



Metrics for Educational and Crowdsourcing Games

Jon Chamberlain | University of Essex | jchamb@essex.ac.uk
Massimo Poesio | Queen Mary University | m.poesio@qmul.ac.uk

Addictive Games



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

score
100

ESP Game

Concentrate...

time
2:21

What do you see?

taboo words

peace
lay



guesses

sheeps...
sheep

The ESP Game

200,000 players, 50 million labels in 2 months

Purchased by Google to improve image search results

<http://ael.gatech.edu/cs6452f13/files/2013/08/labeling-images.pdf>

+ submit → pass



PHRASE • DETECTIVES

USERPROFILE

jon

60 this week
4 decisions
31 agreements
25 extras

69 this month
28247 all time

Level: **Cunning Pirate**

Your rating: **96%**

CASE OPEN

89 tasks remaining

297 completed cases

[EDIT PROFILE](#) | [LOGOUT](#)

Like 88 likes. Sign Up to see what your friends...

WELCOME BACK TO HEADQUARTERS

Hello jon.

You've been doing a great job so far but you need to keep detecting.



Cunning Pirate

You have **28247** total points.

21753 points to the next level.



You have a case open

Noodling (Wikipedia)

[Generate new case](#)



Who agrees with you

livio.robaldo [526]

TLS [249]

carib [136]

papillon [124]

JMS [89]

JRS [85]

Lupian [81]

trelex [64]

aknicho [56]

johnnickel [49]



Your best recruits

jimbo [319]

Earn extra points by recruiting other detectives! [Send them here](#)

TOP SCORES

THIS WEEK
JRS 1100

THIS MONTH
stewart miller 5991

LEADERBOARD

WEEK	MONTH	ALERTING
JRS		1100
stewart miller		1086
MDKorpel		806
papillon		610
thomwd		543
IvoBril_RijkvanBraak		504
Folkert_Patrick_KI		402
StefanenHein		390
myrmidon		388
MAJ		367
VB		277
s2011840		252
AntonMulder		237
michelleburghardt		233
gully		229
julie3164		194
RB_NV_KI		180
MarcoBosman		178
MWithagen		166
rikanna		155

MOST RECENT

Noodling (Wikipedia)
submitted by jon

[Feedback](#)

Phrase Detectives

45k players submitted over 3.5 million labels in 8 years

Is this comparable to The ESP Game?

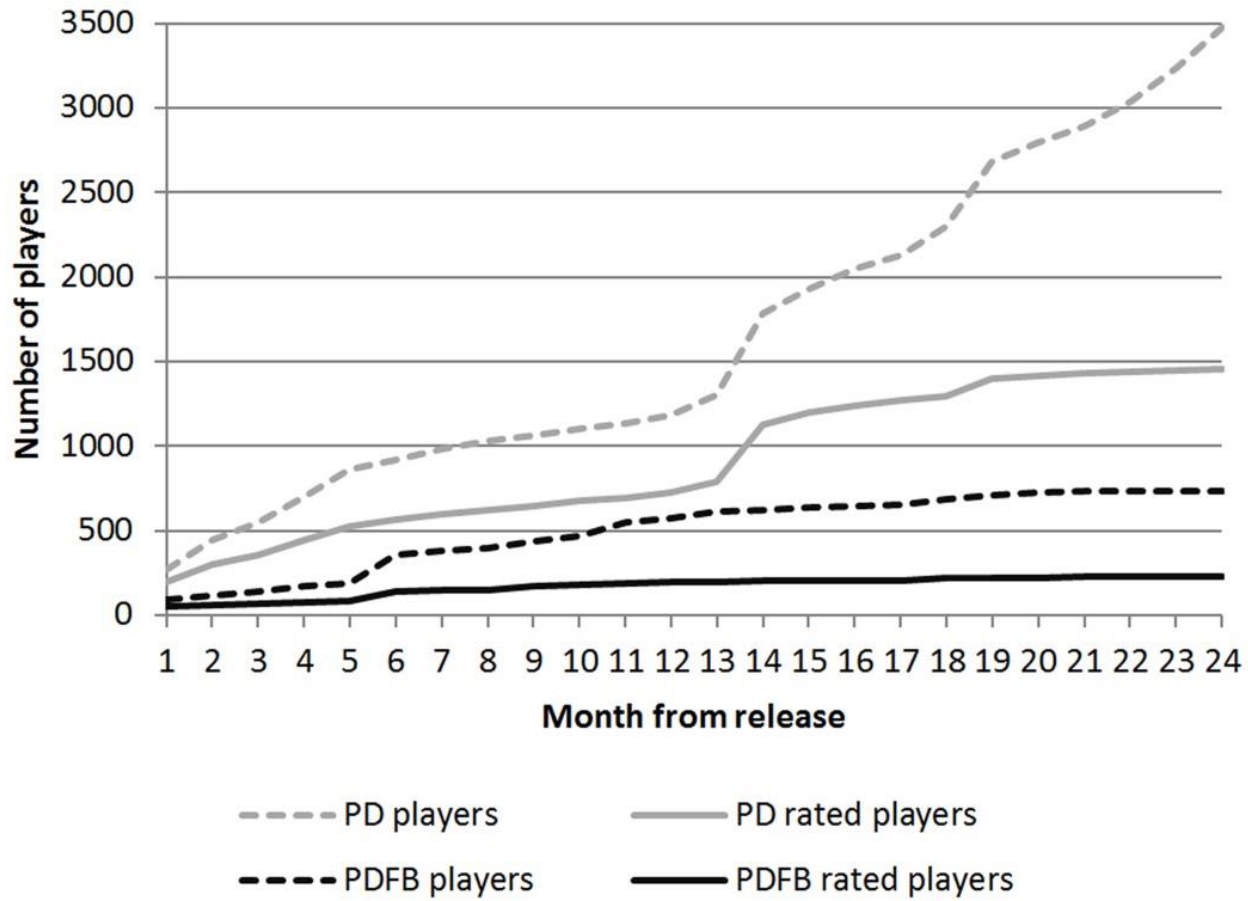
<https://anawiki.essex.ac.uk/phrasedetectives/>

