University of Essex

School of Computer Science and Electronic Engineering

A new life for a dead parrot: Incentive structures in the Phrase Detectives game

Jon Chamberlain
jchamb@essex.ac.uk

Massimo Poesio
poesio@essex.ac.uk

Udo Kruschwitz
udo@essex.ac.uk

20 April 2009

Funded by EPSRC

Overview

- AnaWiki project at the University of Essex, UK
- Phrase Detectives Online Game
  - Citizen Science / Games with a purpose
  - How the game collects annotations
  - Incentive structures in the game
  - Attracting players to the game
  - Evaluating annotations
- Corpus
- Conclusions
- Future work

The AnaWiki Project

- Address the bottleneck in creating large annotated text resources
- Investigate using Web volunteers to annotate text
  - Define criteria for motivating volunteers
  - Define criteria of volunteer evaluation
- Build on proposals to allow for ambiguity whilst still detecting errors
- Create a large, hand-annotated corpus
  - Created from balanced texts
  - Initially annotated with anaphoric information

E.g. This parrot is no more! He has ceased to be!
{antecedent} {anaphor}

Citizen Science

Using humans to perform tasks that computers find difficult.

- Open Mind Common Sense (MIT)
- Crater mapping (NASA)
- Learner
  Learner2
  1001 Paraphrases (Chklovski)
- FACTory (CyCORP)
- Hot or Not (8 Days)

Games with a Purpose

Using humans to complete tasks in an entertaining game.

- ESP
  Phetch
  Verbosity
  Peekaboom
  Tag-a-Tune (Carnegie Mellon University)

- OntoGame
  OntoTube
  OntoPronto (University of Innsbruck)

- Categorilla
  Free Association (Stanford University)

Phrase Detectives – An Overview

- Simple and friendly game interface
- Easy to learn and quick to play
- Used by large numbers of non-expert volunteers

Phrase Detectives – An Overview

- Simple and friendly game interface
- Easy to learn and quick to play
- Used by large numbers of non-expert volunteers
- Text is pre-processed to identify “markable” phrases

Phrase Detectives – An Overview

- Simple and friendly game interface
- Easy to learn and quick to play
- Used by large numbers of non-expert volunteers
- Text is pre-processed to identify “markable” phrases
- Players annotate the markable phrases in 4 ways
- The game presents this task using 2 modes

The annotation task

For each markable phrase the player must decide if it is:

- **Discourse-new (DN)**
  The markable has not been mentioned before in the text

  E.g. A customer enters **a pet shop**.

- **Discourse-old (DO)**
  The markable has been mentioned before and the player must identify the closest previous markable(s).

  E.g. This parrot is no more! **He** has ceased to be!

- **Non-referring (NR)**
  The markable does not refer to anything.

  E.g. Yeah, well **it**'s not easy to pad these Python files out to 150 lines, you know.

- **Property (PR)**
  The markable is a property of another markable which the player must identify.

  E.g. I wanted to be **a lumberjack**!

Game modes

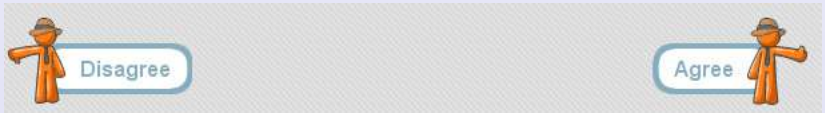
The game collects annotations using 2 game modes:

- **Find The Culprit** (Annotation Mode)

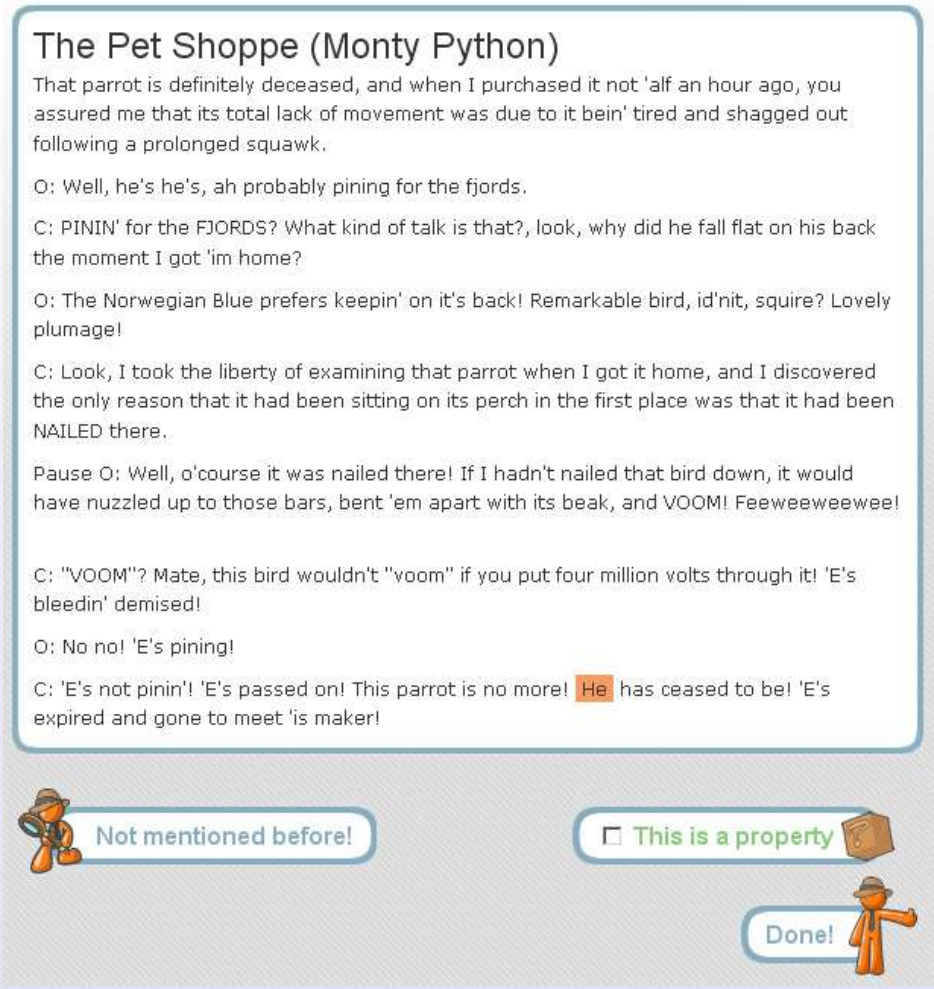
  The player makes an annotation

- **Detectives Conference** (Validation Mode)

