Making WordSleuth Fun

Galina Tucker
The overall goal of this project: to create a database of messages expressing specific social cues

Applications:
- Testing cognitive science models
- Machine learning algorithms
  - Automatic extraction of tone from existing text
  - Word-processors giving “tips” on tone
  - Programs to help people who struggle with tone-detection
Existing corpora were examined
  * Too small
  * Only focused on one type of social information

Machine learning algorithms need a particularly large amount of data
How to Gather Enough Data?

- Newer method of data collection: game-with-a-purpose (GWAP)
- Easier to gather more data per person
- Easier to convince more people to participate
- The game can be created in a way to reward quality data

- Many people already familiar with web-based games, and spend lots of time with them
  - Want to tap into this resource by building a GWAP
Social tone information is something humans are better at than computers

- This is the gap we’d like to close

- Example: “Get over here, boy.”

- Any human reader may observe that the sentence is:
  - Persuasive / commanding in nature
  - Directed towards someone of lower social standing
  - Disdainful, entitled tone
In our GWAP, WordSleuth, there are two types of play:

1. People are given a social tag (such as “persuading”) and a picture, then asked to create a message for the picture that conveys that social information.

2. People are given a message and the picture used to create it, then asked to guess which social tag it was originally created with.
Some benefits of this structure:

* Will get a variety of opinions on what key words or language structure makes a message reflect a tag
* Can analyze tag ambiguity through “guesses” results
* Can pull out sentences that are the clearest examples of their tag
The nature of a GWAP makes it easy to tweak parameters
- A single social tag may be adjusted to be more prominent
- New tags may be added
- Old ones retired
- Hardcoded quality limits imposed through “taboo words”
- Messages can be required to be a certain length
Our GWAP - WordSleuth

- Player-based quality control
  - How many people agree that a message matches its tag
  - Poor quality sentences can also be flagged (as of v3)
- System will grow as the number of players does
Pearl and Steyvers (2010) ran a small-scale test with an offline version of WordSleuth.

- Demonstrated the potential usefulness of this type of data
  - Analyzed which social tags humans most frequently confuse
  - Trained a machine learning classifier that performed decently
My focus: how can we get people to play?
  * We won’t get enough data otherwise

Goals:
  * Make the game function properly online
  * Make the game stable for large numbers of people
  * Add new features to increase enjoyment
**WordSleuth – Functionality**

- Basic game flow did not work right online
  - Was re-implemented with help of another student

- Some features had to be re-worked / added
  - Retrieving forgotten passwords
  - Editing user account information
  - Making the layout work with any size image prompt
  - Not accepting messages that contain taboo words
  - Allowing punctuation in messages
* Most important issue – data storage
* Originally data was kept in text files
  * Very difficult to understand and alter
  * Cannot scale
* Solution: implemented a MySQL database
  * Five initial tables with data on: social tags, members, created messages, guesses made, and the pictures used as prompts
* Substantial code had to be rewritten
This first version of WordSleuth was released to a select group of friends, family, and linguistic research students. They were asked to play 30+ minutes on each mode, and to fill out a survey.
How often do you see yourself playing?

On a scale of 1 to 6, how likely would you be to put more time into this game?

On a scale of 1 to 6, how much fun did you have while playing?

V1 average: 3.88

V1 average: 3.55
9 people took the survey

* Likely positive bias due to the select nature of participants
* The game was not found very fun
* The most common answer to “How often do you see yourself playing?” was “never again!”

Two goals from here
1. Improve numeric results
2. Address specific problems / suggestions noted in free-form survey questions
Von Ahn (2008): “People play [GWAPs] not because they are personally interested in solving an instance of computation problem but because they wish to be entertained.”

