



Phrase Detectives: Understanding Language With Games

DALI End of Project Summary – Sept 2021

Jon Chamberlain

Addictive Games

A person with long blonde hair is seen from behind, wearing a large black headset with a microphone. They are sitting in front of a computer monitor displaying a colorful, top-down strategy game, likely League of Legends. The game shows a battlefield with various characters and structures. The person's hands are visible at the bottom, resting on a desk. The background is dark and out of focus.

By age 21, the average American has spent more than 10,000 hours playing video games, equivalent to five years of working a full-time job.

Marc Prensky, CEO and founder Games2train.com

Games with a Purpose

What if that energy was repurposed?

The ESP Game

200,000 players, 50 million labels
in 2 months

Purchased by Google to improve
image labelling of search results

Luis von Ahn, co-founder of Captcha and DuoLingo



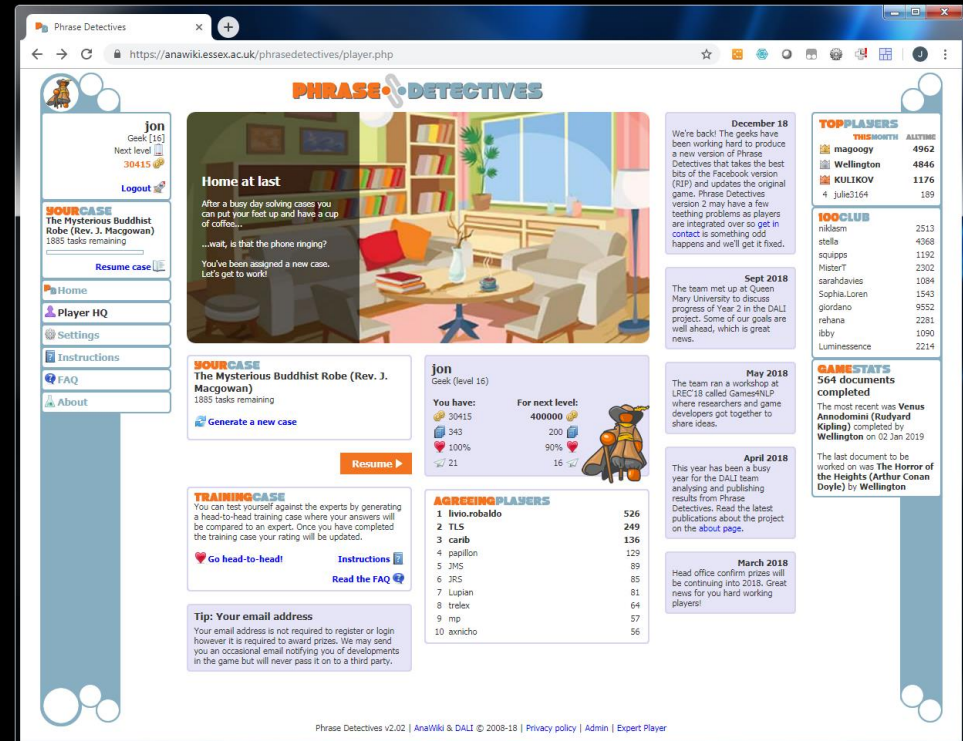
Phrase Detectives

Players annotate linguistic features of a text

Players also validate the opinions of other players

Tested methods of player motivation using game elements and prizes

Developed methods to minimise cheating and poor performance



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*, 3(1):1–44.

Phrase Detectives

Live for over 14 years

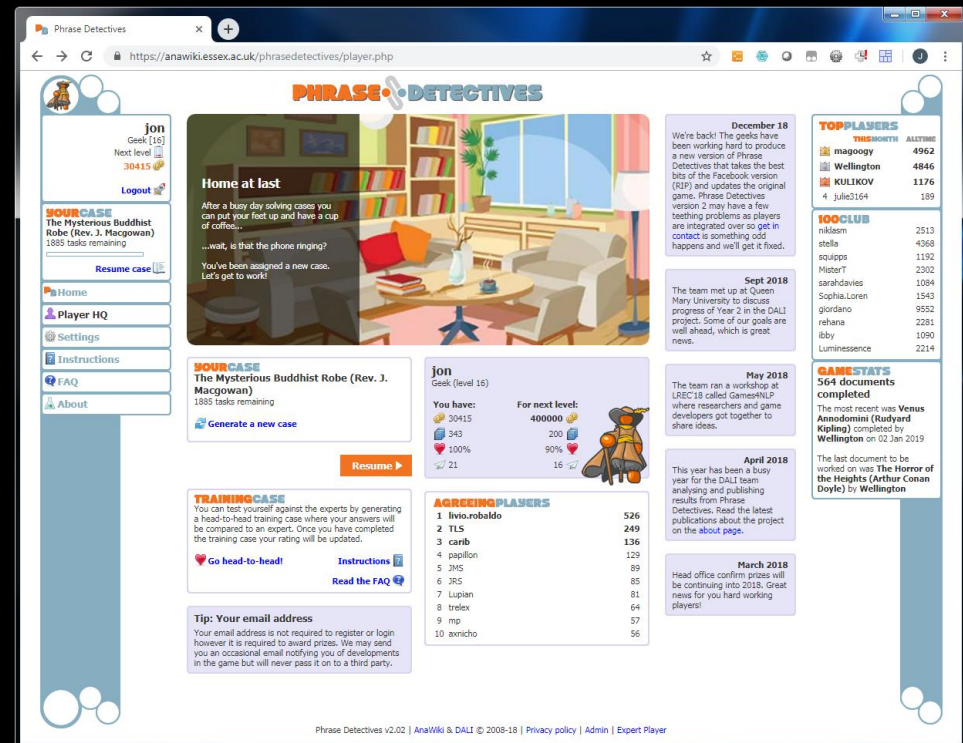
1.1M words in the game

598 docs completed
= 430k words

Over 5.2M decisions

61.3k players

11k human hours of work
= 460 days!