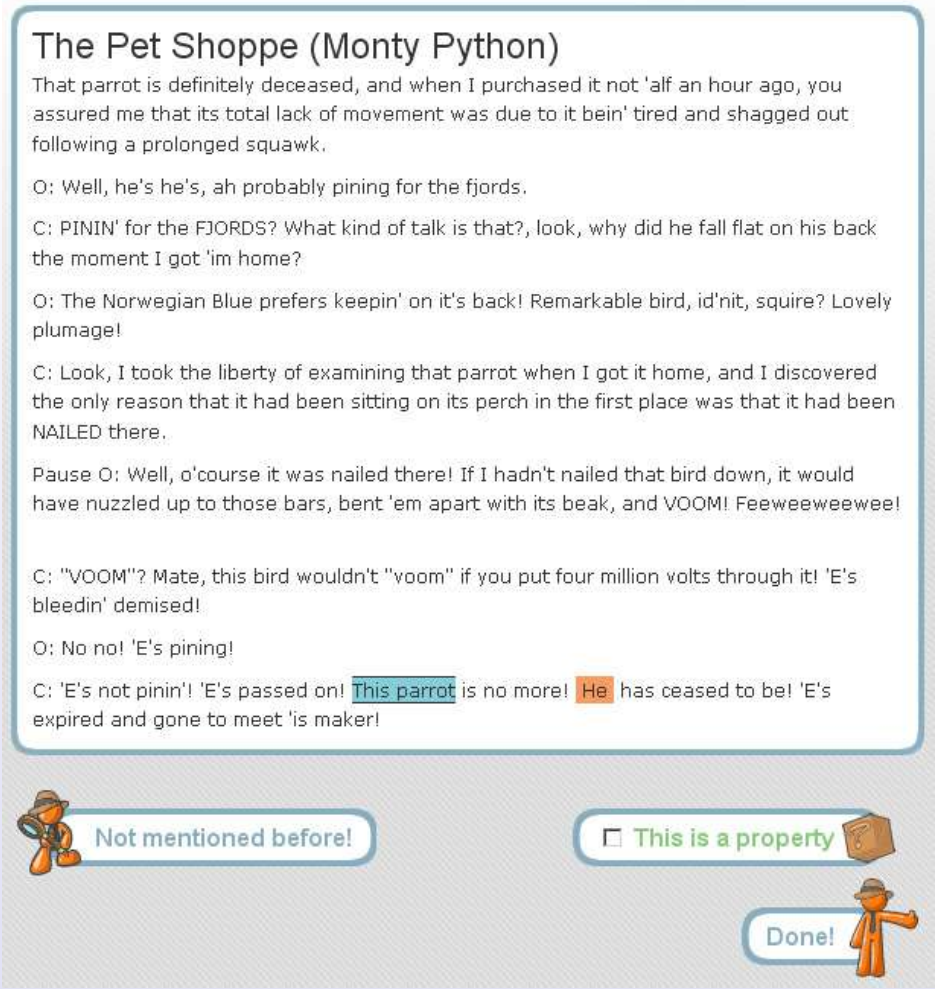
  The player must agree/disagree with a decision made by another player.



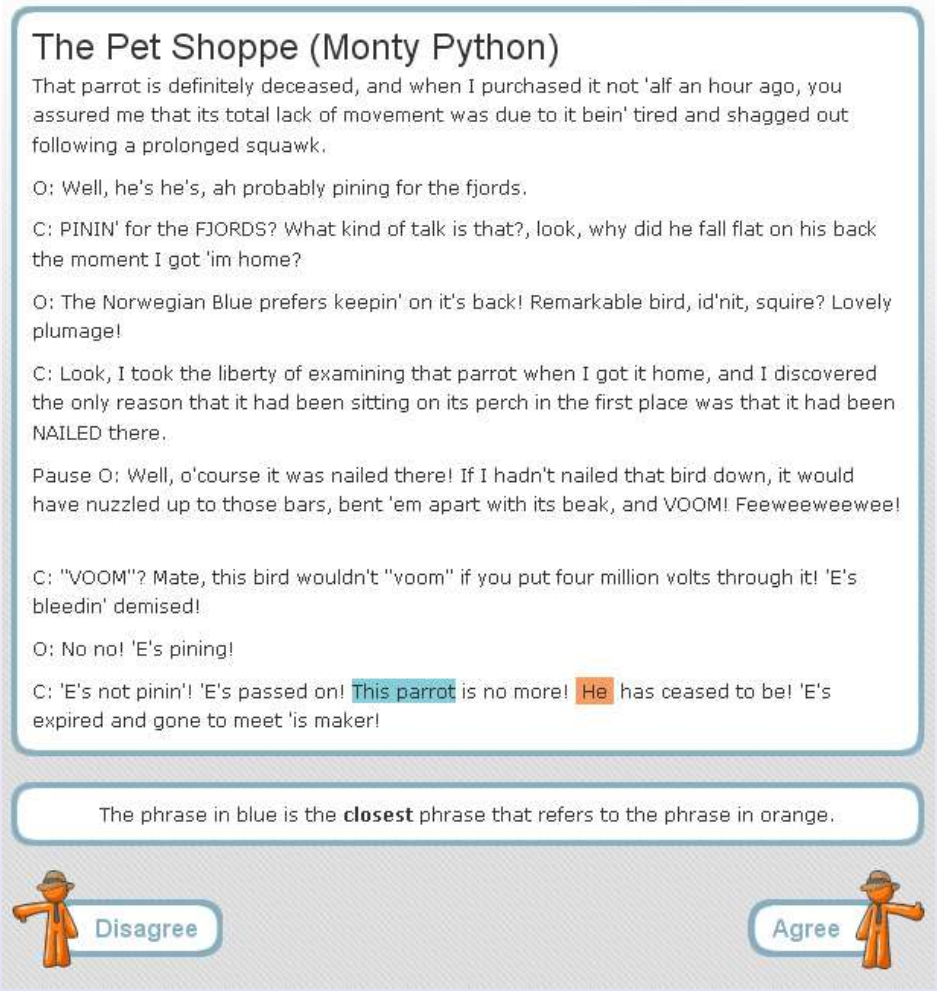
1. Find The Culprit (Annotation Mode)