Four new features were added towards the goal of entertainment:
- A new score (activity points)
- Unlockable features
- Difficulty levels
- High score tables
Original version had four score types
  * Expressive Score/IQ, Receptive Score/IQ
* Problem was: expressive points based on other people
* Instant feedback is an important motivator

* Activity points are a count of actions a user has taken
  * Total number of guesses + total number of creations
* At least have some instant feedback for creating, now

Current score for frawlik
Expressive: 325, Receptive: 600, E-IQ: 126, R-IQ: 113, Activity Points: 139
Survey results supported this addition

- “Creating sentences doesn’t seem rewarding. They haven’t been tagged ... so that score hasn’t moved since I created this account.”
- “Creating sentences gets old after a while.”
Many existing games use instant feedback.

- Example: Chess Tempo (online game)
- Point of game is to solve chess puzzles
- After each puzzle, your rank and the puzzle’s ranks are both immediately adjusted based on how you did
WordSleuth v2 – Unlockable Features

* Activity points also now unlock new variations in gameplay

* Unlockable features give users small goals to achieve
  * “Create” mode must now be unlocked
  * Can also unlock difficulty levels: easy, medium, hard

* Forces user to stagger play of different parts of the game
WordSleuth v2 – Unlockable Features

* Unlockable create mode increases data integrity

* Some users did not know what we wanted from them
  * Sometimes gave definitions, or general statements

* Users must now guess 15 messages before they can create
  * Are forced to see examples
  * Increases quality of their subsequent creations
Research supports this addition
Malone (1980) wrote a paper on what makes learning fun

Unlockable features taps into 2/3 categories he discusses
  * **Challenge**: users want an obvious and compelling goal
  * **Curiosity**: users want to know what harder difficulties are like
Select a difficulty and mode, then click the arrow to play. Harder difficulties will let you earn more points.

Select a difficulty
- Easy - Level 1
- Medium - Level 2
- Hard - Level 3

Select a mode
- Guess - assign tags to sentences
- Create - write sentences based on tags

Complete 76 more plays of either mode to unlock hard difficulty.

Play more!
Many existing games have unlockable content

* Picross (puzzle game) – can unlock harder difficulties after completing a certain amount of easier ones
* Harvest Moon (farming game) – new crops, animals, and people to marry are unlocked as you progress
* Soul Calibur, Super Smash Brothers (fighting games) – can unlock new fighters when you accomplish certain tasks
Difficulty levels allow users to perform harder tasks for greater rewards

* Levels are easy, medium, hard

* For guessing, difficulty is based on how many people guess a message correctly
WordSleuth v2 – Difficulty Levels

* For creating, difficulty is adjusted through taboo words

* Two taboo word lists:
  * Static list contains words no message should ever have
  * Dynamic list contains words commonly associated with particular tags
All harder modes provide the user with more points

- Easy 15, Medium 30, Hard 45

Difficulties do not affect:

- **Activity points** (because they are supposed to be a cohesive count of actions)
- **IQ** (because it is based currently on accuracy alone)
Research supports this addition
Malone’s paper (1980) on what makes learning fun

Difficulty levels relate to the challenge category
Allows the user to select how much of an “uncertain outcome” they desire
Many existing games have difficulty levels

- Mass Effect (roleplaying game) – users can choose “Casual,” “Normal,” or “Veteran” (and later unlock “Hardcore” and “Insanity”)
- Many shooter games provide similar basic structure
- World of Warcraft (MMORPG) – provides two versions of much of its end game content, “regular” and “hard mode”
High score tables relate to both goal-driven and socially-driven motivations.

Five high score tables were introduced in v2 (correspond to 5 existing scores).
### All-Time Top Scores

#### Top Receptive Scores

<table>
<thead>
<tr>
<th>Position</th>
<th>Player</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Shamu</td>
<td>15615</td>
</tr>
<tr>
<td>2</td>
<td>helan</td>
<td>3675</td>
</tr>
<tr>
<td>3</td>
<td>upup</td>
<td>3660</td>
</tr>
<tr>
<td>4</td>
<td>brendatucker</td>
<td>3670</td>
</tr>
<tr>
<td>5</td>
<td>Lauren</td>
<td>3195</td>
</tr>
<tr>
<td>6</td>
<td>hunning</td>
<td>1350</td>
</tr>
<tr>
<td>7</td>
<td>up90</td>
<td>1035</td>
</tr>
<tr>
<td>8</td>
<td>owagenaa</td>
<td>1005</td>
</tr>
<tr>
<td>9</td>
<td>top777</td>
<td>915</td>
</tr>
<tr>
<td>10</td>
<td>Yin</td>
<td>885</td>
</tr>
</tbody>
</table>

