How to Get People to Want to Give You Data: Making WordSleuth Fun

Galina Tucker (May 2011)

Abstract

With the increasing prevalence of online games in people's lives, a new effective way to gather data has emerged called a game-with-a-purpose (GWAP). We have created a GWAP by the name of WordSleuth to accumulate a database linking text messages to specific pieces of social information. One important purpose of gathering this data is to train machine learning algorithms so that computers may one day reliably recognize social information in text, such attitudes, intentions, and emotions. However, this would only work if the game was fun enough to play repeatedly. Numerous new features, selected based on an analysis of existing research and currently popular games, were added to WordSleuth. The enjoyment of these new features was analyzed through surveys players took at various points in the game's development. The addition of these features has been successful at greatly increasing the rate in which WordSleuth is accumulating social language data.

This paper was written under the guidance of Professor Lisa Pearl in the department of Cognitive Science at the University of California, Irvine.

1. Introduction

Our goal is to create a social information repository consisting of messages that express specific social cues. This database would have a variety of applications, spanning a variety of disciplines as well as the full range from theoretical to practical. For example, cognitive scientists could use it to test various models of how people communicate tone (like embarrassment or disbelief) in language. In addition, machine learning algorithms can be created based on this database. Assuming that these become reliably accurate, programs could be developed for a variety of different functions. For example, social information could be automatically extracted from articles or websites, word-processors could note what sentences don't match the intended tone, and perhaps programs could even be written to help people who struggle with tone-detection, such as children suffering from autism or Asperger's syndrome.

Unfortunately, it is not feasible to use already existing corpora of to achieve the various goals outlined above. Even if someone identifies a general tone for a document or set of data, we require social information on a much finer level, an annotation per sentence or two. Collecting raw data about these mappings is necessary for developing machine learning algorithms that can automatically extract this information at or above human performance levels. There are some existing corpora that annotate at an appropriate granularity, but only consider a single type of social information. For example, the Linguistic Data Consortium (http://www.ldc.upenn.edu) recently added a Language Understanding Annotation Corpus (LUAC) through the efforts of Diab et al. (2009). This corpus annotates whether a statement is a belief, speculation, or considered a fact. While this is one specific type of social information, we are seeking something more broadly examining multiple types of tones. An additional problem is the fact that this is a very small sample containing only about 7000 words in English. Language corpora need to be much larger to provide sufficient data. For example, the Gigaword corpus from the Linguistic Data Consortium contains 1.75 million words. Some open-source data can be found containing social data as well. For example, there is a set of data regarding what was said on an online gaming forum where gameplay involved the intent to deceive (e.g., Zhou and Sung 2008: Mafia game forums). Once again, we encounter the problem that only one type of social information has been focused on – deception, in this case. As we desire a corpus with multiple pieces of social information, the best way seems to be to build it ourselves.

In working towards this goal, we must overcome a very frequent problem in research: generating sufficient data to draw conclusions. This problem is often a result of the amount of effort it takes to recruit enough subjects for an experiment, or even just to fill out simple survey questions. The trick is to get people to want to take part in whatever study it is. This has

typically been accomplished by trying to interest potential subjects in the study and sometimes by offering monetary or school grade incentives for participating. Even once people are physically participating, however, there is never any guarantee as to the quality of the data. Money can get them into the seat, but it can't force them to pay attention instead of just rushing through to reach the end.

In our specific case, the problem is exacerbated because more than the typical amount of data is required. Humans can easily extrapolate social information out of small amounts of data, even if their claims may have less statistical significance. However, when using computers to develop machine learning algorithms, massive amounts of data are often required to make these algorithms accurate. Machines don't bring background knowledge to the problem the way that humans do. Quality control also becomes a concern, as computers will accept all properly-formatted data, without questioning the strangeness of any responses. When humans analyze data, they can throw out pieces that are invalid or faulty. Thus, human-based quality control must exist to ensure the integrity of developed algorithms, adding another level of tedium.