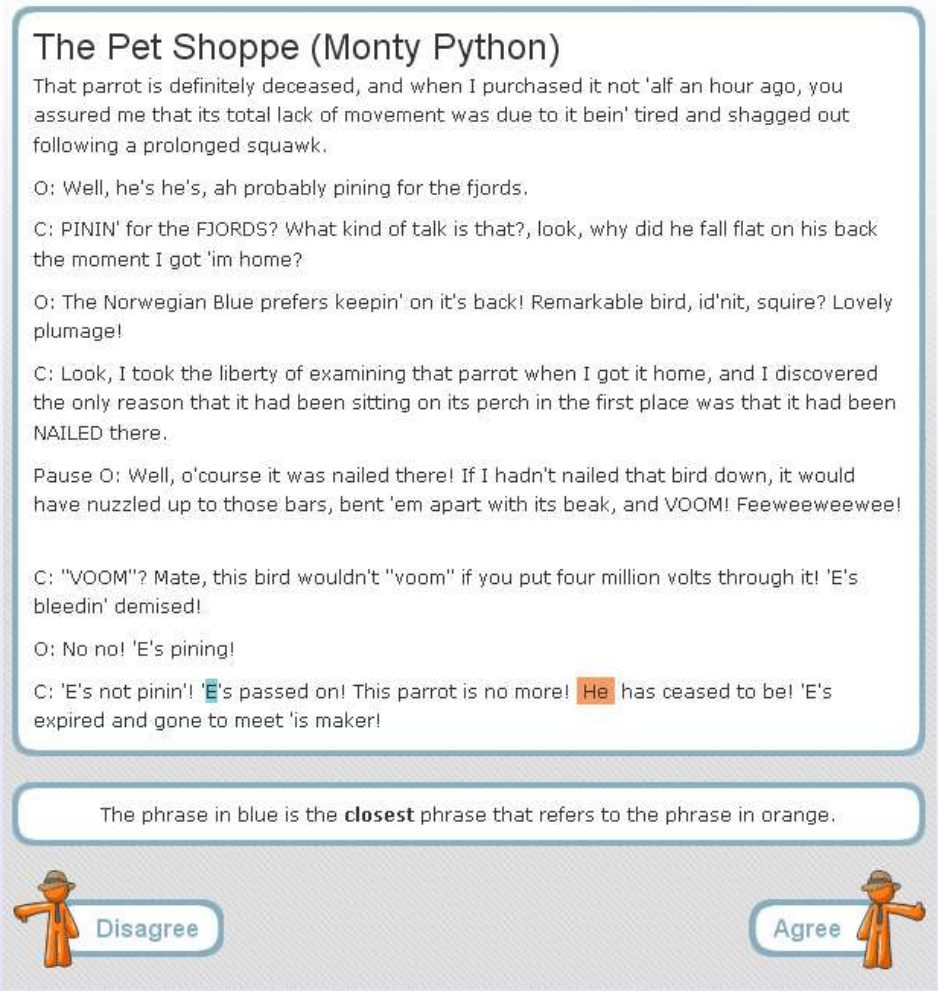
1. Find The Culprit (Annotation Mode)



2. Detectives Conference (Validation Mode)



2. Detectives Conference (Validation Mode)



Training players

- Players must complete a training phase
  - Shows how the tasks work and what to look for
  - Gives feedback on right or wrong answers
  - Provides an initial rating for the player
  - The player is kept in training until their rating is above the training threshold
  - Their current rating is recorded with all their annotations
- Players are rated at random with a Gold Standard text
  - The user will return to training mode if their rating is too low

Scoring in the game

Scoring designed to reward **quality** as well as **quantity**

How do you reward good decisions?

Scoring in the game

Scoring designed to reward **quality** as well as **quantity**

How do you reward good decisions?

- **Comparative scoring**
  - The player's decision is compared to a known answer.

Scoring in the game

Scoring designed to reward **quality** as well as **quantity**

How do you reward good decisions?