Phrase Detectives

Phrase Detectives corpus v2.0

2,235,664 judgments from
1958 players, of which:

1,358,559 annotations and
867,844 validations.

20.6 judgments per markable

Compared to:

600K judgments for
Ontonotes (~3 per markable)
10M judgments for PRECO
(also ~3 per markable)

		Docs	Tokens	Markables
C _{gold}	Gutenberg	5	7536	1947 (1392)
	Wikipedia	35	15287	3957 (1355)
	GNOME	5	989	274 (96)
	Subtotal	45	23812	6178 (2843)
C _{silver}	Gutenberg	145	158739	41989 (26364)
	Wikipedia	350	218308	57678 (19444)
	Other	2	7294	2126 (1339)
	Subtotal	497	384341	101793 (47147)
All	Total	542	408153	107971 (49990)

Poesio, M., Chamberlain, J., Paun, S., Kruschwitz, U., and Yu, J. (2019 forthcoming). "A Crowdsourced Corpus of Multiple Judgments and Disagreement on Anaphoric Interpretation." In Proceedings of NAACL-HLT 2019, Minneapolis. Association for Computational Linguistics.

Data Quality: An Initial Investigation

Quality (majority voting) is high compared to experts

Can be improved with filtering

	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%	94.1%	89.4%
	($\kappa=0.93$)	($\kappa=0.88$)	($\kappa=0.88$)

Agreement between experts

	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)

Agreement between experts and the majority-voted answer from players

Class Difficulty and Distribution

Contextual difficulty
Readability and document
length do not impact
accuracy.

Interpretation difficulty

- DN markables are common and easy to identify.
- DO are less common and harder to identify.
- NR rare but easy to identify.
- PR uncommon and difficult to identify.

	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%

Accuracy of crowd based on class

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

Distribution of class within corpora

Incorrect vs Ambiguous (Maj Vote)

	n	mn	sd	min	med	max	intersect	overlap
G1 $A + V_a - V_d$							5.6	23.6%
Gold standard	1,814	10.5	4.3	-3	10	32		
Incorrect	1,979	0.8	3.9	-3	0	20		
Possible	9	2.7	5.5	-3	2	12		
Same Entity	346	-0.2	3.7	-3	-1	15		
G1 $A_6 + V_a$							4.6	30.4%
Gold standard	1,760	6.4	1.7	1	6	9		
Incorrect	1,553	2.8	1.8	1	2	9		
W1 $A + V_a - V_d$							5.1	34.3%
Gold standard	3,537	8.6	3.8	-3	8	28		
Incorrect	4,027	2.0	4.8	-3	1	29		
Possible	28	1.0	3.9	-3	0	14		
Same Entity	395	-0.6	2.8	-3	-1	14		
W1 $A_6 + V_a$							4.4	47.6%
Gold standard	3,300	5.8	1.4	1	6	9		
Incorrect	2,538	3.4	2.1	1	3	9		

Aggregation / Ambiguity

The background of the slide is a dense crowd of stylized human figures. Most figures are dark grey or blue, but one figure in the center is glowing bright yellow, standing out from the crowd. The figures are arranged in a way that suggests a large gathering or a crowd.

Majority voting produces an answer set comparable to expert
Few systems have probabilistic answer set with ambiguity
Hard to distinguish a correct minority opinion from an error

Bayesian Models of Annotation

A Bayesian model of annotation specifies the probability of a particular label on the basis of parameters specifying the behavior of the annotators, the prevalence of the labels, etc.

Metrics:

- Annotator accuracy
- Item difficulty
- Item distribution

Paun, S., Carpenter, B., Chamberlain, J., Hovy, D., Kruschwitz, U., and Poesio, M. (2018). "Comparing bayesian models of annotation." Transactions of the Association for Computational Linguistics.

Mention Pair Annotation (MPA)

Anaphoric information, in which the 'labels' are not a discrete set, but coreference chains.

On the Phrase Detectives v2.0, it achieves an accuracy of 91% (as opposed to 84% for Majority Voting)

To improve we need to understand where the interface fails to capture user intent and allow the user to highlight interesting linguistic phenomena.

Paun, S., Chamberlain, J., Kruschwitz, U., Yu, J., and Poesio, M. (2018). "A probabilistic annotation model for crowdsourcing coreference." In Proceedings of EMNLP18, pages 1926–1937, Brussels, Belgium. Association for Computational Linguistics.