1.1. Games with a purpose

To overcome the problem of getting sufficient data, a newer method of data collection was investigated. This method is the idea of a *game-with-a-purpose*, or GWAP (von Ahn 2006). If people honestly enjoy the process of providing data, there will be a much higher level of motivation to do so and larger amounts of data can be easily gathered. Assuming proper design, with precautions built directly into the system, participants can receive not just motivation to provide data but motivation to provide *quality* data. A subject who enjoys a game enough will not just run through the requested number of questions, but continually return to the game and play more. Thus, GWAPs have the theoretical benefits of attracting more people to provide data, providing more incentives to produce quality data, and returning a higher amount of data per each subject. If developed properly, GWAPs have the potential to be extremely effective at easily gathering massive amounts of high-quality data.

Another issue of quality arises when one considers that there are numerous tasks that humans can consistently perform better than machines. Many of von Ahn's GWAPs strive to collect data in these areas specifically. One example of this is visual processing, which the ESP game (http://www.gwap.com/gwap/gamesPreview/espgame/) tries to gather data on. In the game, two people are presented a picture and possibly a few taboo words. Both users simultaneously type in words to describe the picture until there is a word both of them typed in. The idea is to gather data on these "tags", which are likely to name either a salient object in the image or an important feature of it (von Ahn et al. 2006a). A second example is the game Verbosity (http://www.gwap.com/gwap/gamesPreview/verbosity), which attempts to gather a

database of common-sense facts such as "milk is white." It achieves this through gameplay where one user is given a word to write sentences about in an attempt for the other user to guess the word. These sentences are intended to be common-sense facts about the word itself (von Ahn et al. 2006b).

These are all examples of the concept of "human computation" (Thompson 2007), where computers outsource parts of a process to humans. Computers may then analyze and sometimes utilize the responses. Human computation does not just occur in GWAPs, but is also frequently performed through "crowdsourcing," where a question is simply put out to many people who are indiscriminately asked to answer. A classic example of crowdsourcing human computation is Amazon's Mechanical Turk, where people can sign up and perform various "human intelligence tasks" for small amounts of money (Mason et al. 2010).

With the increasing presence of the internet in the average person's life, there is a large and varied audience already familiar with the structure of small web-based games. For example, many people devote hours and hours of their lives towards Facebook applications such as Farmville and Mafia Wars. A GWAP, on the surface, can easily be built to mirror these diversions that many people are already very comfortable with. The wide variety of people who play web games ensures the diversity of our subject pool, allowing us to crowd source our own research needs.

1.2. A GWAP for social information in language

Our research focuses on one particular task that humans tend to be good at: extracting social information (such as tone, intentions, and social status – referred to hereafter as **social tags**) from a written message. For example, let us examine the sentence "Get over here, boy." The reader may easily intuit that the sentence is both persuasive in nature and directed towards someone of lower social standing (likely through age, possibly through rank), due to its somewhat disdainful and entitled tone. In our GWAP, diverse groups of people are given a social tag and asked to create a message, giving us a variety of opinions on what key words or specific language structure makes an utterance express that tag. With enough of this data filling our repository, the goals of creating machine learning algorithms and cognitive models can become achievable.

Using a GWAP makes it very easy to tweak parameters, such as forcing a single social tag to be more prominent if we need more data, or adding entirely new ones with minimal difficulty. Additionally, it also allows us to provide specific hardcoded quality limits on the data we want to receive and store, such as restricting which words may be used and imposing a minimum length on their messages. Beyond what is hardcoded, there is also a crowdsourced element of quality control. This is very beneficial for the system's scalability, as the quality

control system will grow necessarily and automatically with the amount of data being inserted. The most basic level of our crowdsourced information is the fact that we have data containing statistics on how many people agree that a given a message matches with its associated piece of social information. Beyond that basic information, poor quality sentences can be flagged (as of the latest version of the game) creating a system where users are checking on the work of other players for us.