- **Comparative scoring**
  - The player's decision is compared to a known answer.
- **Collaborative scoring**
  - Many games use partner matching for scoring

    i.e. if you put in the same answer as your game partner you both score points
  - Phrase Detectives uses group collaboration for scoring

    i.e. if your answer agrees with other players then you all score more points

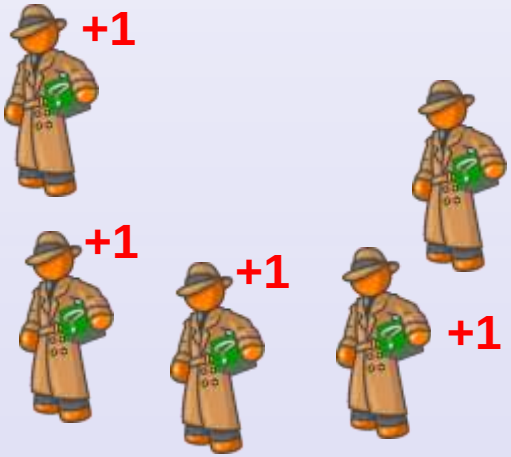
Scoring in the game: Annotation

A player makes a decision about a markable phrase in Annotation Mode

Scoring in the game: Annotation

A player makes a decision about a markable phrase in Annotation Mode

Other players also make decisions



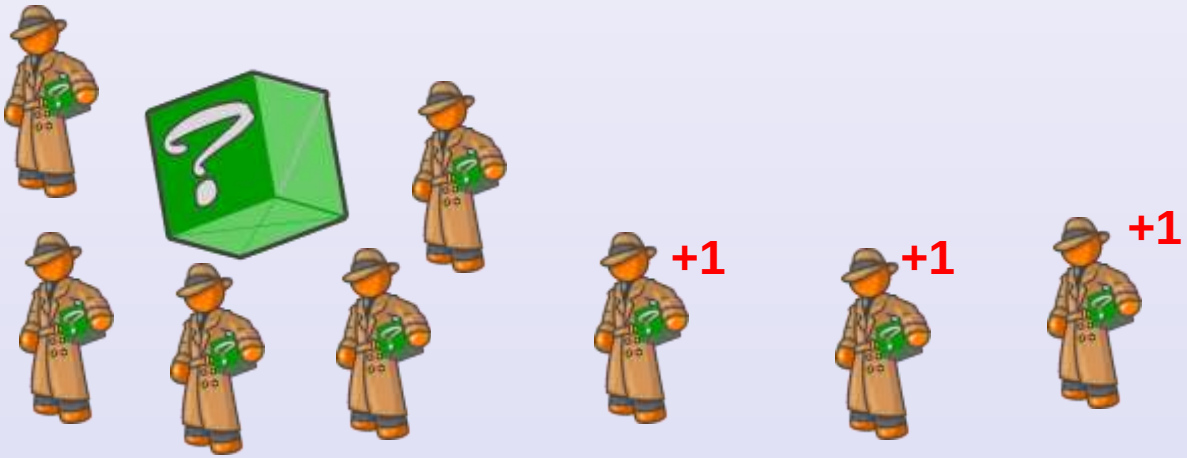
Scoring in the game: Annotation

A player makes a decision about a markable phrase in Annotation Mode

Other players also make decisions

Currently 8 players in this group

If they all agree then the markable is considered complete



Scoring in the game: Annotation

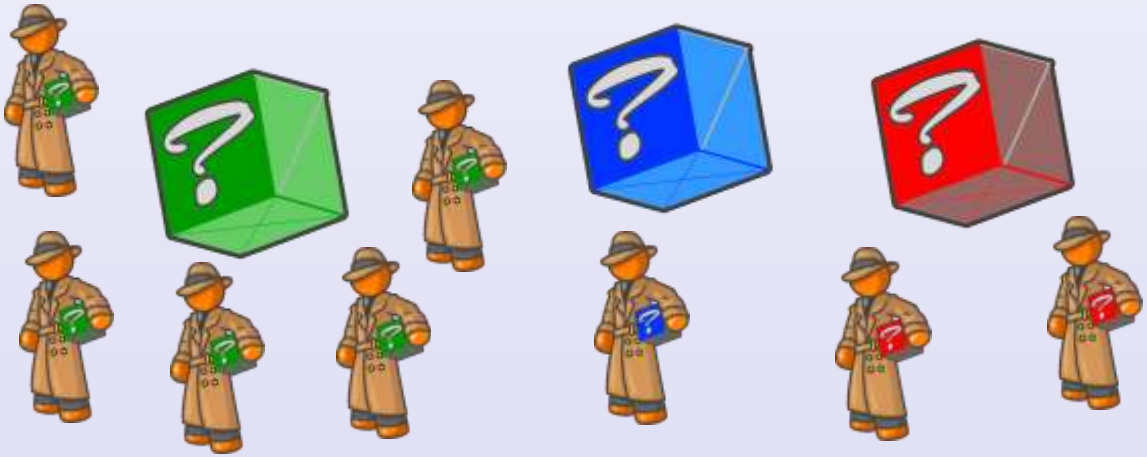
A player makes a decision about a markable phrase in Annotation Mode

Other players also make decisions

Currently 8 players in this group

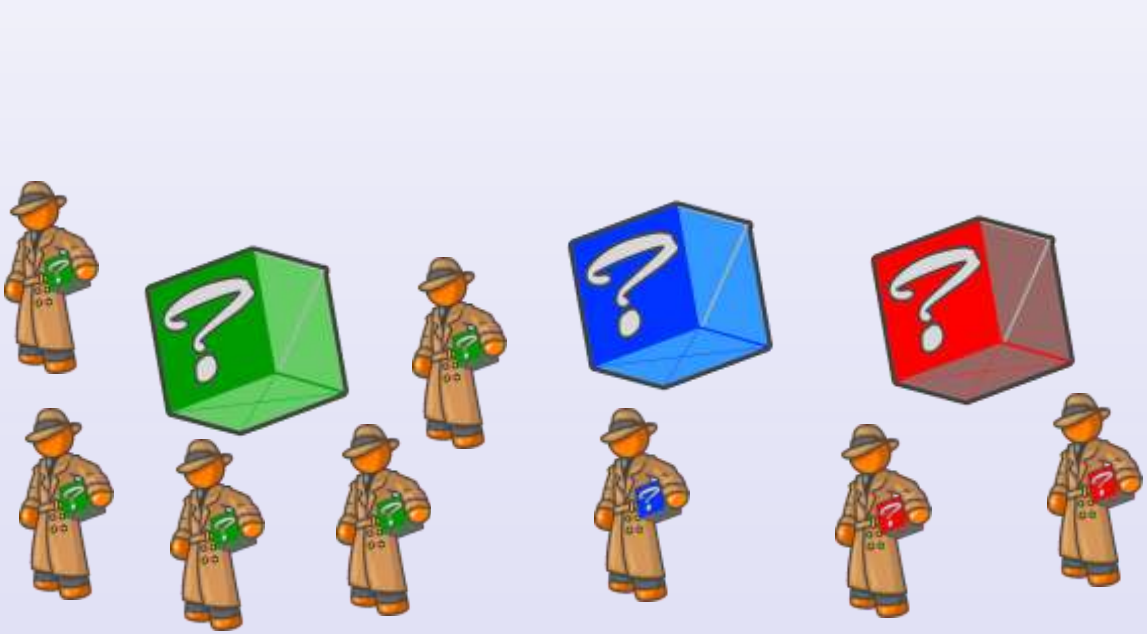
If they all agree then the markable is considered complete