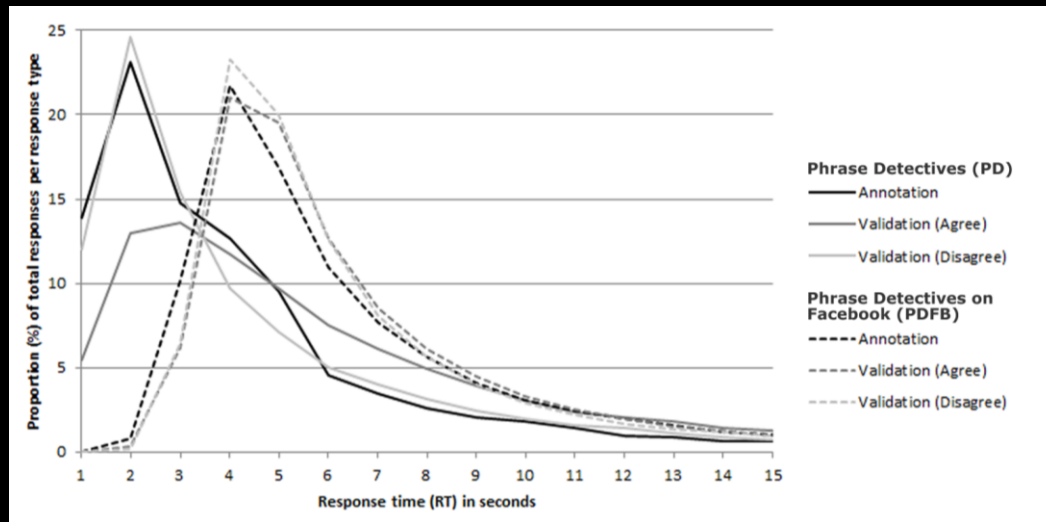
Physical Performance Indicators



1. input processing (sensory processing) where the user views the input (text or image) and comprehends it;
2. decision making (cognitive processing) where the user makes a choice about how to complete the task;
3. taking action (motor response) to enter the response into the system interface (typically using a keyboard or mouse).

User Performance Indicators In Task-based Data Collection Systems.
Chamberlain & O'Reilly, 2014.
Proc. MindTheGap'14, Berlin.

Physical Performance Indicators



	Correct	Incorrect
Annotation*	10.1s	12.8s
Validation (Agree)*	13.5s	17.7s
Validation (Disagree)	14.5s	15.0s

Incorrect answers take longer but there a large amount of fast spam responses.

The interface plays a major role in response times.

Prior Knowledge and Specificity

Different levels of knowledge held by the users can create ambiguous results, eg

Sun birds = Andean condor

How to mark up knowledge that reveals itself through discourse, eg the butler conundrum

Mr Smith = the butler = the murderer

How to mark up specificity, eg:

John, the king of England ...

King John, at age 21, ...

After being dethroned, **John** ...

User Interface Restrictions

Restricted context of the system in order to make it game like and not an annotation tool, eg:

- No cataphors (selecting beyond the antecedent)
- 1000 character context
- Cant create or edit markables

Limiting expressions makes the processing easier and reduces errors but we miss useful information

Constraining Inputs

Speaking Outside the Box: Exploring the Benefits of Unconstrained Input in Crowdsourcing and Citizen Science Platforms

Jon Chamberlain | Udo Kruschwitz | Massimo Poesio

LREC workshop for Citizen Linguistics in Language Resource Development (2020)



Modes of data collection

To create a game-like experience the text is pre-processed and inputs are constrained.

Rhinogradentia (Wikipedia)

Rhinogradentia (also known as snouters or Rhinogrades or Nasobarnes) is a fictitious mammal order documented by the equally fictitious German naturalist Harald Stumpke. The order's most remarkable characteristic was the Nasorium, an organ derived from the ancestral species's nose, which had variously evolved to fulfill every conceivable function.

Both the animals and the scientist were allegedly creations of Gerolf Steiner, a zoology professor at the University of Karlsruhe. A mock taxidermy of a certain Snouter can be seen at the Musee zoologique in Strasbourg.

The order's remarkable variety was the natural outcome of evolution acting over millions of years in the isolated Hi-yi-yi islands in the Pacific Ocean.

NAME THE CULPRIT

Has the phrase shown in orange been mentioned before in this text or is it a property of another phrase? Select the closest phrase(s) within the text if it has been mentioned before and click "Done".



Not mentioned before



This is a property



Done



Rhinogradentia (Wikipedia)

Rhinogradentia (also known as snouters or Rhinogrades or Nasobarnes) is a fictitious mammal order documented by the equally fictitious German naturalist Harald Stumpke. The order's most remarkable characteristic was the Nasorium, an organ derived from the ancestral species's nose, which had variously evolved to fulfill every conceivable function.

Both the animals and the scientist were allegedly creations of Gerolf Steiner, a zoology professor at the University of Karlsruhe. A mock taxidermy of a certain Snouter can be seen at the Musee zoologique in Strasbourg.

The order's remarkable variety was the natural outcome of evolution acting over millions of years in the isolated Hi-yi-yi islands in the Pacific Ocean.

DETECTIVE CONFERENCE

Another detective has said the phrase in orange has been mentioned before and its nearest mention is highlighted in blue. Do you agree with them?



Disagree

Agree



Constrained annotation mode

Constrained validation mode


User Interface Restrictions

Players try to express themselves using the limited interface eg discourse diexis







The Andean Condor is a scavenger, feeding mainly on carrion. Wild condors inhabit large territories, often traveling more than 200 km (100 miles) a day in search of carrion. In inland areas, they prefer large carcasses, such as those of dead farm animals or wild deer, while their diet consists mainly of beached carcasses of marine mammals when near the coast. It is no use trying to catch one. **This behaviour** is typical of condors.

Comments and skips also allow user expressions to some extend

Modes of data collection

 Comment on this phrase

Submit comment

-  Skip - error in the text
-  Skip this one
-  Skip - closest phrase is no longer visible
-  Skip - closest phrase can't be selected
-  Skip - this is discourse deixis
-  Skip - this is a quantifier

Semi-constrained input

Freetext field for comments

Constrained "skip" reasons

Allow for error detection and capture of unknowns

Capturing the Unknowns

Classification	Skip	Comments
Not selectable	[5]	31,846
Out of context window	[4]	21,732
Parse error	[2]	15,707
Discourse deixis	[6]	328
Ambiguous		49
Non-referring		24
Nearest mention embedding		237
Bridging reference		11
Quantifier	[7]	50
Unclassified		6,899
TOTAL		76,883

Errors in the pre-processing

Interface limitation prevented some markables from being selectable:

- Embedded in another markable
- No longer in the selectable document content

Capturing the Unknowns

Classification	Skip	Comments
Not selectable	[5]	31,846
Out of context window	[4]	21,732
Parse error	[2]	15,707
Discourse deixis	[6]	328
Ambiguous		49
Non-referring		24
Nearest mention embedding		237
Bridging reference		11
Quantifier	[7]	50
Unclassified		6,899
TOTAL		76,883

Discourse deixis (DD) and Quantifiers (QQ) were detected by users and used the comments to tell us.

