ignored;
- The average (mean) proportion of markables that were deleted (*%M del*) or edited (*%M edit*) per document;
- The average (mean) readability of each document's content as calculated by an online assessment[13] of the Flesch Reading Ease Score (FRES) (Kincaid et al., 1975). The score is calculated as weighted averages of words per sentence and syllables per word:

$$206.835 - 1.015\frac{total\_words}{total\_sentences} - 84.6\frac{total\_syllables}{total\_words}$$

The two largest corpora (G1 and W1) were then compared (the three other corpora were considered too small to show any significant differences).
G1 had a significantly longer average sentence length than W1 (25.0 words compared to 21.3; unpaired t-test, $p<0.01$) but was significantly easier to read (FRES=84.3 compared to 50.7; unpaired t-test, $p<0.01$). They had the same number of words per markable: 3.5 sd(1.0).
There was no difference between the proportion of markables with errors that needed deleting; however, G1 required fewer markables per document to be edited (4.5% compared to 9.1%; z-test, $p<0.01$).
It might be reasonable to assume that documents that are easier to read are also easier to process using automatic parsing; however, readability in this context only weakly correlates to the proportion of markables deleted (n=34 R=0.16 R^2=0.024; Pearson, weak positive correlation) and edited (n=34 R=0.28 R^2=0.077; Pearson, weak positive correlation).

5. Annotations

5.1. Coding scheme

The corpus was annotated according to the linguistically-oriented approach to anaphoric annotation that is currently prevalent, having been adopted in OntoNotes (Pradhan et al., 2007), the ARRAU corpus (Poesio and Artstein, 2008)

[11] http://opennlp.apache.org

[12] No markables were actually deleted, to ensure database integrity they were instead flagged to be ignored by system outputs.

[13] https://readability-score.com

and in all the corpora used in the 2010 SEMEVAL anaphora evaluation (Recasens et al., 2010).
Markables can be assigned four types of interpretation:

- DN (discourse-new): this markable refers to a newly introduced entity;
- DO (discourse-old): this markable refers to an entity already mentioned in the text;
- NR (non-referring): this markable does not refer to anything (e.g. pleonastic *it*);
- PR (property attribute): this markable represents a property of a previously mentioned entity (e.g. *a teacher* in 'He is a teacher').

Annotations can be examined at three levels of granularity: class; entity or specific. At the **class** level a markable can be assigned one of the four definitions above. At the **entity** level the two classes DO and PR allow for a referring entity to be selected, for example, *he* referring to the entity *Dave* in 'Dave was the best he could be.' At the **specific** level the closest mention of the entity in terms of character distance is considered correct, which allows for linear anaphoric chaining to occur. An example would be *she* referring to the markable *her* in 'Kate wondered if her suit was the best she had.' which are both mentions of the entity *Kate*.
Analysis of specific annotations are presented in this paper and represent the most difficult of the three levels of annotation.

5.2. Export format

The PD-MAS-XML format used to export Phrase Detectives data is a modified version of the Minimum Anaphoric Syntax (MAS-XML) format, a form of inline XML in which the basic information required to carry out resolution is marked (Poesio, 2004b).
As an example, the representation in MAS-XML of the noun phrase *four little rabbits* is as follows:

```
<ne id="ne4" AAcat="num-np" AAgen="neut" AAnum="plur"
    AAper="per3">
  <mod id="AAm2" AAcat="AApre">
    <W Lpos="CD">four</W>
    <W Lpos="JJ">little</W>
  </mod>
  <nphead id="AAh4">
    <W Lpos="NNS">rabbits</W>
  </nphead>
</ne>
```

PD-MAS-XML allows all interpretations for the markables to be stored, leaving it to subsequent processes to select which interpretations to use. The PD-MAS-XML file includes:

- the original text;
- the markup of sentences, NPs, their features and constituents as automatically computed by the import pipeline (MAS-XML format);
- the gold standard expert annotations;