Pearl and Steyvers (2010) ran a small-scale test with an offline version of our GWAP, called WordSleuth. One point of the experiment was to demonstrate the potential usefulness of the types of data we want to collect. For example, they analyzed which social tags are most frequently confused by humans, and demonstrated how a machine learning classifier trained on that dataset could draw close to reproducing human behavior.

The offline game used was restricted in functionality, but revolved around WordSleuth's two core gameplay modes. In the first mode, users were presented with a tag and a picture, and told to create a message for the picture that conveys the social information of the tag. This was called the create mode. In the second mode, users were presented with a message and the picture used to create that message, and asked to guess which tag that message was originally created with. This was called the guess or tag mode. A few benefits are gained from including the latter gameplay mode as well, in particular the quality control aspect discussed briefly above. For example, we can check the percentage of guessers that agreed with the creator's original social tag, and easily pull out which messages most clearly exemplify their tag. Additionally, guess data provides valuable insight on which tags are most easily confusable given only text and a simple context picture (in speech, additional non-verbal cues such as voice pitch, gesture, and facial expression play important roles).

We cannot hope to get enough data unless people are actually motivated to play the game. This project focused on two specific tasks to accomplish this goal. The first was to make the game actually function properly online (allowing the internet population access) and to be stable for larger numbers of people (supporting a high level of internet attention when popularity is reached). These factors are both very important, but do not motivate people on their own. Their presence simply means that users will not be given extraneous motivation to *stop* playing. As far as enticing people to continue playing once they start, and to ensure that they want to return in the future, new gameplay features to increase enjoyment had to be added.

2.1 Transitioning our original GWAP to the web

A limited version of WordSleuth written in Matlab was used to run some preliminary experiments. An initial Perl code base was created to transition this limited version to a web application. This code base was not complete, however, so I began by addressing the problems of stability and functionality.

As far as stability goes, the most important change to the code involved altering the way that data were stored. In the preliminary version, code was written and retrieved from text files. The nature of this meant the code was very difficult to understand and alter. More importantly, however, text files simply cannot scale with the large amounts of data we are aiming to acquire. We needed to implement an actual database, so we chose MySQL due to its ease of use, good performance, and high reliability. I installed MySQL on the server, then designed and created all the necessary tables. The database initially contained five tables that stored information on the social tags, the members, the created messages, the guesses made, and the pictures used as prompts.

After the database was set up, a substantial amount of code needed to be rewritten to correctly access the new tables instead of the text documents. Additionally, new wrapper functions were created to make accessing the database simpler. Constraints were added to the database to ensure data integrity (for example: requiring that the id of the message associated with a guess actually exists in the message table). The basic game flow of creating and guessing sentences did not initially function correctly on the live website, and was implemented in part with the assistance of another student.

The code went through several rounds of testing and debugging, which included the addition of a few features not included in the preliminary version:

- allowing users to retrieve forgotten passwords through emails
- allowing users to edit their account information
- making the layout display properly no matter the size of the image prompt
- forbidding taboo words
- making punctuation in messages work properly

Additionally, the starting data (from the preliminary experiment) was cleaned up during this phase. Low-quality and testing messages were removed, as were all initial users that had not created or guessed anything. The pictures listed in the database were synched with the pictures that actually existed on the server.

2.2 First round beta testers

Once this was completed, the first version of WordSleuth was functional. Rather than releasing it to the public, we found a group of beta testers consisting of friends, family, and students interested in linguistic research. A survey was created for this first round of beta testers, and Professor Pearl and I contacted them with the following request:

> WordSleuth is now ready and open for beta testers!
>
> Please go to http://gwap.ss.uci.edu/ to make an account and start playing. Play as frequently as you wish over the next week (with a minimum of 30 minutes on each of the two modes: creating and tagging), then fill out the beta tester survey.
>
> *[...information on how to find the survey...]*
>
> If you encounter any errors, please note the details of them. You may either save these notes and submit them in the final survey, or email the issues to me (gtucker@uci.edu) as you encounter them.