SHARE THIS



[Start >>](#)





Phrase Detectives on Facebook

Far fewer players in the first 2 years of release.
 What are those players doing?



00:00
5/5

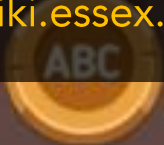
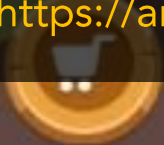
sFiz7LD%J4 cV82 rGjM5z !MKq v% fzKuu0
tw08p6 erk152!J ql5Elw00u0

RoboCorp

232 players in 2 months

64 played mini-game, 57 made in-game purchases

https://anawiki.essex.ac.uk/dali/games4nlp17/papers/03_Games4NLP_EACL17_Making.pdf



Spanish skills

Shop



Intro



Phrases



Travel

Crown Level



Daily Goal



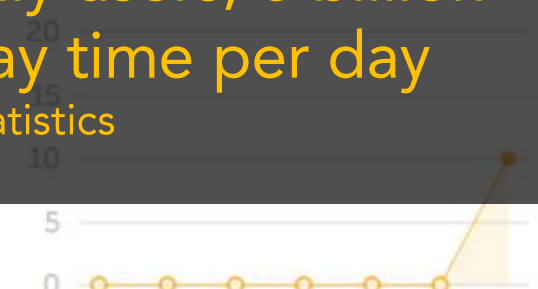
1 day streak

13 hours left

DuoLingo

200 million active users, 25 million monthly users, 6 billion exercises per month, 10 mins average play time per day

<https://expandedramblings.com/index.php/duolingo-facts-statistics>



Goal

Classify games based on their aims and user motivations

Define a set of metrics to compare games with similar aims

Adapt existing metrics to learn from the games industry (serious games, F2P, etc)

Game Classification

Classify games based on their aims and user motivations

What are the differences between a crowdsourcing game and an educational game?

Game Classification

	Crowdsourcing	Educational
Aim	Collect data	Educate users
Developer motivation	Convert an existing task for crowd	Teach/educate
Player motivation	Financial, social, personal	Personal (learning)
Task	Somewhat defined (annotation scheme for language)	Clearly defined (based on reading levels)
Progression	Somewhat defined (based on document difficulty)	Clearly defined (based on reading levels)
Solution	Some gold standard for training	Solutions known and presented to help learner
Learning	Side product	Direct product

Game Metrics

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



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.

Cost per Acquisition (CpA)

Cost to get a player to start playing the game

CpA = Advertising budget / New users

Spillover effect?

Viral games?

New vs active users?