If they do not agree then the markable needs validating (62% need validating)



Scoring in the game: Validation

Each unique relation for the markable phrase is presented to more players



Scoring in the game: Validation

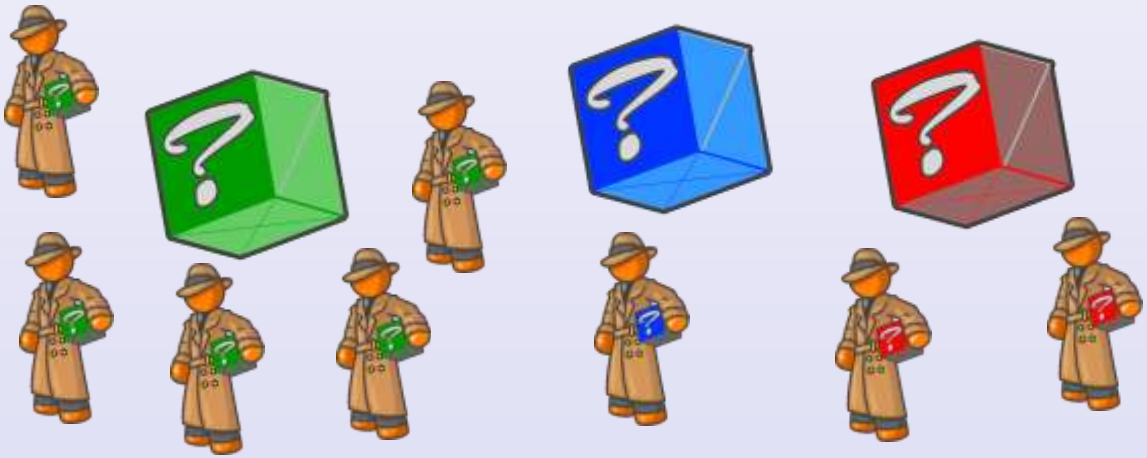
Each unique relation for the markable phrase is presented to more players

The player agrees or disagrees with the relation

The player scores a point for every player from the first group who agree

AGREE!

+5



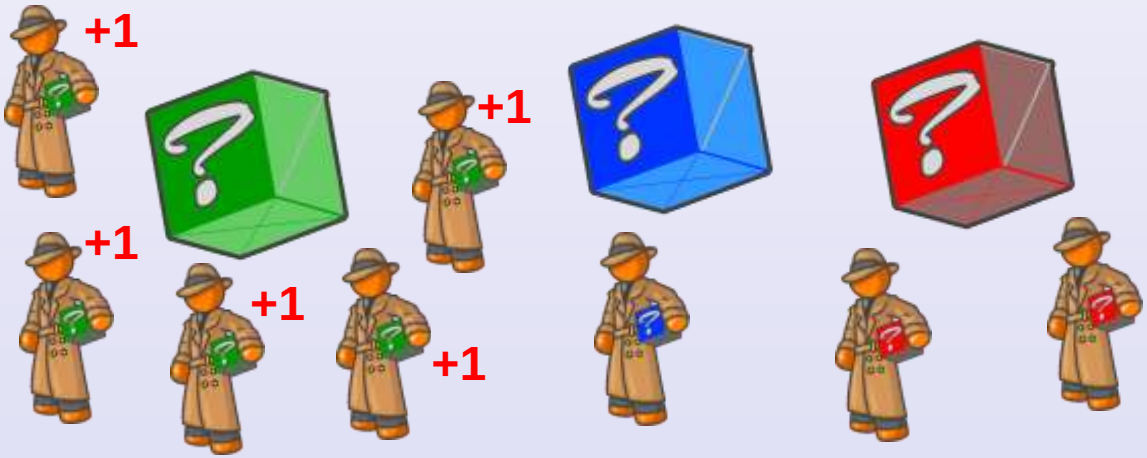
Scoring in the game: Validation

Each unique relation for the markable phrase is presented to more players

The player agrees or disagrees with the relation

The player scores a point for every player from the first group who agree

These players also score a point



AGREE!