The results of the beta testers are as follows:

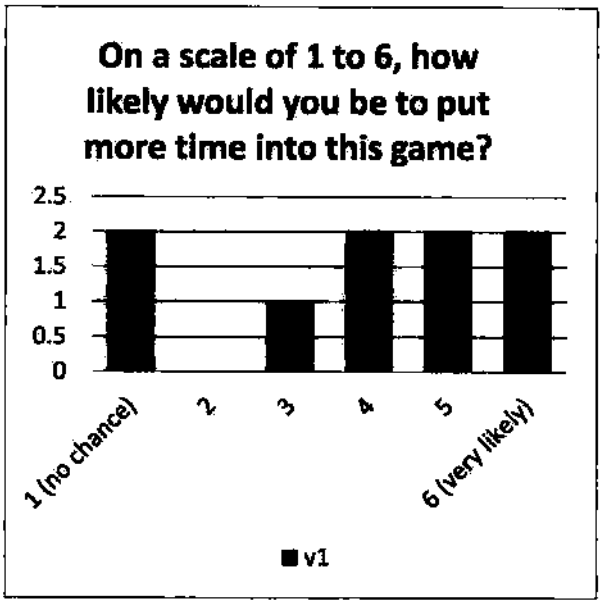

Figure 1: Further time commitment

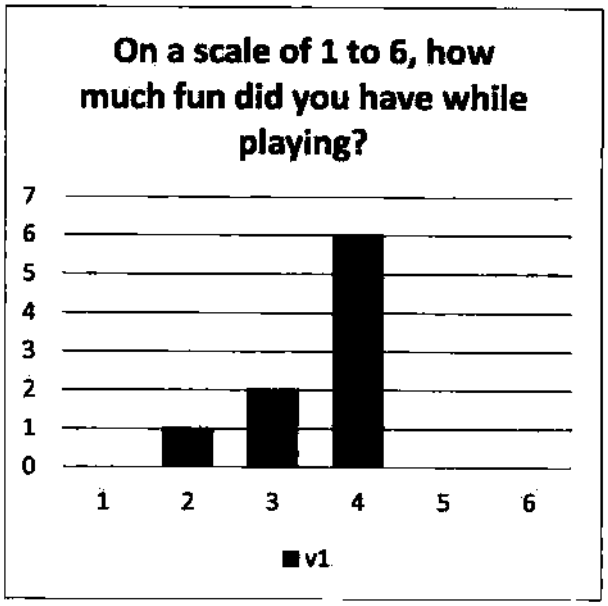

Figure 2: Enjoyment

These graphs show the results of three key questions on the survey. In total, 9 people took the survey. All were close friends or were working with linguistic research. Thus, the results were bound to have a positive bias. Even so, many people appeared to find the game mediocre. The most common answer to "How Often Do You See Yourself Playing" was "never again!" The average likelihood of putting more time into the game, on a scale of 1-6, was 3.88. The average amount of fun, also on a scale of 1-6, was 3.55. These were all numbers that we wanted to improve, in addition to a secondary goal of addressing specific problems and implementing suggestions noted in free-form questions on the survey.