Lifetime Judgements (LTJ)

=Customer Lifetime Value (CTV)

$CTV = \text{Revenue generated} - CpA$

LTJ = Total contribution to the game

Monetary value of contribution?

Time span between plays?

Same user, different accounts?

Average Judgements/Person (AJpP)

=Average Revenue Per User (ARPU)

ARPU = Total revenue / Total active users

AJpP = Average Judgements per Player

= Total judgements / Total active players

Account for Zipfian distribution of work?


Average Lifetime Play (ALP)

ALP = How long players continue to contribute

What is the definition of lifetime?

Contribution time vs actual time

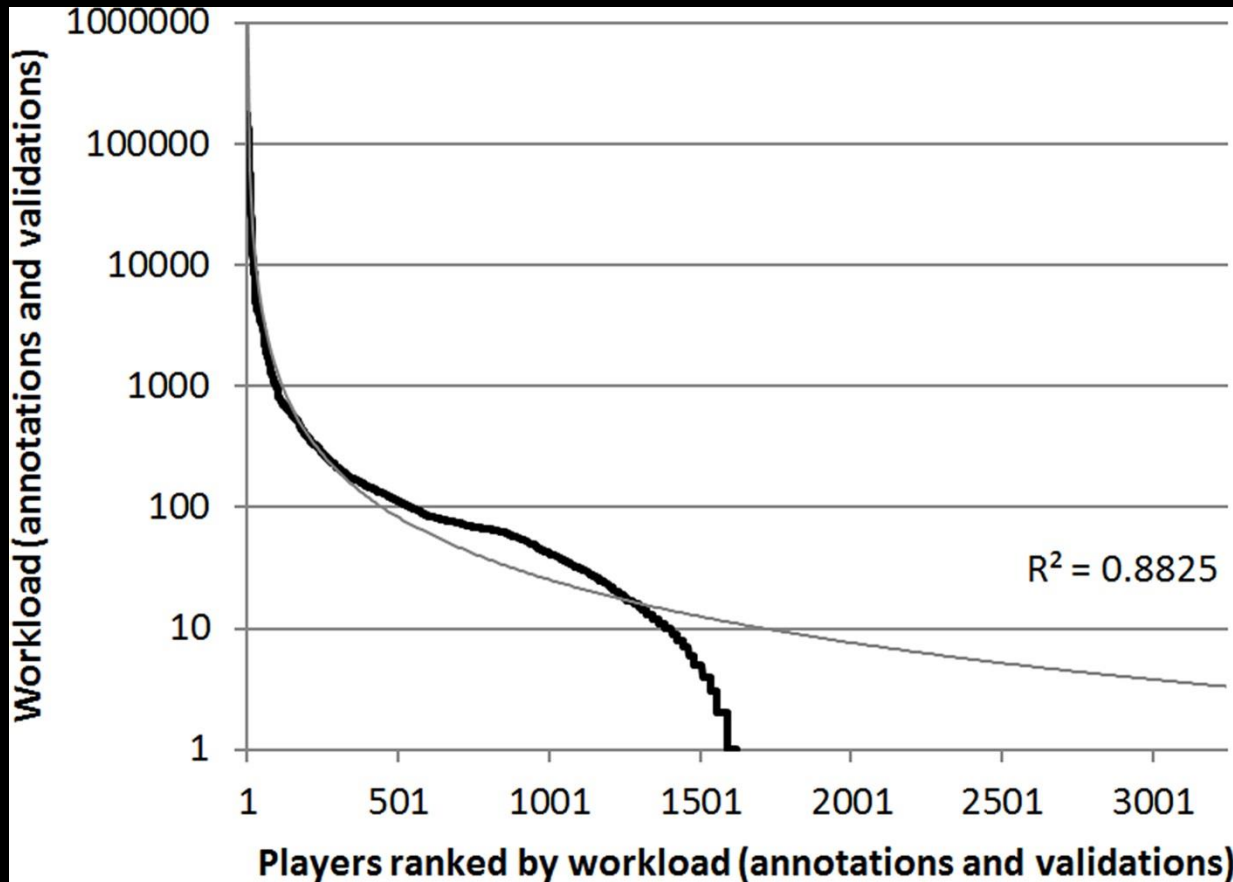
Player Metrics

A humpback whale is captured in the middle of a breach, leaping out of the dark blue ocean. The whale's dark, ribbed back and white, mottled pectoral fin are prominent against the sky. Water is splashing around the whale's head and tail. The background shows a clear blue sky and the horizon line.