+5

Incentive structures

- Personal

  Participation is the reward

- Social

  Participation improves position amongst peers (players)

- Financial

  Participation is rewarded with money

Incentive structures: Personal

- Contribution to a worthwhile project
- Their contribution is valued

Incentive structures: Personal

- Contribution to a worthwhile project
- Their contribution is valued
- Text is interesting
- Player can choose topics to read (stories, science, travel etc)

Incentive structures: Personal

- Contribution to a worthwhile project
- Their contribution is valued
- Text is interesting
- Player can choose topics to read (stories, science, travel etc)
- Reading speed (450 annotations per hour)
- Timed tasks

Incentive structures: Personal

- Contribution to a worthwhile project
- Their contribution is valued
- Text is interesting
- Player can choose topics to read (stories, science, travel etc)
- Reading speed (450 annotations per hour)
- Timed tasks
- User contributed text

Incentive structures: Social

- Weekly, monthly and all-time leaderboards
- Medals and cups for weekly/monthly high-scores
- Named levels

Incentive structures: Social

- Weekly, monthly and all-time leaderboards
- Medals and cups for weekly/monthly high-scores
- Named levels
- Agreement leaderboard

Incentive structures: Social

- Weekly, monthly and all-time leaderboards
- Medals and cups for weekly/monthly high-scores
- Named levels
- Agreement leaderboard
- Communication
  - comments
  - feedback
- Encourage a player community

Incentive structures: Financial

- Weekly prize draw
  (£15 Amazon voucher)

  Motivates low scoring players as every annotation has a chance

- Monthly prize for the 3 highest scorers
  (£75, £50 and £25 Amazon vouchers)

  Motivates all players at the beginning of the month

  Motivates high scoring players (with more experience)

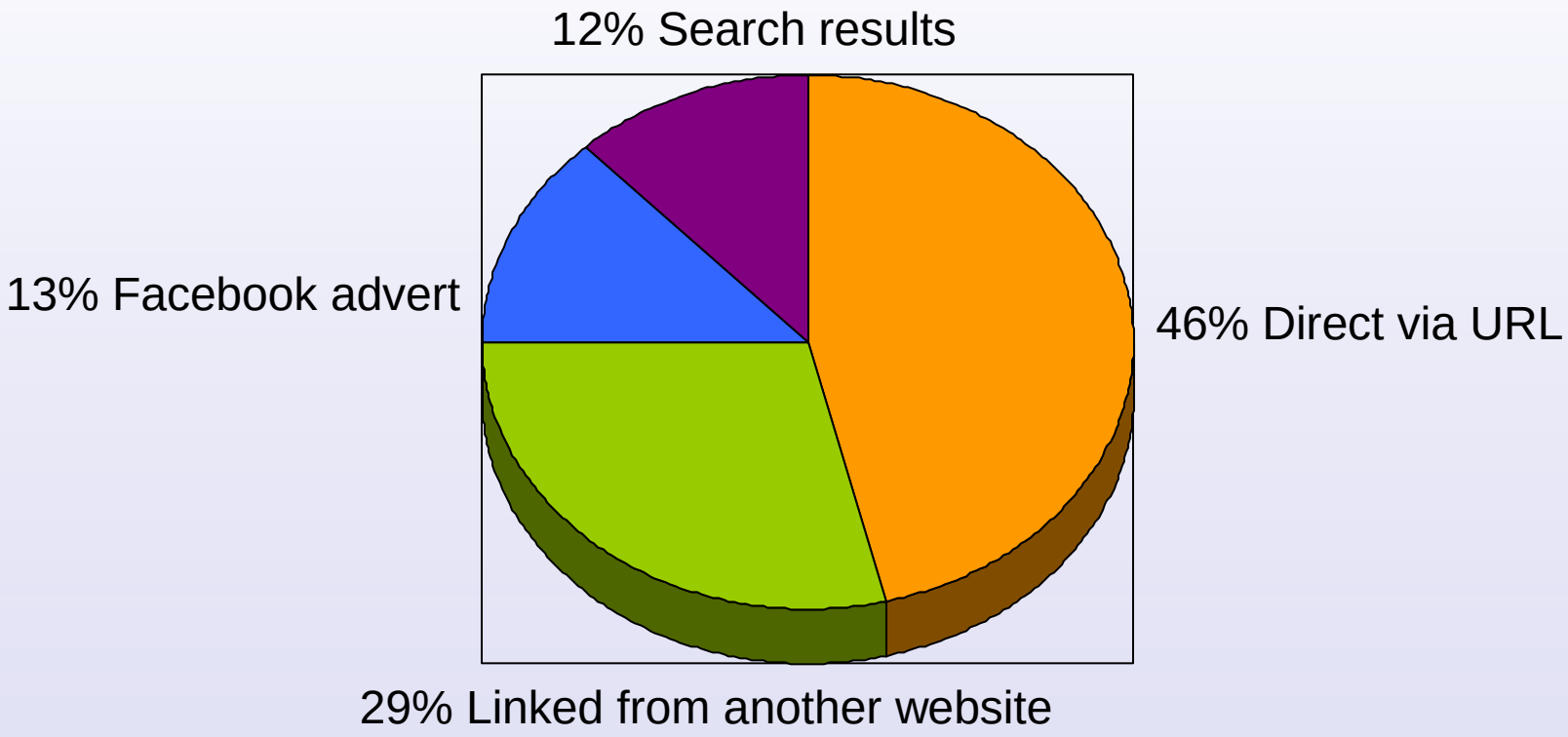
Attracting users

Target audience are English speakers who spend lots of time online

- Press releases
  - local and national press and radio
  - science journals and websites
- Writing on related blogs
- Bookmarking websites (digg, del.icio.us etc)
- Gaming forums
- Pay per click advertising on Facebook ($5 per day budget)