Interface forces user to make a single decision but they can also explicitly state the label is ambiguous.

Constrained vs Unconstrained

- A constrained decision may be easiest to process but may not be sufficient to answer the question/task
- Participants in citizen science want their contribution to be valued and get frustrated if they cannot express themselves
- It is the most difficult and unusual tasks that intrigue and motivate participants...

...and these are the most value for future systems to improve the state of the art


Game Metrics

A person with long blonde hair is seen from behind, wearing a large black headset. They are looking at a computer monitor that displays a League of Legends game. The game shows a top-down view of a battlefield with various champions and structures. The person is wearing a dark-colored shirt.

- 1) Player focused
- 2) Community focused
- 3) Item (annotation) focused

Metrics of games-with-a-purpose for NLP applications. Chamberlain, Bartle, Kruschwitz, Madge & Poesio, 2017. Games4NLP Workshop, co-located at EACL17, Valencia.

Player Metrics

A humpback whale is captured in the middle of a breach, its massive body arched out of the dark blue ocean. The whale's head is at the top right, with its blowholes visible. Its pectoral fin is extended upwards and outwards on the left side, showing a white underside with black spots. The whale's back is dark with white mottled patterns. A large splash of water is visible around the whale's head and pectoral fin. The background is a clear blue sky.

How engaged are the players in the game?
How effective are advertising methods?
Is it better to focus on whales or minnows?

Player Metrics

Cost per Acquisition (CpA)

Lifetime Judgements (LTJ)

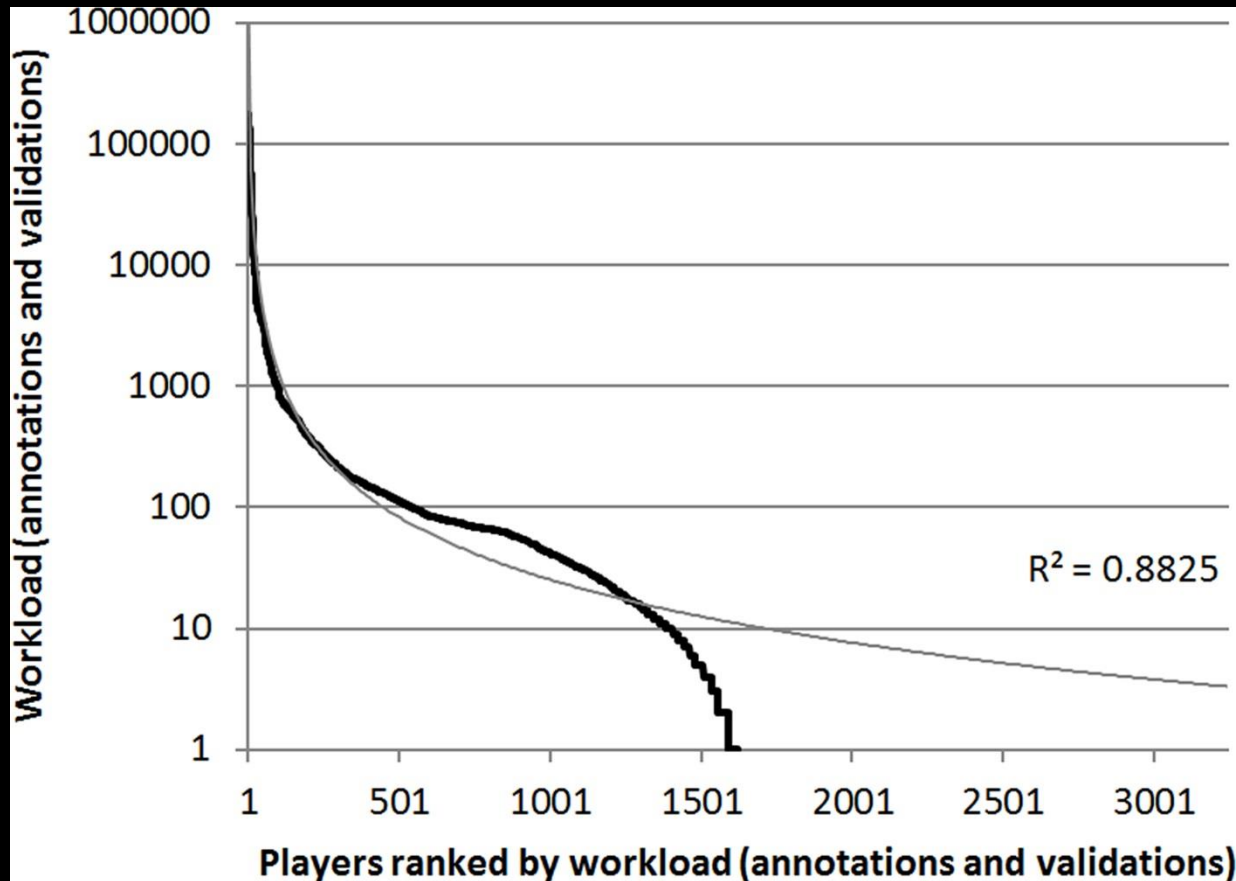
Average Judgements per Person (AJpP)

Average Lifetime Play (ALP)

Metrics to understand the interaction between the player, the platform and outside activity (eg advertising) over given time periods.