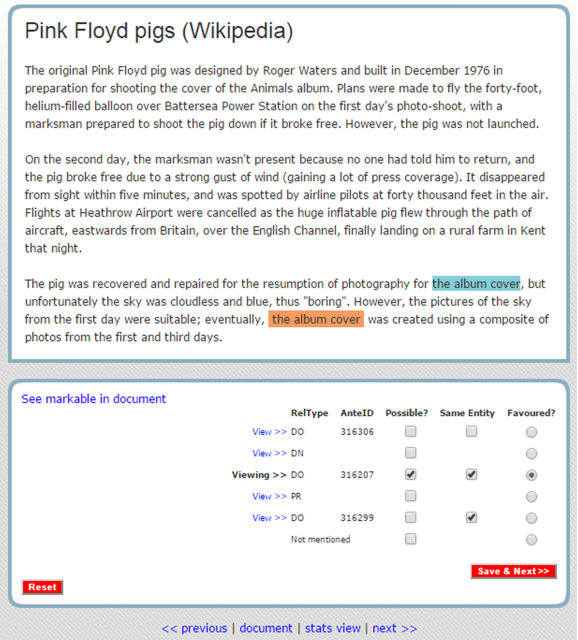

Figure 3: Screenshot showing the expert annotation administration interface.

- the annotations produced by the players including the user ID, the user rating, the time it took to make the decision, whether the decision is an agreement, in what mode the decision occurred (annotation or validation), timestamp, and the interface that was used;
- any player comments about the markable;
- and any time a player skipped the markable.

The following is a simplified example of how player annotations are appended to the MAS-XML to create PD-MAS-XML:

```
<PDante id="ne14817">
  <interpretation>
    <anchor type="DN" user_id="281" user_rating="75"
        annotation_time="2" agree="y" mode="a" timestamp
        ="2009-02-07 17:16:46"/>
    <anchor type="DO" user_id="728" user_rating="58"
        annotation_time="2" agree="y" mode="a" ante="
        ne18253" timestamp="2009-02-07 17:14:46"/>
    <anchor type="PR" user_id="718" user_rating="60"
        annotation_time="5" agree="y" mode="a" ante="
        ne18253" timestamp="2009-02-07 17:14:46"/>
    <anchor type="DN" user_id="1364" user_rating="59"
        annotation_time="4" agree="y" mode="v" timestamp
        ="2009-02-07 17:12:46"/>
    <anchor type="DN" user_id="163" user_rating="80"
        annotation_time="2" agree="y" mode="v" timestamp
        ="2009-02-07 17:11:46"/>
    <anchor type="DN" user_id="165" user_rating="85"
        annotation_time="2" agree="n" mode="v" timestamp
        ="2009-02-07 17:10:46"/>
    <anchor type="DN" favoured="y" user_id="2" mode="e"
        timestamp="2009-04-07 17:10:46"/>
    <anchor type="DN" favoured="y" user_id="18" mode="e"
        timestamp="2009-05-02 12:11:42"/>
  </interpretation>
  <comment timestamp="2011-07-22 04:23:59" user_rating="
      92" user_id="123" type="not_selectable" type_id="
      1"/>
  <skip timestamp="2011-07-22 04:23:59" annotation_time=
      "8" user_rating="92" user_id="123"/>
</PDante>
```

	GN n(59)	**W2** n(154)	**G2** n(57)
DN	-	99.0%	85.7%
DO	93.2%	84.8%	91.6%
NR	-	100%	-
PR	-	72.7%	-
Overall	93.2% (κ=0.93)	94.1% (κ=0.88)	89.4% (κ=0.88)

Table 3: Inter-expert agreement between e2 and e18 (DN = discourse-new, DO = discourse-old, NR = non-referring, PR = property attribute).

6. Gold Standard

In order to create the gold standard, the expert annotator was shown a list of all markable interpretations that had been entered by the players for a particular markable and could view each interpretation as if in Validation Mode (see Figure 3). By default the expert could not see how many annotations or validations each interpretation had scored. The markables were annotated in order of appearance in the text.

The expert selected the best interpretation for the markable (the 'favoured' radio button) and selected the checkbox of any possible interpretations due to ambiguity. Additionally, if the markable was referring, the expert selected the checkboxes of any other interpretation that was for the same entity. If the most appropriate interpretation was not mentioned in the list submitted by the players the expert could indicate the best markable interpretation as 'Not mentioned'. Markables that were marked as deleted did not require an expert annotation.