#### Top Expressive Scores

<table>
<thead>
<tr>
<th>Position</th>
<th>Player</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LisaEx</td>
<td>1612</td>
</tr>
<tr>
<td>2</td>
<td>labsubject10</td>
<td>1575</td>
</tr>
<tr>
<td>3</td>
<td>labsubject15</td>
<td>1200</td>
</tr>
<tr>
<td>4</td>
<td>up90</td>
<td>1110</td>
</tr>
<tr>
<td>5</td>
<td>brendatucker</td>
<td>975</td>
</tr>
<tr>
<td>6</td>
<td>labsubject10</td>
<td>720</td>
</tr>
<tr>
<td>7</td>
<td>labsubject13</td>
<td>675</td>
</tr>
<tr>
<td>8</td>
<td>labsubject15</td>
<td>675</td>
</tr>
<tr>
<td>9</td>
<td>labsubject10</td>
<td>665</td>
</tr>
<tr>
<td>10</td>
<td>labsubject5</td>
<td>800</td>
</tr>
</tbody>
</table>

#### Top Receptive IQs

<table>
<thead>
<tr>
<th>Position</th>
<th>Player</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>kaezera</td>
<td>140</td>
</tr>
<tr>
<td>2</td>
<td>Edwina</td>
<td>139</td>
</tr>
<tr>
<td>3</td>
<td>imryel</td>
<td>135</td>
</tr>
<tr>
<td>4</td>
<td>thepagemaster99</td>
<td>130</td>
</tr>
<tr>
<td>5</td>
<td>notsmoch</td>
<td>134</td>
</tr>
<tr>
<td>6</td>
<td>Yin</td>
<td>132</td>
</tr>
<tr>
<td>7</td>
<td>jamashob</td>
<td>133</td>
</tr>
<tr>
<td>8</td>
<td>chara</td>
<td>131</td>
</tr>
<tr>
<td>9</td>
<td>Imwashaw</td>
<td>129</td>
</tr>
<tr>
<td>10</td>
<td>hunning</td>
<td>125</td>
</tr>
</tbody>
</table>

#### Top Expressive IQs

<table>
<thead>
<tr>
<th>Position</th>
<th>Player</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>up90</td>
<td>140</td>
</tr>
<tr>
<td>2</td>
<td>labsubject10</td>
<td>139</td>
</tr>
<tr>
<td>3</td>
<td>labsubject13</td>
<td>137</td>
</tr>
<tr>
<td>4</td>
<td>hunning</td>
<td>136</td>
</tr>
<tr>
<td>5</td>
<td>helen</td>
<td>133</td>
</tr>
<tr>
<td>6</td>
<td>labsubject10</td>
<td>132</td>
</tr>
<tr>
<td>7</td>
<td>labsubject5</td>
<td>131</td>
</tr>
<tr>
<td>8</td>
<td>chrisnara</td>
<td>131</td>
</tr>
<tr>
<td>9</td>
<td>labsubject12</td>
<td>130</td>
</tr>
<tr>
<td>10</td>
<td>labsubject19</td>
<td>129</td>
</tr>
</tbody>
</table>