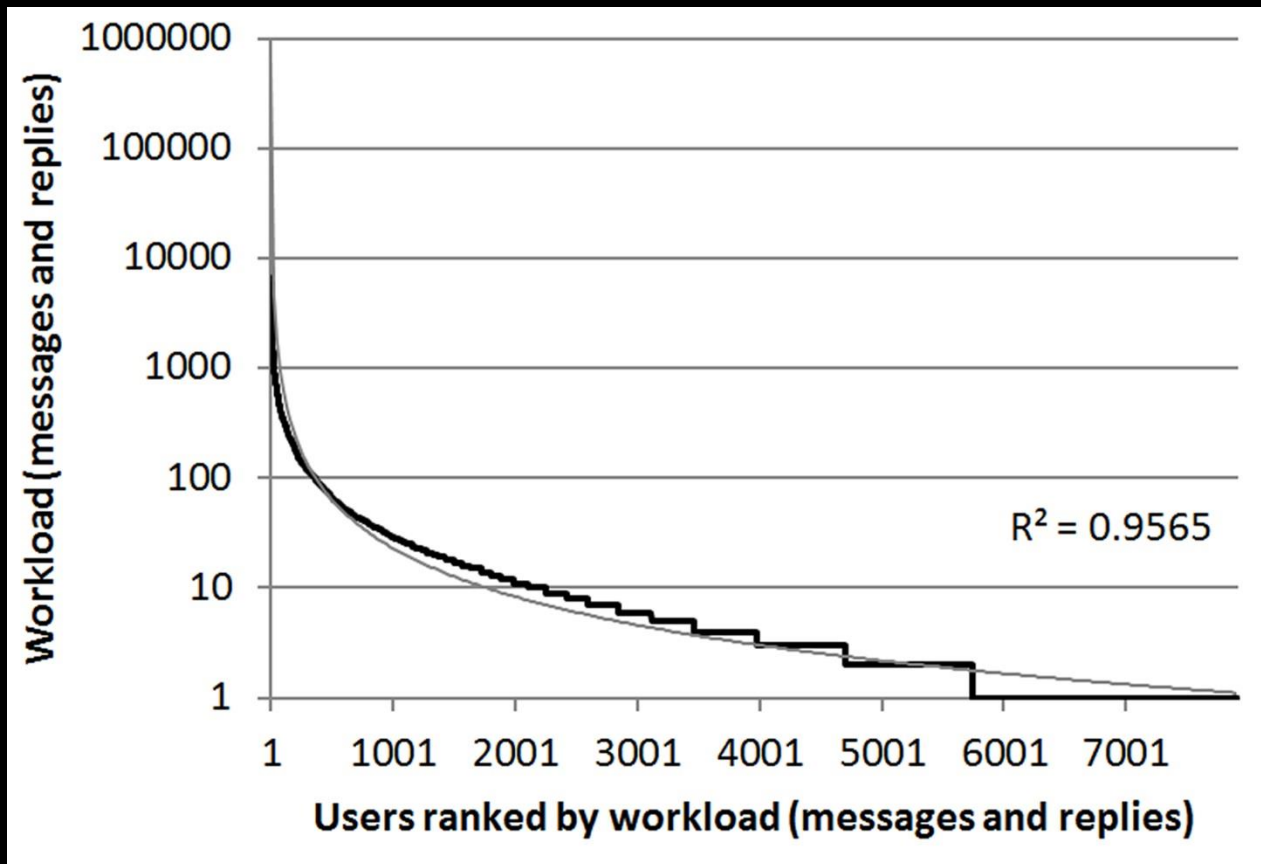
Whales vs Minnows

Ranked contribution in Phrase Detectives



Whales vs Minnows

Ranked contribution on social media groups



Player Metrics

A humpback whale is captured in the middle of a breach, its massive head and upper body emerging from the dark blue ocean. The whale's skin is dark with characteristic white patches and stripes. Its pectoral fin is visible, showing a white underside with dark spots. Water is splashing around the whale's head, and the sky is a clear, pale blue.

Workload/contribution follows a Zipfian distribution.
Very few users contribute most of the work/revenue.
This may be an issue if you need a diverse crowd.

Community Metrics

A close-up photograph of several hands of different skin tones stacked together in a huddle, symbolizing teamwork and community. The hands are positioned in the center of the frame, with fingers interlaced. A black wristwatch is visible on one of the wrists. The background is a plain, light-colored wall.

How fast is the game growing?
How "sticky" is the game (do players return)?
Are incentive methods working?

Community Metrics

Monthly Active Users (MAU)

Number of users who contribute in a calendar month.

