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
The ESP Game
200,000 players, 50 million labels in 2 months
Purchased by Google to improve image search results
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/
Phrase Detectives on Facebook
Far fewer players in the first 2 years of release. What are those players doing?
RoboCorp
232 players in 2 months
64 played mini-game, 57 made in-game purchases
DuoLingo
200 million active users, 25 million monthly users, 6 billion exercises per month, 10 mins average play time per day
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

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

\[ \text{AJpP} = \frac{\text{Total judgements}}{\text{Total active players}} \]

= Average Revenue Per User (ARPU)

\[ \text{ARPU} = \frac{\text{Total revenue}}{\text{Total active users}} \]

AJpP = Average Judgements per Player

\[ = \frac{\text{Total judgements}}{\text{Total active players}} \]

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

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

![Graph showing workload vs players ranked by workload (annotations and validations)]
Whales vs Minnows
Ranked contribution on social media groups

![Graph showing the relationship between users ranked by workload and the number of messages and replies. The graph illustrates a power-law distribution with a coefficient of determination ($R^2 = 0.9565$).]
Whales vs Minnows

Ranked revenue on mobile games

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

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)

\[ Cpl = \text{Cost of to completely annotate an item} \]
\[ Cpl = CpJ \times JR \]

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
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?

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

Games with a Purpose (for NLP and for other purposes)
Microwork (Amazon Mechanical Turk, Crowdflower)
Community QA (YahooAnswers, StackOverflow)
Educational? (DuoLingo)
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)
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

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?

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

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.

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