#### Most Active Players

<table>
<thead>
<tr>
<th>Position</th>
<th>Player</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Shamu</td>
<td>711</td>
</tr>
<tr>
<td>2</td>
<td>upup</td>
<td>3350</td>
</tr>
<tr>
<td>3</td>
<td>brendatucker</td>
<td>309</td>
</tr>
<tr>
<td>4</td>
<td>helen</td>
<td>280</td>
</tr>
<tr>
<td>5</td>
<td>kaezera</td>
<td>248</td>
</tr>
<tr>
<td>6</td>
<td>Lauren</td>
<td>245</td>
</tr>
<tr>
<td>7</td>
<td>LisaEx</td>
<td>224</td>
</tr>
<tr>
<td>8</td>
<td>top777</td>
<td>155</td>
</tr>
<tr>
<td>9</td>
<td>imwashaw</td>
<td>139</td>
</tr>
<tr>
<td>10</td>
<td>hunning</td>
<td>121</td>
</tr>
</tbody>
</table>
Research supports this addition

Peter Vorderer (2011) calls the player versus player concept “social competition”

Argues that doing well can cause a user to have increased self esteem, causing positive feelings towards the game

High score tables allow the user to directly see how they compare to other people
Many existing games rely heavily on high score tables

- Arcade games is the classic example
- Bejeweled Blitz on Facebook – the game is centered around being the highest scorer out of your friends, reset every week
- Audiosurf from Steam – levels are created from a user’s songs, each song has a high score table
This second version of WordSleuth was also released to a select group of people.

However, this group was broader in scope.

Players were asked to:
- Play 15+ minutes on each mode, and to fill out a survey.
- Create new accounts if they had played previously, to experience the new beginner game.
WordSleuth v2 – Survey Results

How often do you see yourself playing?

On a scale of 1 to 6, how much fun did you have while playing?

On a scale of 1 to 6, how likely would you be to put more time into this game?

V1 average: 3.88
V2 average: 2.83

V1 average: 3.55
V2 average: 3.16
WordSleuth v2 – Survey Results

* 12 people took the survey
  * 6 had played v1, 6 had not
  * Less of a positive bias this time

* The most common answer to “How often do you see yourself playing?” is now “a few times per year”

* Same two goals
  1. Improve numeric results
  2. Address specific problems / suggestions noted in free-form survey questions
Comments support belief that v2 was an improvement

Noted favorite differences were:

- “High score tables are neat.”
- “I loved the activity points count!”
- “Activity points leading to new levels.”
- “Things to unlock were my favorite.”
For v3, we wanted to add

- Features to reduce frustration at ambiguous tags
  - Skipping messages
  - Flagging messages
  - Clarifying create message page

- More features to increase enjoyment
  - Achievements
  - Profile pages
• “What part of the game did you enjoy least?”
  • “The most frustrating part of the game is when I got the answer wrong for what I still believe should have been the correct answer.
  • “Some of the sentences with the answer just didn’t make sense. Deception was a big one: not sure people know what that means.”
  • “Stupid people that upload sentences that have nothing to do with the tone they were given.”
57% of all responses to that question, in first and second surveys combined, touched on this issue

Addressed in multiple ways
- Letting people skip messages
- Tweaking the create message process
- Letting people flag messages
Skipping Messages

* Users may click a link to skip a message they don’t like
* Technically not new functionality
  * Made the process transparent, however

* This allows people to avoid the problem, but does not really solve it
Modifying Create Message Process

- Addresses the cause directly
- Create page was changed to show all tags, not just the one the user needs to create for
- They are also advised to avoid ambiguity

- Suggested by survey comment:
  - “I think that when creating a caption, the user should be able to view all 8 possible tags ... This way they can make sure that their caption is most like the one they are aiming for ...”
Flagging messages

* Addresses the effect of the issue directly
* Users are encouraged to flag bad messages
* A bad message:
  * Is nonsensical or very poorly spelled
  * Gives away the tag in an explicit fashion
  * Very strongly does not match the given tag
* Messages with enough flags are marked in the database
You are playing on medium difficulty. You will earn 2x the base number of points when people guess correctly. Additionally, you will earn 5 points just for creating!

Write a message that more clearly expresses rudeness than any other tag.