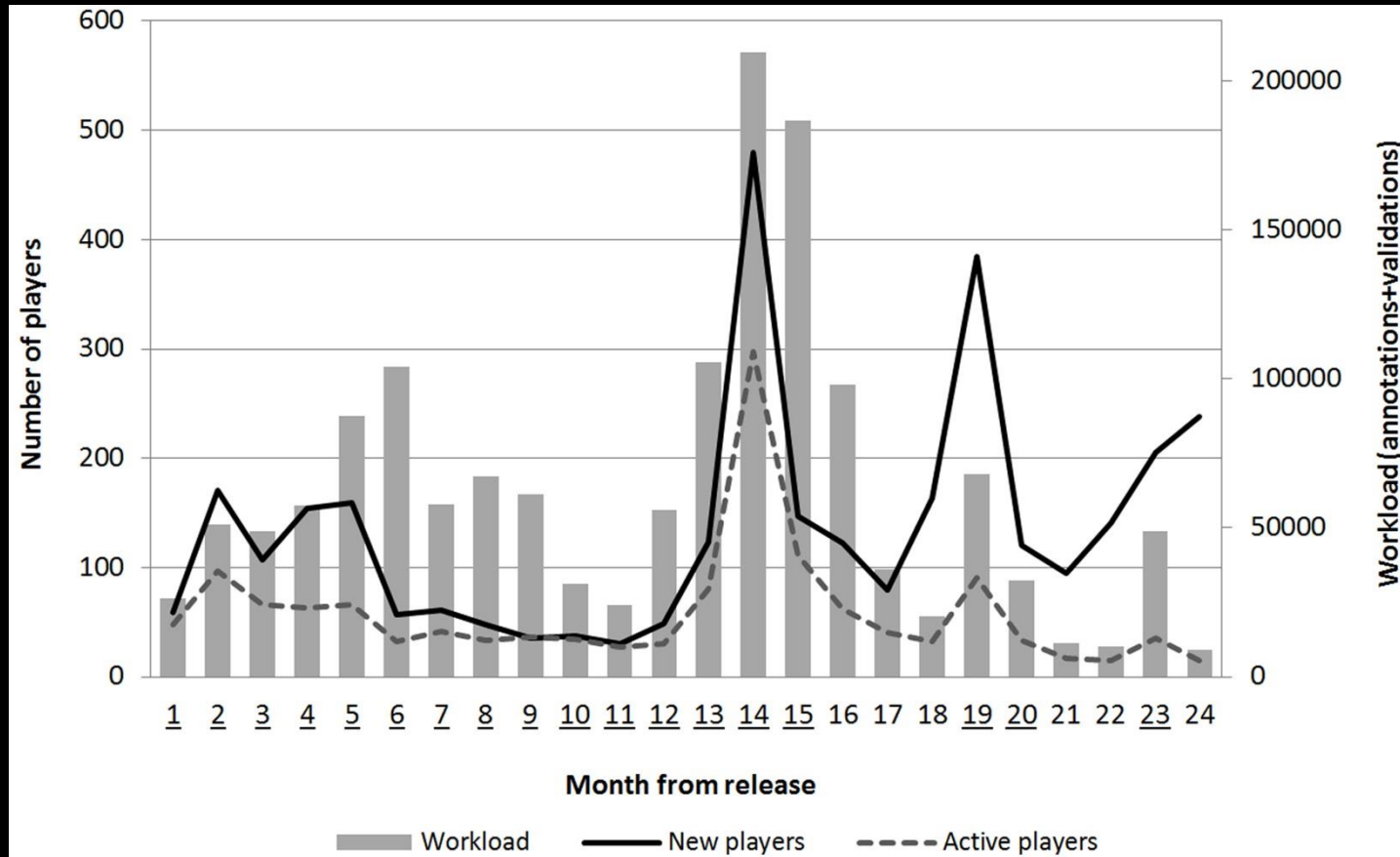
Definition of “active” varies.

Retention / Churn

Percentage of players who continue to play /

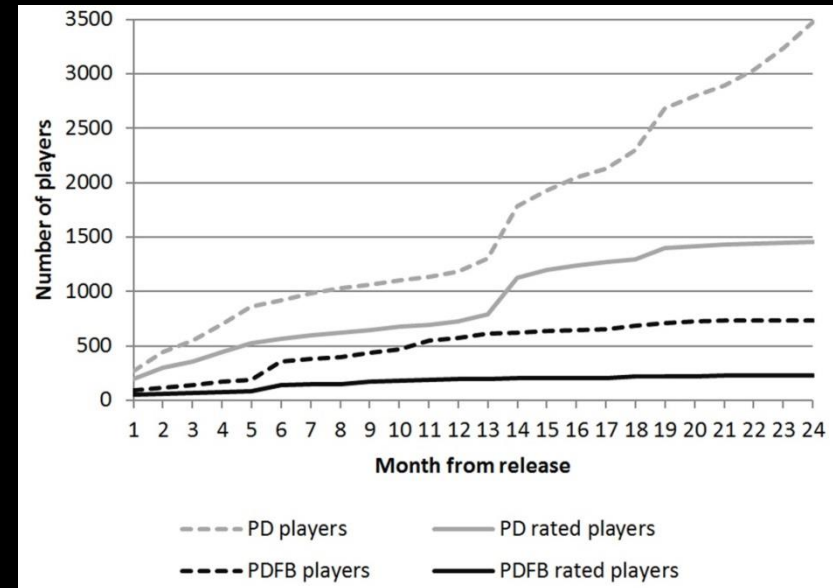
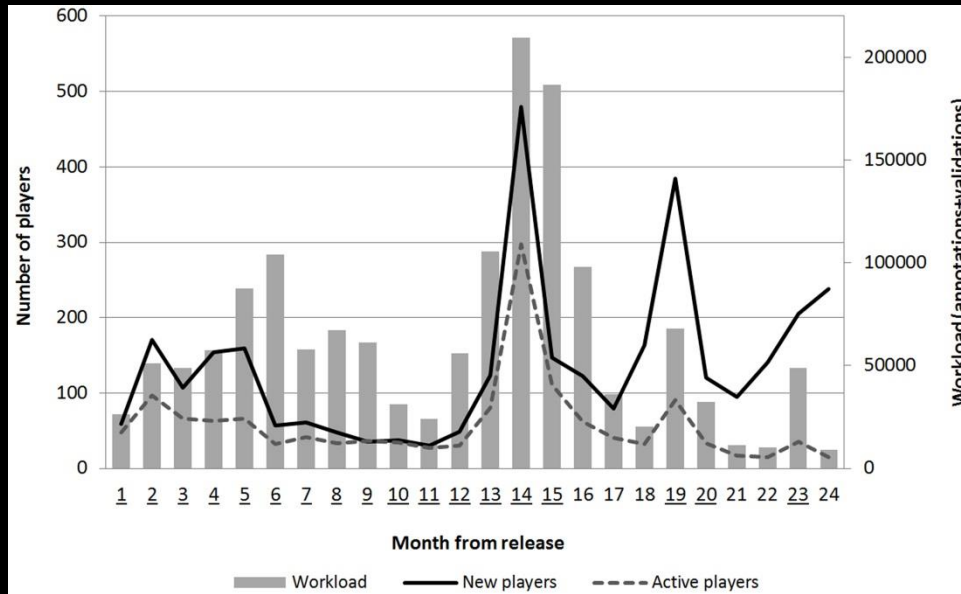
Percentage of players who stop playing