Complete instructions and examples for experts on how to annotate apposition, discourse diexsis, out-of-context errors, questions, names, compound entities, bridging entities, temporal revelations, numerators and dates are detailed in the supplementary files attached to the resource.

6.1. Inter-expert agreement

Five documents from the Wikipedia corpus, containing 154 active markables (W2), and one document from the Gutenberg corpus, containing 57 active markables (G2), were manually annotated by two experts operating independently, called e2 and e18.[14]

Five documents from the GNOME (GN) corpus were annotated by e2 and compared to the consolidated annotations of the GNOME corpus (in which e18 was the main annotator).[15] DN and PR annotations were not recorded and there were no instances of NR markables. The GNOME annotations were manually converted into the Phrase Detectives corpus scheme under the expert ID e39181. In total there were 59 markables that e2 and GNOME produced an annotation for (see Table 3).

Overall, agreement between experts in the three corpora was very high although not complete: 93.2% (GN), 94.1% (W2) and 89.4% (G2), for a chance-adjusted κ value (Artstein and Poesio, 2008) of $\kappa = .93$, $\kappa = .88$ and $\kappa = .88$ respectively, which is extremely good. This value can be seen as an upper boundary on what we might expect to get out of a crowdsourcing system.

There was no significant difference between the inter-expert agreement of the three corpora (GN n(59) 93.2%; W2 n(154) 94.1%; G2 n(57) 89.4%; p=0.810, p=0.238, p=0.465, z-test). This indicates that the expert annotations created by e2 are what could be considered a gold standard between document domains and when compared to an existing gold standard or another linguistic expert. Expert annotator e2 also created the gold standard for W1 and G1.

6.2. Baseline crowd agreement

Traditional methods of measuring annotation accuracy generally assume a singularity of correct answers and for comparative purposes the baseline agreement is presented in this way. Measuring accuracy of a multi-dimensional annotation set is more complex and is the subject of future work. *Quality* is measured as the level of agreement between an expert and the highest scoring system answer. *Noise* is defined as the average (mean) number of wrong interpretations per markable.

The annotations and validations of each markable from each corpus were analysed and either aggregated to produce a best answer or were excluded because:

- the markable has been marked by an administrator to be deleted;
- the expert did not provide an answer (therefore an answer was not possible);
- the markable was skipped by enough players (the markable did not have eight annotations when considered complete).

All annotations and validations for each interpretation of a markable were combined:

$$A + V_a - V_d$$

where A is the number of players initially choosing the interpretation in Annotation Mode, V_a is the number of players agreeing with that interpretation in Validation Mode, and V_d is the number of players disagreeing with it in Validation Mode. This formula is used to score each interpretation of a markable, with the highest scoring interpretation called the 'best' or game interpretation (Chamberlain, 2014).

The baseline agreement in the three corpora where two experts provided a gold standard show very high agreement, comparable to pairwise inter-expert agreement (see Table 4). These values are also comparable to those obtained when comparing an expert with trained annotators (usually students) that are typically used to create medium-quality resources (Poesio et al., 2013). Both W1 and G1 have lower

[14]The two experts were Jon Chamberlain (e2), who developed the game and wrote the instructions, and Massimo Poesio (e18), a linguistic expert in anaphoric coreference.

[15]The GNOME annotation scheme records DO annotations as 'ident' variables. Plural DO was only annotated once (as an 'element-inv' variable) and was not imported here to avoid additional conflict in the annotation instructions.

	GN		W2		G2		W1	G1
	e2	e39181	e2	e18	e2	e18	e2	e2
Markables	264	61	176	160	63	58	3,729	1,844
Agreement	93.9%	85.2%	84.0%	81.8%	96.8%	93.1%	79.1%	86.6%
Kappa κ	0.86	0.85	0.63	0.59	0.96	0.92	0.52	0.85
$Noise_{mean}$	1.6		2.7		2.6		1.3	1.4
	sd(2.0)		sd(3.4)		sd(2.1)		sd(1.6)	sd(1.3)

Table 4: Baseline agreement between experts and the best answer from the game.