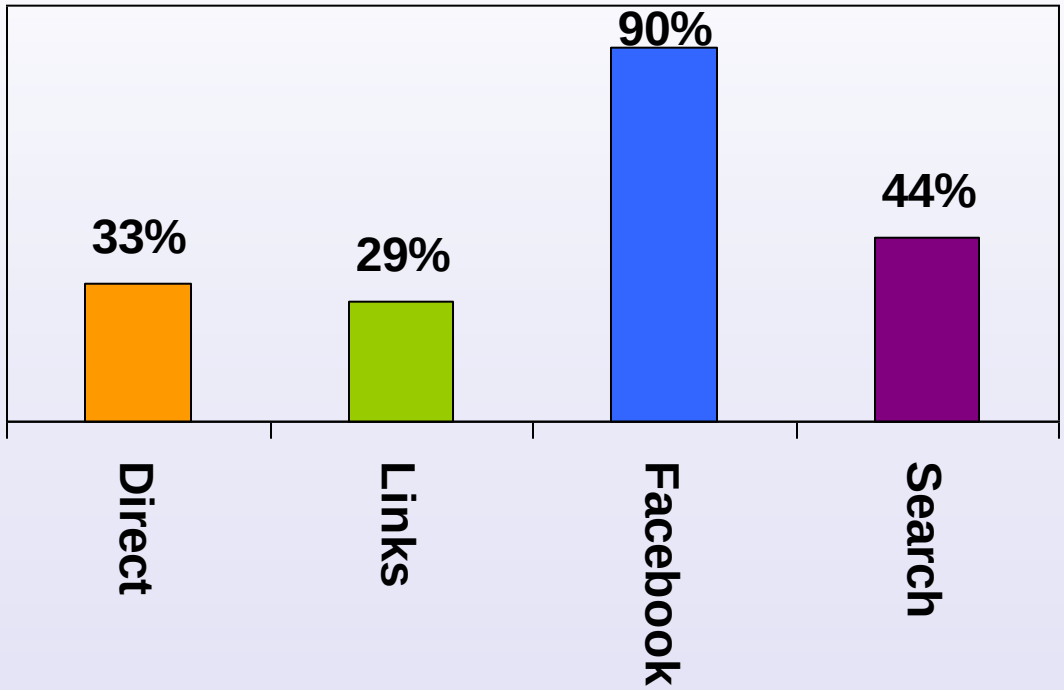
Attracting users

Traffic analysed by Google Analytics (Feb 2009): Incoming traffic



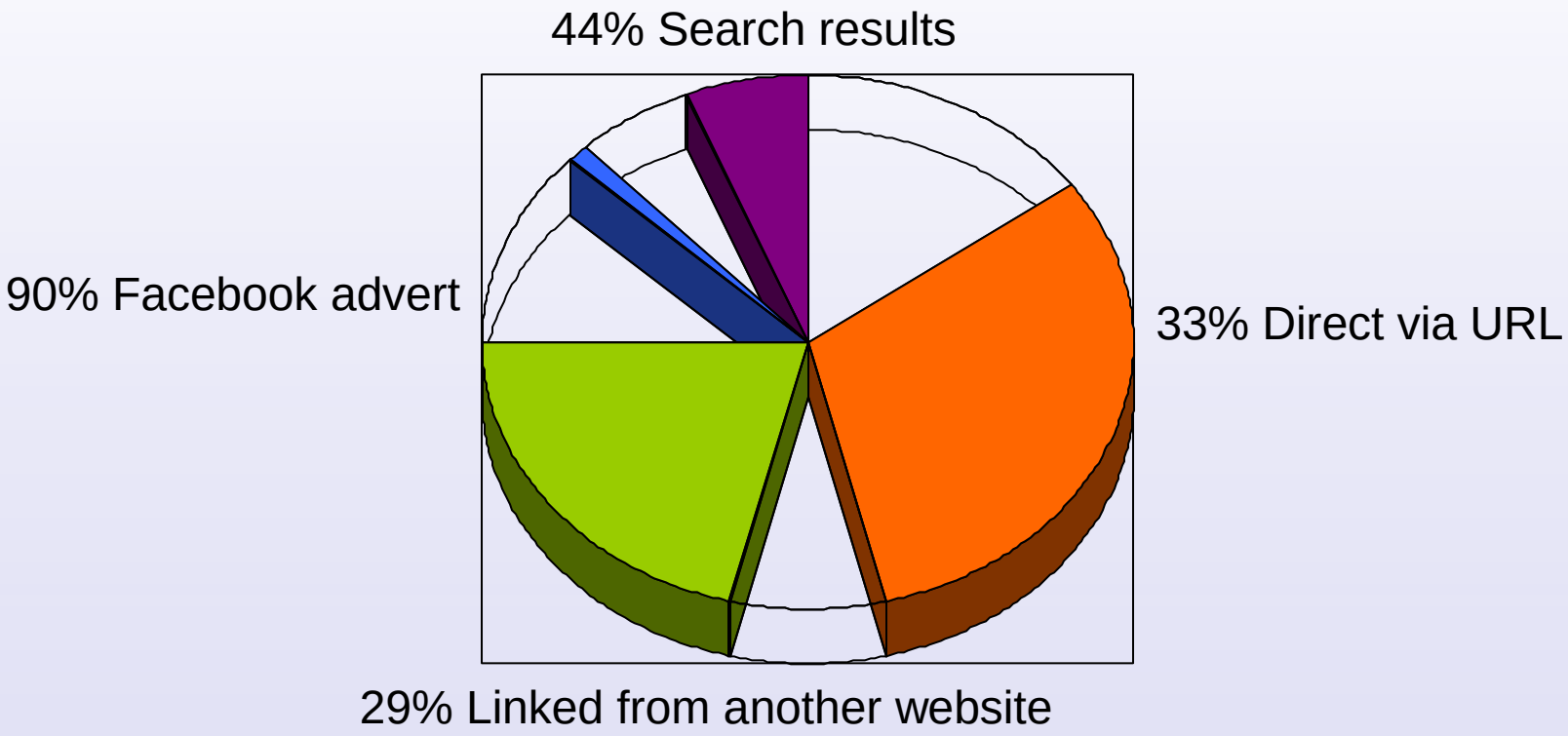
Attracting users

Bounce rate
- single page visits where the user leaves on the page they entered



Attracting users

Bounce rate
- single page visits where the user leaves on the page they entered



New sources of players

New source of traffic: prize directories (March 2009)

	% of site traffic	Bounce rate
www.loquax.co.uk	9%	40%
www.contesthound.com	9%	60%
www.theprizefinder.com	5%	67%
Facebook advert (Feb 2009)	12%	90%