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

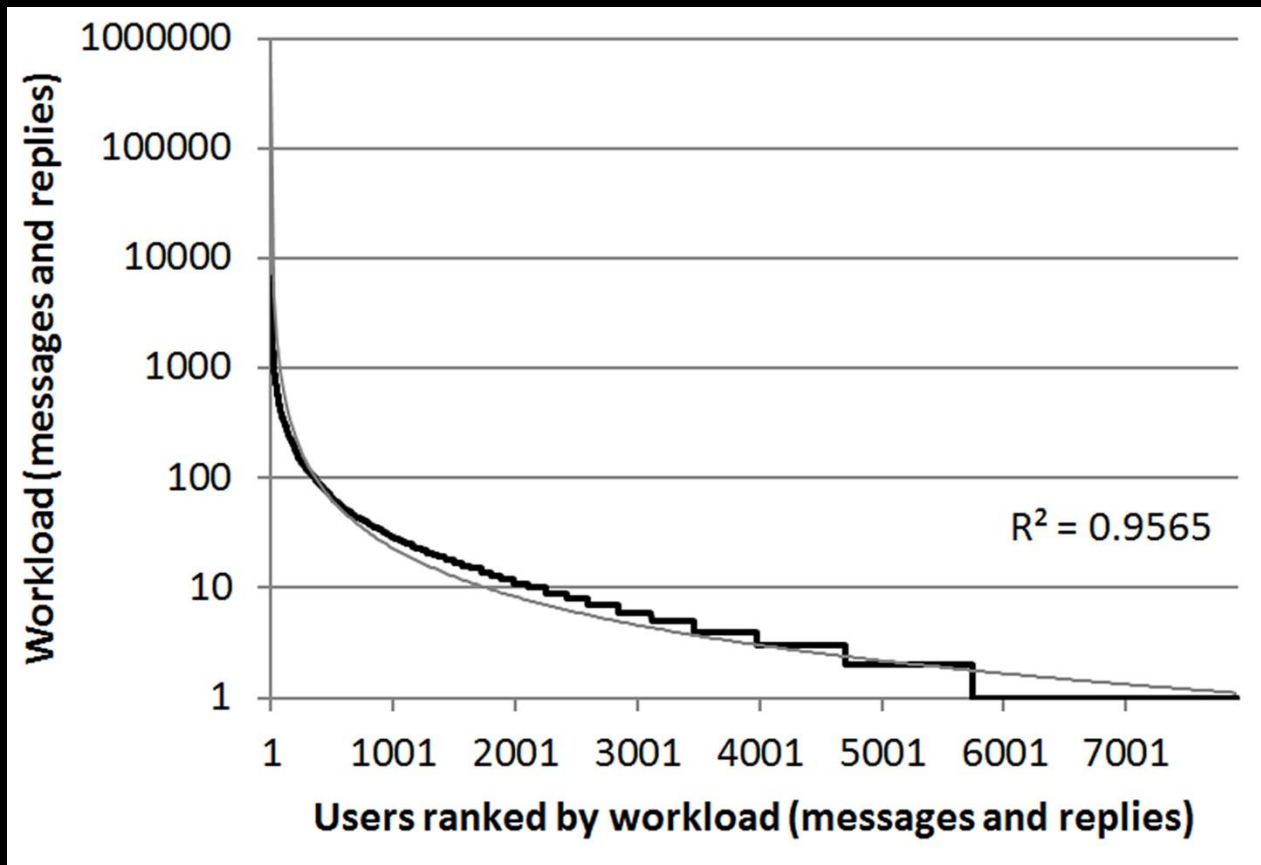
Whales vs Minnows

Ranked contribution in Phrase Detectives



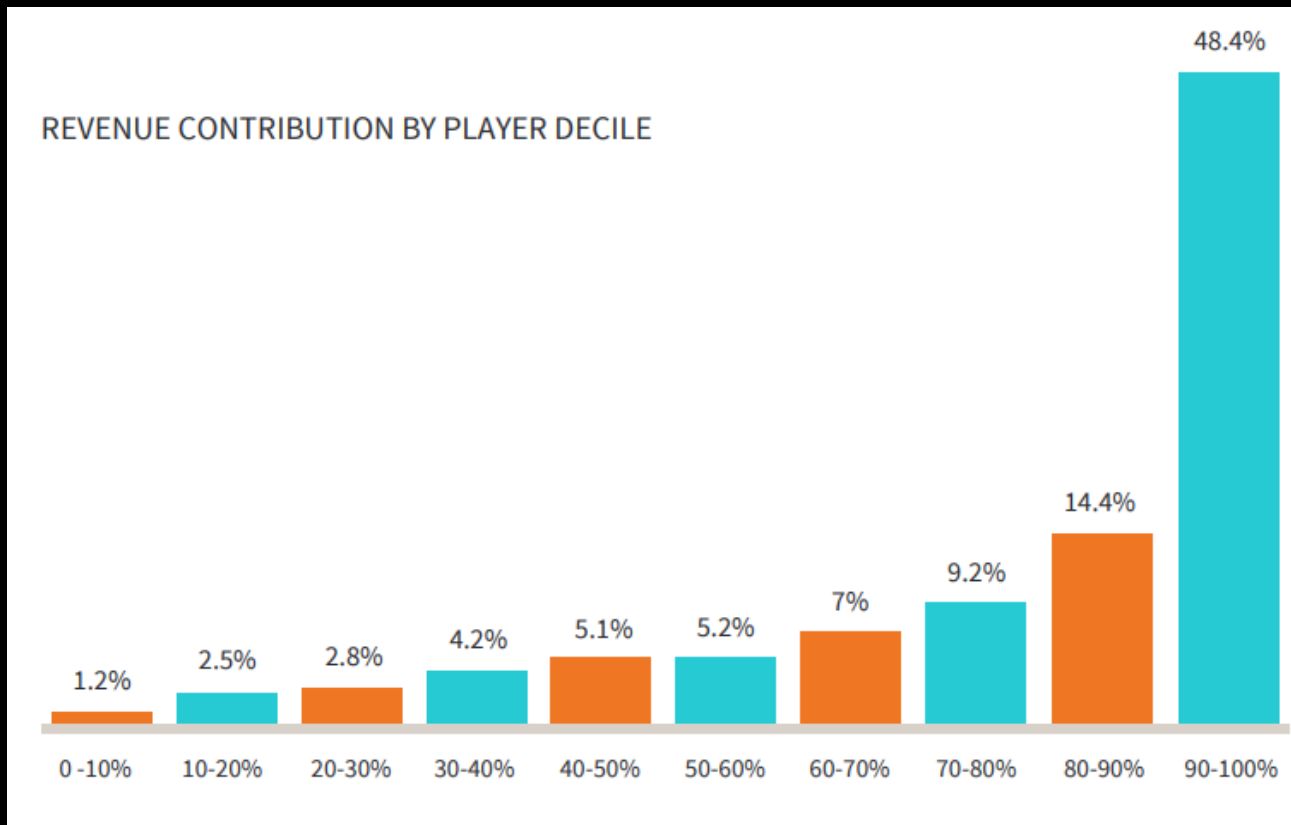
Whales vs Minnows

Ranked contribution on social media groups



Whales vs Minnows

Ranked revenue on mobile games



https://www.swrve.com/images/uploads/whitepapers/swrve-monetization-report-2016.pdf?utm_source=blog&utm_medium=organic

Player Metrics

A humpback whale is captured in the middle of a breach, its dark, ribbed back and white, mottled pectoral fin cutting through the blue ocean. Water is splashing around the whale's head and back. 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