Community Metrics



Growth of Phrase Detectives in the first 2 years

Community Metrics



More informative than cumulative growth (right)
Player specific retention/churn for deeper analysis

Item Metrics

Can the system produce enough data fast enough?
How many players will you need?
Would another approach be better? (e.g., microworking)

Item Metrics

Cost per Judgement (CpJ)

Judgements Required (JR)

Cost per Item (CpI)

Throughput

Metrics indicating the overall performance of the system

Collaboration vs Collusion



Phrase Detectives ensured all decisions were independent.
Reduce collusion, bias and cheating to score points.
This allowed us to explore the dataset and test algorithms.

Collaboration vs Collusion



Moving towards directive models:

- We know the players better and give more challenges
- Players can direct us to interesting phenomena

Collaboration vs Collusion

A close-up photograph of several hands of different skin tones stacked together in a huddle, symbolizing teamwork and collaboration. The hands are positioned in the center of the frame, with fingers interlaced. A black wristwatch is visible on one of the wrists. The background is a plain, light-colored wall.

- Social contract of crowdsourcing and volunteering:
- Players give up their time and effort
 - They must be entertained and rewarded in return

What makes games fun?

Bartle, R. Hearts, Clubs, Diamonds, Spades: Players Who suit MUDs (1996)

Killers

Also known as “griefers”

Achievement comes from another person's loss

Value knowledge for its applications

Prize reputation and recognition



Achievers

Seek to improve power and status

Fun comes from points and leveling up.

Point of playing is to master the game

Enjoy recognition of their achievements



Acting



Interacting

Players



World

Socializers

Enjoy meaningful social interaction with other players

Point of playing is to make friends

Game is simply a backdrop

Enjoy recognition of their followers, contacts, influence



Explorers

Love to “figure out” games

Fun comes from discovery

Collectors of knowledge and little-known facts

Enjoy teaching others

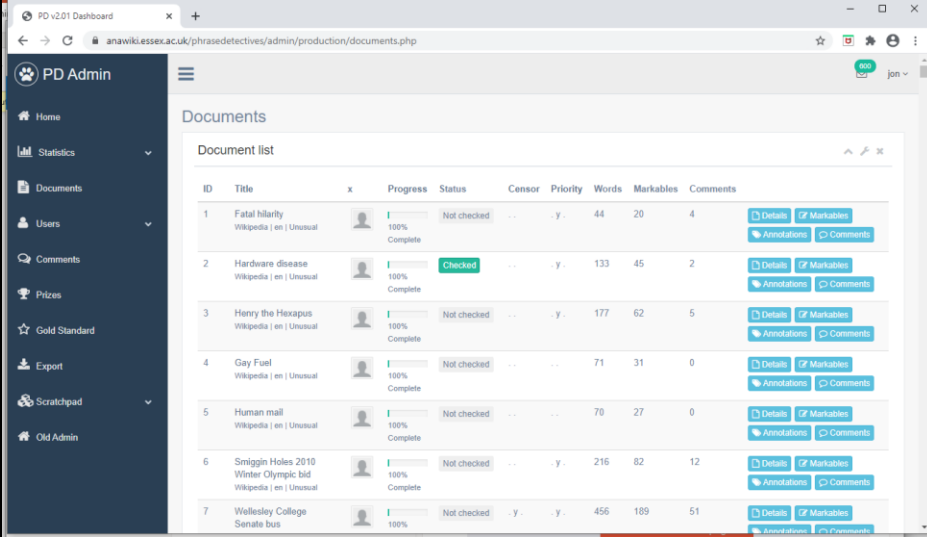


Phrase Detectives v2: Admin

A number of legacy admin functions being combined into a single admin function

Document page is now central to overview of document status

Moving away from the concept of "document completion" and more towards a directive/dynamic scheme



The screenshot shows the PD Admin interface. On the left is a sidebar with navigation links: Home, Statistics, Documents, Users, Comments, Prizes, Gold Standard, Export, Scratchpad, and Old Admin. The main content area is titled 'Documents' and contains a 'Document list' table. The table has columns for ID, Title, x, Progress, Status, Censor, Priority, Words, Markables, and Comments. Each row represents a document with its details and action buttons for Details, Markables, Annotations, and Comments.