	DN	DO	NR	PR	NM
GN n(275)	189 (68.7%)	65 (23.6%)	0	4 (1.4%)	17 (6.1%)
W2 n(176)	128 (72.7%)	33 (18.7%)	1 (0.5%)	13 (7.3%)	1 (0.5%)
G2 n(63)	27 (42.8%)	36 (57.1%)	0	0	0
W1 n(3,729)	2,502 (67.0%)	912 (24.4%)	23 (0.6%)	108 (2.8%)	184 (4.9%)
G1 n(1,884)	638 (33.8%)	1,160 (61.5%)	25 (1.3%)	21 (1.1%)	40 (2.1%)

Table 5: Summary of the distribution of interpretations for active markables.

agreement (quality) than W2 and G2, significantly so in the Gutenberg corpus (G1-G2, z-test, p=0.02; W1-W2, z-test, p=0.12) which may be because the latter documents were worked on by more linguists, rather than the former documents which were worked on by a non-expert crowd.[16]

The Gutenberg corpus had a higher agreement than the Wikipedia corpus (G1 n(1,844) 86.6%, W1 n(3,729) 79.1%, p<0.01, z-test) showing that the narrative texts of Gutenberg are perhaps easier to annotate.

The prevalence of genuinely ambiguous interpretations is low (0.5% in G1 and 0.8% in W1); however, learning to identify these cases automatically may be of most interest to linguists (see Section 8.).

7. Task distribution and difficulty

During the data collection all markables were treated in the same way; however, it is clear that some markables are easier to annotate than others, either because of the text itself (contextual difficulty) or because of the type of relation it has with the other markables (interpretation difficulty).

7.1. Contextual difficulty

It could be assumed that the more complex the text, the more difficult the players would find the task of annotating the markables, and therefore the quality would be lower. However, agreement per document shows a weak positive correlation to readability (n(45) R=0.19 R^2=0.037; Pearson, weak positive correlation) implying readability has little impact on the user's ability to perform annotation tasks. The documents in the corpus were not complex and other documents might show different results.

To investigate whether document length had an impact on difficulty the W1 corpus was split into two groups, one with long documents (WL1, >700 words, n(8) mean=819.4 sd(199.3)) and one with short documents (WS1, <700 words, n(22) mean=272.6 sd(50.2)). There was no difference between the agreement in the Wikipedia long and short corpora (WL1 n(1,947) 79.9%; WS1 n(1,782) 78.1%; z-test, p=0.18) which suggests that document length also does not seem to impact on a player's ability to annotate the text.

7.2. Interpretation difficulty

In order to explore whether some types of interpretation are harder to detect and annotate than others, first we looked at how the classes of interpretation are distributed through the corpora, then looked at the agreement of each class.

The distribution of annotation class was calculated as a proportion of interpretations of active markables as determined by an expert (e2). When there was no correct interpretation the markable interpretation would be classed as NM (Not Mentioned), see Table 5.

The documents in G1 have more coreferring DO markables (61.5%) than in the documents in W1 (24.4%), with the reverse being true for DN markables. NR and PR markables are rare in both corpora (W1 n(3,729); G1 n(1,884); χ^2=763.6, p<0.01). One explanation might be that as Wikipedia articles, which are explanatory in nature, become longer they introduce more entities to explain the topic of the document. The reverse could be true for Gutenberg documents, which are mainly narratives, that will introduce entities and continue to refer to them throughout the discourse.

A closer look at the breakdown of agreement between the best game answer and e2 shows a significant difference between the performance of players on the Gutenberg and Wikipedia corpora on different tasks (see Table 6). These results suggest that DN is an easier task and as W1 has more true DN markables it could be expected that the W1 corpus would be annotated to a higher quality. However, this is not the case due to the poor performance of interpretations of DO markables in the W1 corpus. This indicates that task difficulty has a considerable impact on the quality that can be achieved by a crowd.

[16]The intial deployment of the game was publicised at language conferences, forums and linguist blogs.