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 /

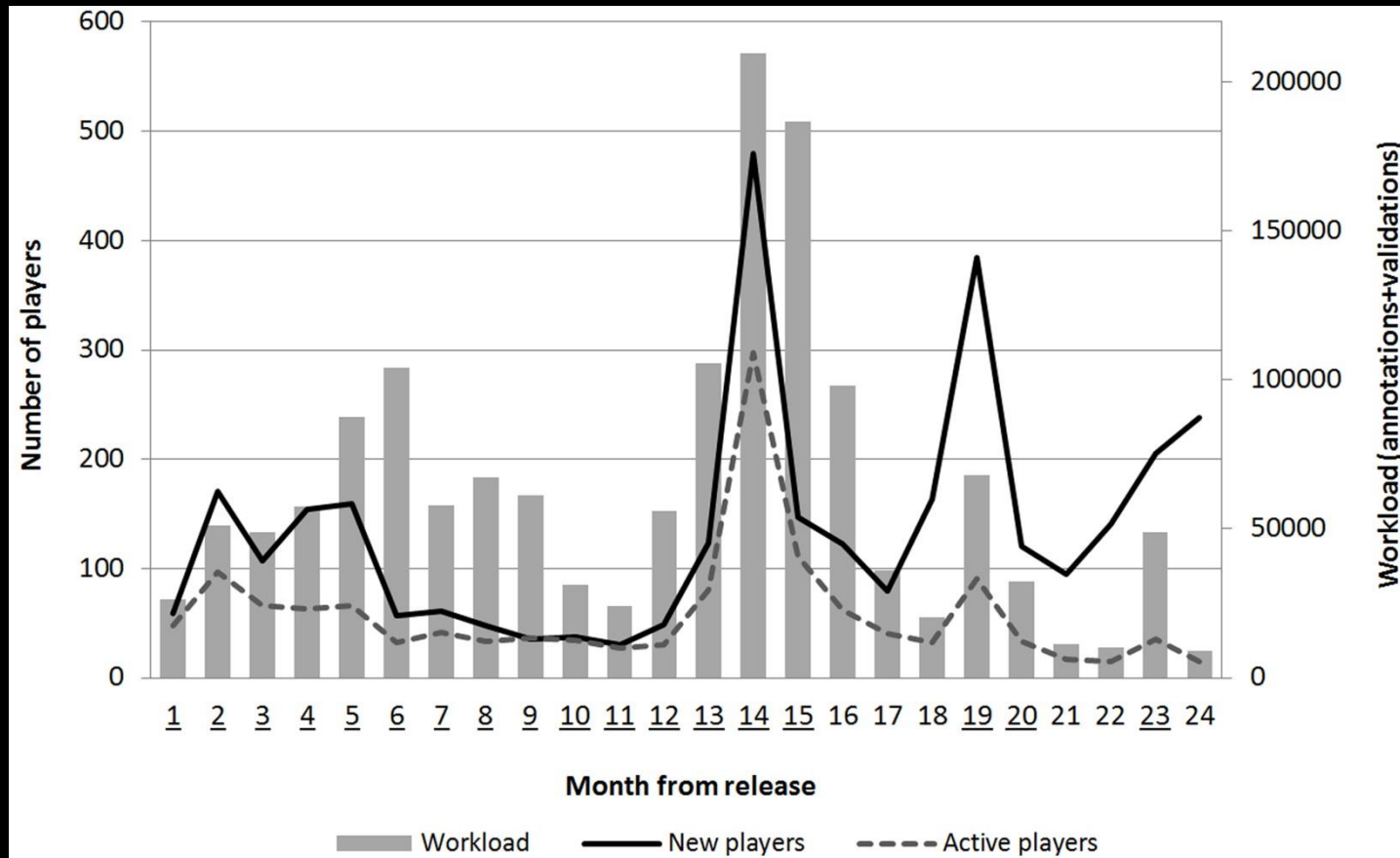
Percentage of players who stop playing

Community Metrics

A close-up photograph of several hands of different skin tones stacked together in a pyramid shape. The hands are positioned with fingers pointing upwards, creating a sense of unity and teamwork. The background is a plain, light-colored wall. The image is overlaid with a semi-transparent dark grey banner at the top and bottom, containing text.