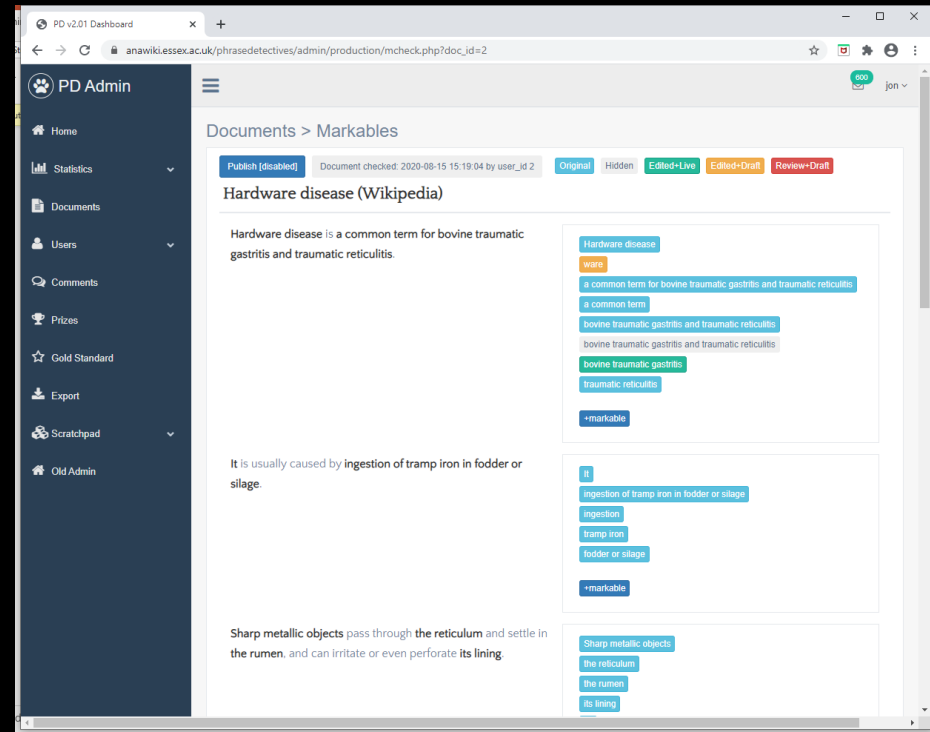
ID	Title	x	Progress	Status	Censor	Priority	Words	Markables	Comments
1	Fatal hilarity Wikipedia en Unusual		100% Complete	Not checked	...	y	44	20	4
2	Hardware disease Wikipedia en Unusual		100% Complete	Checked	...	y	133	45	2
3	Henry the Hexapus Wikipedia en Unusual		100% Complete	Not checked	...	y	177	62	5
4	Gay Fuel Wikipedia en Unusual		100% Complete	Not checked	71	31	0
5	Human mail Wikipedia en Unusual		100% Complete	Not checked	70	27	0
6	Smuggin Holes 2010 Winter Olympic bid Wikipedia en Unusual		100% Complete	Not checked	...	y	216	82	12
7	Wellesley College Senate bus		100%	Not checked	y	y	456	189	51

Phrase Detectives v2: Markables

Markable checker creates a draft version of markables with a change log.

Markable changes are sanity checked (don't exceed sentence boundaries or overlap other markables)

Character length based, not token based



Phrase Detectives v2: Markables

Markable edits can be consumed from other sources i.e., Tile Attack or Wormingo and added to draft (todo)

Draft changes are deployed live on publish. Sanity checked again to make sure nothing breaks (todo)

Impact on document needs processing ie do markables need to be done again?

Original Hidden Edited+Live Edited+Draft Review+Draft

Hardware disease

ware

Hardware disease is a common term for bovine traumatic gastritis and traumatic reticulitis.

ID: 465849 Min: 3 Start: End: Max: 94

Hide ☒ Divert Review ☐ Submit

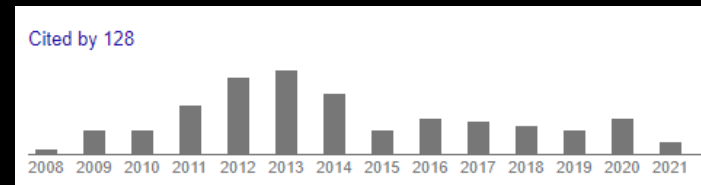
Timestamp	user	action	divert	span	review	status
2020-08-15 15:15:37	2	add		7-11		
2020-08-15 15:16:15	2	hide		7-11		
2020-08-15 15:16:27	2	hide	27	7-11		
2020-08-15 15:16:33	2	hide	27	7-11	y	
2020-08-15 15:18:59	2	hide		7-11		draft

Final Plans

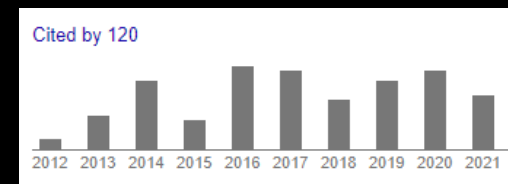
- Convert look and feel (game todo list)
- Finish the admin migration (admin todo list)
- Gold Standard creation admin tool
- Task allocation, document creation, annotation toolbar
- Games and NLP at LREC'22
- Shared database, communication between games.

Academic Impact

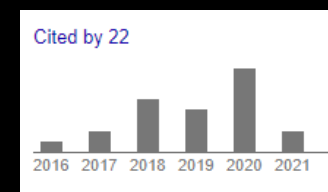
Chamberlain, J., Kruschwitz, U., and Poesio, M. (2008). **Phrase detectives: A web-based collaborative annotation game**. Proceedings of iSemantics08, Graz, Austria.



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, 3(1):1–44.



Chamberlain, J., Poesio, M., and Kruschwitz, U. (2016). **Phrase Detectives corpus 1.0: Crowdsourced anaphoric coreference**. In Proceedings of LREC, Portoroz, Slovenia.



Over 600 citations of Phrase Detectives papers...



Thanks for listening!
Now start playing...