	G1	W1
Markables	1,844	3,729
DN	91.5% (584 of 638)	98.5% (2,466 of 2,502)
DO (specific)	88.0% (1,021 of 1,160)	49.8% (455 of 912)
NR	96.0% (24 of 25)	65.2% (15 of 23)
PR (specific)	19.0% (4 of 21)	12.9% (14 of 108)
Overall agreement	86.6%	79.1%

Table 6: Breakdown of agreement between each interpretation type (as determined by e2) and the best game answer on the Wikipedia and Gutenberg corpora, showing a difference in all classes of interpretation ($p<0.01$, z-test).

8. Discussion

Despite the high performance of annotators, it is worth discussing the types of errors that were made during the annotation of the corpus.

Annotators make **genuine errors** caused by a slip of attention or fatigue. Experts could review and rectify their mistakes, but players were not allowed to go back and correct annotations they knew they had made incorrectly (as this would have influenced the scoring of the game).

Other errors were introduced because annotators were not sure of the correct way to mark up the text and numerous examples of **markup ambiguity** were added to the guidelines. Whilst expert annotators were expected to read and understand all types of markup, most players only had a general understanding of what was required.

Of the 12 markables on which the experts did not agree in the W2 and G2 corpora, only one was a genuine error in which the entity had been correctly identified but not the closest mention. The remaining disagreements fell into four categories of markup ambiguity:

The first category was the **specificity** of the antecedent. The preprocessing chunks markables in a way that allows different levels of specificity to be selected, for example, in *Henry the Hexapus (Wikipedia)* the markable *the Blackpool Centre* refers to an earlier mention of *the Blackpool Sea Life Centre in North West England*; however, a less specific markable within this markable was also selectable: *the Blackpool Sea Life Centre*.

The second category relates to **assumptions** the reader makes regarding the role of entities, for example in *Gay Fuel (Wikipedia)* it could be assumed that the acronym *LLC* and *Its maker* refer to the manufacturer of the drink *Florida-based Speciality Spirits*; however, this is not explicitly stated in the context so the reader has to make an assumption about the role of the entity.

A further example of this type of markup ambiguity was for temporal revelations. Revelations that are made during the context of the document should have been marked up at the point of revelation and not retrospectively marked throughout the text, for example, *the bridegroom* would not be marked up as also being *the robber* throughout *The Robber Bridegroom (The Brothers Grimm)* when it is only revealed at the end of the text.

The third category was confusion over what constitutes a **property** of another markable and what is in fact another entity, for example, in *Gay Fuel (Wikipedia)* whether *bright pink and elderberry flavored* is a property of *the liquid* in '...the liquid was dyed bright pink and elderberry flavored.' Examples of this type of ambiguity explain why property markables are more difficult to annotate.

The final category was for entity **generalisations**, in which perhaps coreference is not an appropriate annotation, for example, in *Human Mail (Wikipedia)* the mentions of *a person* in the sentences 'Human mail is the transportation of a person through the postal system.' and '...is the mailing of a part of a person...'

Examples of **genuine ambiguity**, where there are truly two (or more) indistinguishable interpretations for a markable were quite rare in this corpus (as indicated by the expert gold standard showing there to be at least two interpretations possible for a markable). Developing ways to identify true linguistic ambiguity from other forms of error and noise is a key goal for future research.

It is our hope that the overall quality of the crowd decisions in this corpus are high enough to use this data in more sophisticated models of annotation (Raykar et al., 2010) to understand annotator bias and automatically identify genuine ambiguity over poor decisions. Next steps include developing methods for cleaning up the data with filtering and optimisation and for using the data to train anaphoric models. A larger corpus will also be released in the future.

9. Conclusion

Phrase Detectives was one of the first GWAPs applied to language resource creation and in quantitative terms has been one of the most successful, collecting over three million judgements in six years. The baseline figures for the five gold standard corpora show high quality at near-expert annotator performance. Analysis of the corpus also suggests that factors such as document length and readability do not impact agreement; however, users find it harder to detect and annotate different types of interpretation.