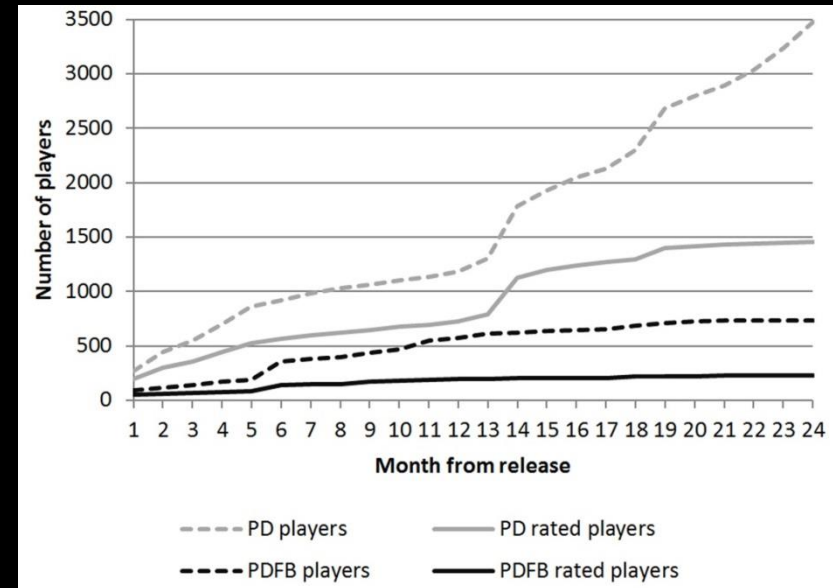
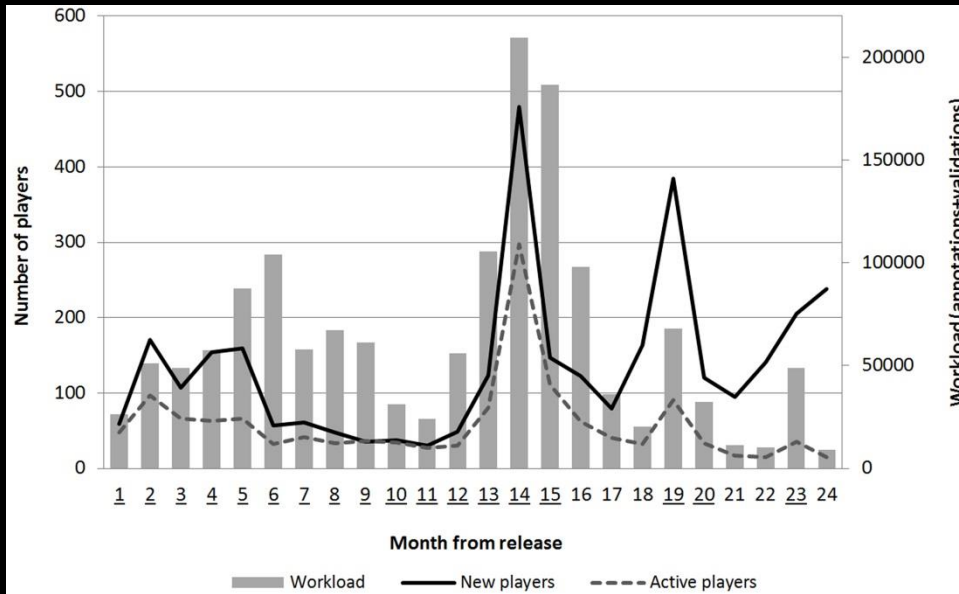
How fast is the game growing?
How "sticky" is the game (do players return)?
Are incentive methods working?

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

Cost per Judgement (CpJ)

Judgements Required (JR)

Cost per Item (CpI)

Throughput

Metrics indicating the overall performance of the system

Cost per Judgement (CpJ)

CpJ = Cost to get a useful contribution from a player

Financial cost of engagement, eg prizes

Ongoing cost of project (researchers, hosting, etc)

Judgements Required (JR)

JR = How many useful judgements are required to complete an item

Wastage from spam, training, attention slips

Aggregation method used

Difficulty of item compared to skill of players

Cost per Item (Cpl)

Cpl = Cost of to completely annotate an item

$$\text{Cpl} = \text{CpJ} * \text{JR}$$

Headline figure to estimate cost of complete corpus
Comparable across games, adjusting for size of data
collection

Throughput

Throughput = Speed of data collection

= Number of completed items per hour

Von Ahn defined the headline figure but not the distribution

Time to complete items will vary by difficulty and crowd skill

Community Metrics



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

Can we extend these metrics?

Metric	Description in relation to GWAP
Cost per Judgement (CpJ)	Average cost to get a player to provide a useful judgement.
Judgements Required (JR)	Average judgements required to complete an item.
Cost per Item (CpI)	Cost to acquire a completely annotated item.
Cost per Acquisition (CpA)	Cost to have someone start to play a game.
Lifetime Judgements (LTJ)	Total judgements made in the game per player.
Average Judgements per Player (AJpP)	Judgements per player.
Average Lifetime Play (ALP)	How long players play a game.
Monthly Active Users (MAU)	Total players who have submitted a judgement in a month.
Retention and Churn	Percentage of players retained/lost over a time period.
Throughput	Number of completely annotated items produced per hour.