Money spent on pay-per-click advertising might be better spent on prizes

Evaluating annotations

- Data expected to be noisy
  - Annotations are timed – too fast or too slow can be discarded
  - Annotations from users with low ratings can be discarded

Evaluating annotations

- Data expected to be noisy
  - Annotations are timed – too fast or too slow can be discarded
  - Annotations from users with low ratings can be discarded
- Do players from different sources provide higher quality data?
  - Current top players have language backgrounds

Evaluating annotations

- Data expected to be noisy
  - Annotations are timed – too fast or too slow can be discarded
  - Annotations from users with low ratings can be discarded
- Do players from different sources provide higher quality data?
  - Current top players have language backgrounds
- Ambiguity can be explored
  - Implicit ambiguity (several annotators create different relations for the markable)
  - Explicit ambiguity (an annotator creates several different relations for the markable)
- Plausible interpretations | Anaphoric chains | Incorrect decisions

Evaluating annotations: Easy ones!

The Pet Shoppe (Monty Python)

A customer enters a pet shop. Customer: 'Ello, I wish to register a complaint. The owner does not respond. C: 'Ello, Miss?

Owner: What do you mean "miss"?

Pause

C: I'm sorry, I have a cold. I wish to make a complaint!

O: We're closin' for lunch.

C: Never mind that, my lad. I wish to complain about this parrot what I purchased not half an hour ago from this very boutique.

	RelType	AnteID	Annotations	Agree	Disagree	Total	Expert
Viewing >>	DN		8	0	0	8	

Evaluating annotations: Easy ones!

The Pet Shoppe (Monty Python)

A customer enters a pet shop. Customer: 'Ello, I wish to register a complaint. The owner does not respond. C: 'Ello, Miss?

Owner: What do you mean "miss"?

Pause

C: I'm sorry, I have a cold. I wish to make a complaint!

	RelType	AnteID	Annotations	Agree	Disagree	Total	Expert
Viewing >>	DO	320535	8	0	0	8	

<< previous | next >>

Evaluating annotations: Easy ones!

The Pet Shoppe (Monty Python)

A customer enters a pet shop. Customer: 'Ello, I wish to register a complaint. The owner does not respond. C: 'Ello, Miss?

Owner: What do you mean "miss"?

Pause

C: I'm sorry, I have a cold. I wish to make a complaint!

O: We're closin' for lunch.

	RelType	AnteID	Annotations	Agree	Disagree	Total	Expert
Viewing >>	DO	320581	14	4	0	18	
View >>	PR	320581	1	0	4	-3	
View >>	DN		2	0	4	-2	
View >>	PR		1	2	2	1	

Evaluating annotations: Ambiguous ones!

The Pet Shoppe (Monty Python)

A customer enters a pet shop. Customer: 'Ello, I wish to register a complaint. The owner does not respond. C: 'Ello, Miss?

Owner: What do you mean "miss"?

Pause

C: I'm sorry, I have a cold. I wish to make a complaint!

O: We 're closin' for lunch.

	RelType	AnteID	Annotations	Agree	Disagree	Total	Expert
View >>	DO	320637	3	2	2	3	
View >>	DN		6	3	1	8	
Viewing >>	DO	320490	2	4	0	6	

Evaluating annotations: Ambiguous ones!

The Pet Shoppe (Monty Python)

A customer enters a pet shop. Customer: 'Ello, I wish to register a complaint. The owner does not respond. C: 'Ello, Miss ?

	RelType	AnteID	Annotations	Agree	Disagree	Total	Expert
View >>	PR	320534	1	4	0	5	
Viewing >>	DO	320534	6	4	0	10	
View >>	DN		1	4	0	5	

<< previous | next >>

Quality of data

- Ongoing analysis of completed documents
  - A selection of completed documents are also annotated by experts
- We compare
  - Expert vs game (top answer)
  - Expert vs expert
  - Player (top 5 all-time) vs game (top answer)

Quality of data

- Ongoing analysis of completed documents
  - A selection of completed documents are also annotated by experts
- We compare
  - Expert vs game (top answer)
  - Expert vs expert
  - Player (top 5 all-time) vs game (top answer)
- Initial results show agreement is consistent between experts and top game answers
- We anticipate validation will produce negative scores for poor annotations
- Filtering of annotations (by time or player rating) will be the second stage of analysis

Corpus

- Over 1 million words in the live game
- Sources of text include:
  - Fiction e.g. Project Gutenberg
  - Non-fiction e.g. Wikipedia
  - Non-traditional e.g. ENRON Corpus (emails)
  - Dialogue e.g. Film and TV scripts
  - Existing corpora e.g. GNOME and ARRAU
- Automatic detection of markables has proved non-trivial
  - Compatibility of processors
  - Inconsistent source data
  - Size of corpus
  - However, a pipeline for text processing is now available
- Over 200,000 annotations and validations

Conclusions

- Web-based games can be used to do huge computation tasks if users are incentivised and motivated
- The incentive structures of Phrase Detectives motivated users to provide large quantities of high quality annotations
- Collaborative and social elements hold the most promise especially if linked with existing social networks
- Financial rewards are a better use of money than pay-per-click advertising
- More innovative incentive structures will be needed as this methodology becomes more popular
- Will there be a point of saturation when Web users will no longer contribute to projects?

Future work

- Continue selecting and processing the corpus text (aiming for 100M words) including Italian, German and Spanish texts.
- Develop new tasks to create a community of users who can self-manage the game
- Analyse the data to assess the quality
- Ultimately, show that anaphora resolution algorithms perform better when trained on the data
- Create a generic framework to use this methodology for other NLP tasks e.g machine translation, generation, summarisation
- Data will be made available at the www.anaphoricbank.org

Play the game at:

www.phrasedetectives.com