Phrase Detectives, and other GWAPs for language resource creation, have shown that for large-scale, persistent annotation efforts, a game-based crowdsourcing approach could be considered based on factors such as cost and scalability.

Acknowledgments

The creation of the original game was funded by EPSRC project AnaWiki, EP/F00575X/1. Dr Chamberlain would also like to acknowledge the EPSRC Doctoral Training Award that enabled the analysis of the corpus.

10. Bibliographical References

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

Attardi, G. and the Galoap Team. (2010). Phratris. Demo presented at the 9th International Semantic Web Conference (ISWC'10) tutorial INSEMTIVES'10.

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008). Phrase Detectives: A web-based collaborative annotation game. In *Proceedings of the 2008 International Conference on Semantic Systems (I-Semantics'08)*.

Chamberlain, J., Kruschwitz, U., and Poesio, M. (2009). Constructing an anaphorically annotated corpus with non-experts: Assessing the quality of collaborative annotations. In *Proceedings of the 4th International Joint Conference on Natural Language Processing (IJCNLP'09) Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*.

Chamberlain, J., Fort, K., Kruschwitz, U., Mathieu, L., and Poesio, M. (2013). Using games to create language resources: Successes and limitations of the approach. In *ACM Transactions on Interactive Intelligent Systems*, volume The People's Web Meets NLP: Collaboratively Constructed Language Resources. Springer.

Chamberlain, J. (2014). The Annotation-Validation (AV) model: Rewarding contribution using retrospective agreement. In *Proceedings of the 1st International Workshop on Gamification for Information Retrieval (GamifIR'14)*.

Chklovski, T. and Gil, Y. (2005). Improving the design of intelligent acquisition interfaces for collecting world knowledge from web contributors. In *Proceedings of the 3rd International Conference on Knowledge Capture (K-CAP'05)*.

Chklovski, T. (2005). Collecting paraphrase corpora from volunteer contributors. In *Proceedings of the 3rd International Conference on Knowledge Capture (K-CAP'05)*.

Green, N., Breimyer, P., Kumar, V., and Samatova, N. F. (2010). Packplay: Mining semantic data in collaborative games. In *Proceedings of the 4th Linguistic Annotation Workshop (LAW IV)*.

Hladká, B., Mírovskỳ, J., and Schlesinger, P. (2009). Play the language: Play coreference. In *Proceedings of the 4th International Joint Conference on Natural Language Processing (IJCNLP'09)*.

Howe, J. (2008). *Crowdsourcing: Why the power of the crowd is driving the future of business*. Crown Publishing Group.

Jovian, L. T. and Amprimo, O. (2011). OCR correction via human computational game. In *Proceedings of the 44th Hawaii International Conference on System Sciences (HICSS'11)*.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Technical report.

Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics*.

Poesio, M. and Artstein, R. (2008). Anaphoric annotation in the ARRAU corpus. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2012). The Phrase Detective multilingual corpus, release 0.1. In *Proceedings of LREC'12 Workshop on Collaborative Resource Development and Delivery*.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2013). Phrase Detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*.

Poesio, M. (2004a). Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*.

Poesio, M. (2004b). The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue (SIGDIAL'04)*.

Pradhan, S. S., Ramshaw, L., Weischedel, R., MacBride, J., and Micciulla, L. (2007). Unrestricted coreference: Identifying entities and events in OntoNotes. In *Proceedings of the 1st IEEE International Conference on Semantic Computing (ICSC'07)*.

Rafelsberger, W. and Scharl, A. (2009). Games with a purpose for social networking platforms. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322.

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010). Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of 5th International Workshop on Semantic Evaluations (SEMEVAL'10)*.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast - but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*.

Venhuizen, N., Basile, V., Evang, K., and Bos, J. (2013). Gamification for word sense labeling. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS'13)*.

von Ahn, L. (2006). Games with a purpose. *Computer*, 39(6):92–94.

11. Language Resource References

Chamberlain, J., and Poesio, M., and Kruschwitz, U. (2016). *Phrase Detectives Corpus 1.0*. AnaWiki, University of Essex, ISLRN 955-898-221-547-9.