Games with a Purpose (for NLP and for other purposes)

Microwork (Amazon Mechanical Turk, Crowdfunder)

Community QA (YahooAnswers, StackOverflow)

Educational? (DuoLingo)

Aggregation / Ambiguity

A large crowd of stylized human figures, rendered in a light blue-grey color, is scattered across the frame. In the center, a single figure is highlighted with a bright yellow glow, making it stand out from the rest of the crowd. The background is a dark, textured blue-grey.

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

Player Skill and Context



Value of a single good player vs many bad players
Contextual/real world knowledge required by the task
Can language learners provide useful language data?

Disagreements and Language Interpretation (DALI)

A 5-year, €2.5M project on using games-with-a-purpose and Bayesian models of annotation to study ambiguity in anaphora

A collaboration between Queen Mary, Essex, LDC, and Columbia

Funded by the European Research Council (ERC)

Ambiguity in anaphora

15.12 M: we're gonna take the engine E3

15.13 : and shove it over to Corning

15.14 : hook [it] up to [the tanker car]

15.15 : _and_

15.16 : send **it** back to Elmira



(from the TRAINS-91 dialogues collected at the University of Rochester)

Workplan

WP1: Improved GWAPs for Anaphora

WP2: Analyzing Multi-Judgment Data

WP3: An anaphorically annotated corpus with multi-judgment data
(from Y2)

WP4: A Linguistic theory of disagreements in anaphoric interpretation
(from Y3)

WP5: Models of anaphora resolution trained and evaluated with multi-judgment data (from Y3)

Our Games

Phrase Detectives

Collects data on anaphoric coreference, nearly 10 years old, new version to be released by the end of 2018.

TileAttack!

Platform to investigate player motivations around named entity tagging.

WordGems (under development)

Language learning game to introduce concept of noun phrases

Wormingo

Language learning game that combines data collection with non-data collection games

Lingo Boingo

The screenshot shows the Lingo Boingo website interface. At the top, there is a browser address bar with the URL <https://lingoboingo.org>. Below the address bar is a navigation bar with the site's logo, which consists of the letters 'L I N G O B O I N G O' in colorful circles. To the right of the logo, the text 'World Language Games' is displayed. Below the navigation bar, there are tabs for 'All', 'English', and 'French'. The main content area features a grid of four game cards:

- Jeux de mots** (French): Lexical and semantic games with a purpose in French.
- Phrase Detectives** (English): Compete against other detectives by identifying the relationships between words and phrases in a variety of texts including literature, history, travel.
- Tile Attack** (English): Go head-to-head against another player competing to identify the noun phrases of a text.
- Zombilingo** (French): Identify syntactical dependencies, collect brains and eat them! This language game is fun for both fans of grammar and zombies.

What else can we learn?

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

Players

World

Interacting

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



Who Are The Players?

User Personas created during a DALI team gathering Dec 2017.

Name	Age	Employment	Personality	Motivation	Interface	Social	We must	We must not	Bartle type
Chris Gelhead	18	Student	Extroverted, sporty	Needs to be thrilled, likes team play	phone, tablet	yes	engage quickly; quick progression and mastery; exciting gameplay	frustration; ask for registration details	KS
Hector Lector	29	Tax inspector	Unfulfilled, restless, trapped	Freedom, duty	phone, tablet, laptop	no	allow him to be creative; appreciate his efforts; make him feel liked	boss him around; give boring tasks	S
Hailey Bailey	37	Programmer	Focused, controlled	Provision for future, family, work	phone, tablet, laptop	some	give short bursts of gameplay that can be abandoned without consequence	dumb down, she likes a challenge	A
Mr Bank	40	City	Regular, commutes	play games on the train	phone, tablet	no	low entry hurdles; no sound	make it too challenging, make sessions too long	A
Sophie King	28	Entry level assistant	Grammar buff, lots of downtime, pedantic, grammar nazi	helpful, prove she knows her onions, maybe make extra money	all	yes	player ranking, collectables, short sessions, interact with other players	everybody wins, reward quantity over quality, remove competitive elements form	AS

Cheap Psychological Tricks

Bartle, R. MMOs from the Outside In (2016)

The Zeigarnik effect: "If you need to collect X objects, you won't want to stop at $X-1$ objects. When you finally get X of them, it is enjoyable – but perhaps only in the same way that stopping hitting your head with a hammer is enjoyable."



Metrics for Educational and Crowdsourcing Games

Jon Chamberlain | University of Essex | jchamb@essex.ac.uk
Massimo Poesio | Queen Mary University | m.poesio@qmul.ac.uk