All tags: confidence, deception, disbelief, embarrassment, formality, persuading, politeness, rudeness

Don't use any of these taboo words: rude, rudely, rudeness, hah, popped, hose

Please write quality messages. Items judged as bad may be removed; if so, points will be taken away. :

My message is complete

Click to Skip
Achievements are ribbons earned for reaching some goal
  * May be trivial (guess 25 messages) or very difficult (guess 100 messages in a row correctly)
  * Gives users non-gameplay affecting goals to pursue
  * Gives users very hard goals they can optionally pursue
Research supports this addition

Mikael Jakobsson (2011) argues that adding achievements is actually adding a second game atop the original one

“...achievements provide a [strong] sense of optional unfinishedness.”

“I can convince myself that further engagement with the game is reasonable ... because the achievement scaffolding stretches further and provides a direction”
Many existing games have strong achievement systems

- Xbox360 achievement system is what Jakobsson wrote on
- Farmville on Facebook – rewards diligent play with ribbons and occasionally little banners for your farm
- World of Warcraft – added achievements two expansions ago with the release of The Wrath of the Lich King
Profile pages were added in v3
- All known scores
- How many messages have been guessed / created
- Visible representation of achievement ribbons

Improves the community aspect of WordSleuth
An easy way for users to view their achievements
Ties in to Vorderer’s “social competition” drive
  * Users can easily see what any given user has achieved
  * People feel more attached to a website when they have a page that is “theirs”
  * In fact, this is the entire foundation for a website as popular as Facebook!
Third version of WordSleuth was released to a general audience

- Link was added to the CoLa lab website
- Advertised through the CHP newsletter

Players were asked to:

- Play 15+ minutes on each mode, and to fill out a survey
- Create new accounts if they had played previously, to experience the new beginner game
WordSleuth v2 – Survey Results

**How often do you see yourself playing?**

![Bar chart showing frequency of gameplay](chart1.png)

V1 average: 3.88
V2 average: 2.83
V3 average: 3.81

**On a scale of 1 to 6, how likely would you be to put more time into this game?**

![Bar chart showing likelihood of increased time](chart2.png)

V1 average: 3.55
V2 average: 3.16
V3 average: 3.18

**On a scale of 1 to 6, how much fun did you have while playing?**

![Bar chart showing fun ratings](chart3.png)
WordSleuth v3 – Survey Results

* 17 responses
  * 6 played v2, 11 had not
* Comments support belief that v3 was an improvement
* Some comments were:
  * “I like the new flagging feature the best, although the achievements are also really cool.”
  * “Achievements are awesome. Flagging is very nice. Good choices.”
  * “Unlocking an achievement was awesome. Then I clicked my name and saw it there and it was GLORIOUS.”
Interesting backlash against skipping from some people

- “Skipping a question is weird. Why not skip everything that you aren’t 100% certain on?”
- “I think having the skip feature defeats the purpose to see what others think of the clues.”
- “least [favorite feature] SKIP”
Some features to implement in future versions

- Basic social networking
  - Being able to add friends, message people, compare profiles
- Present users with trivia related to their guesses / creations
- More visual rewards for playing
- Maybe some fantasy/story-mode elements
- Adding more high score tables
- Adding more achievements
- Having the IQ scores account for question difficulties
The amount of data generated makes this method of collection seem successful so far.

- WordSleuth has been played for about 6 months
  - Much of this was in a limited, beta tested capacity
- 56 seed members are now 171 total members
  - Some of this is users creating duplicate accounts for testing
- 2,873 seed guesses are now 8,569 total guesses
- 1,060 seed messages are now 2,191 total messages
We believe that the addition of these new features has had a substantial affect on the speed of data gathering.

With this rapidly growing amount of data, our applications seem more plausible and not so idealistic.

Practical and theoretical applications:

- Testing cognitive science models
- Machine learning algorithms
  - Automatic extraction of tone from existing text
  - Word-processors giving “tips” on tone
- Programs to help people who struggle with tone-detection
Thank you!