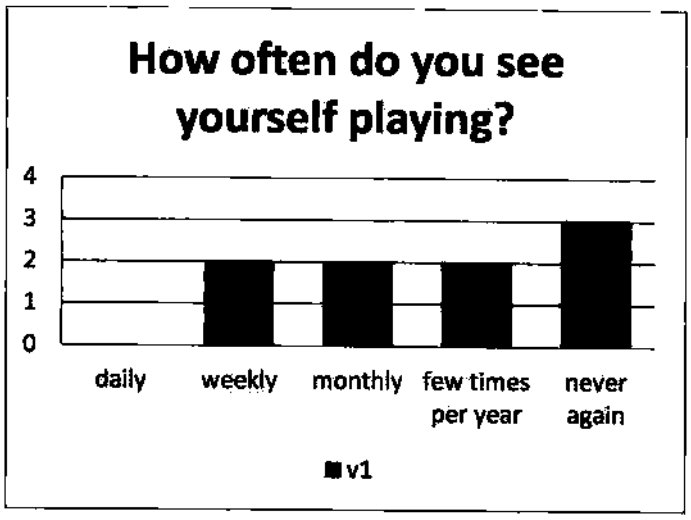

Figure 3: Anticipated frequency of play

3. How to make it more fun – WordSleuth version 2

Our game was functional, but as the responses from our first round of beta testers confirmed, it was not an enjoyable enough experience to cause most people to willingly return to it. In order to get people to want to play, von Ahn (2008) suggests using one of three general game templates when creating a GWAP. The first he calls output-agreement, where two players are given the same input and try to create the same output (seen in the ESP game: http://www.gwap.com/gwap/gamesPreview/espgame/). The second type he calls inversion-problem, where one player is given an input and has to produce output related to this input. The other player then has to guess the original input. This is the template Verbosity (http://www.gwap.com/gwap/gamesPreview/verbosity/) follows. The third he calls input-agreement, where two players are possibly given the same input, but must produce output to determine if their input matches or not. This can be found in his game Tag a Tune (http://www.gwap.com/gwap/gamesPreview/tagatune/).

WordSleuth most closely matches the second template, "inversion-problem." However, all three of von Ahn's GWAP templates assume synchronous play, while WordSleuth uses asynchronous gameplay. He does discuss briefly adapting this so that, sometimes, a single player can participate alone, but this is typically with a pre-recorded set of instant responses. The basic structure remains the same: two "entities" actively engaging for a brief period of time. In WordSleuth, however, one player is given the input of a tag and produces a sentence

guess percentage to the average guess percentages of all users. The calculation is similar to that of an actual IQ, and is based on the following formula:

$$120 + \left(\frac{g - ag}{sg} * 15\right)$$

g = a particular person's guess accuracy
ag = population's average guess accuracy
sg = population's standard deviation of guess accuracy

The two expressive scores, however, are not instantaneous. This is due to the fact that they are calculated based on how frequently *other users* correctly guess a given user's created messages. When this happens, the expressive score increases (also by 15) and the expressive IQ is recalculated in the same fashion as the receptive one. By creating more sentences, a user was able to increase the chance of their sentences showing up for other users to guess, but they did not receive any instant benefit for doing so.

The addition of a new score was supported by the stated desires of our beta testers. In response to the question "What part of the game did you enjoy least?" two of our nine first round beta testers complained:

> "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."

Thus, to motivate users when they were creating messages, activity points were added. This is a count of simply how much a user has done anything on the site. It is calculated as the total number of guesses plus the total number of creations. Even if a user's expressive scores are not showing any change when they are creating in WordSleuth version 2, they now receive the instant feedback of activity points going up.

Figure 4: The new score bar

As all of von Ahn's GWAPs give instant feedback, and it seemed to be an important motivational factor, it seemed prudent to integrate this new score. One example of a game that gives instant feedback is Chess Tempo (http://chesstempo.com/). This game has a large number of chess puzzles of various difficulties that it presents to you based on your current ranking. After each single puzzle is attempted, your rank and the puzzle's difficulty levels are

both immediately adjusted based on whether you successfully solved the puzzle or not. This is satisfying, and a good motivation to keep playing, as each puzzle is brief but provide instant score changes.

Additionally, this new activity points score matters in a new and interesting way in this second version: it is used to unlock new variations in WordSleuth gameplay.

3.2 Unlockable features

Unlockable features were introduced for the purpose of giving people small, tangible, achievable goals. These goals are not just trivial, either, because the unlockable features are game play modes and difficulty levels. Unlockable difficulty levels are easy, medium, and hard. The unlockable game play mode is create mode (guess mode starts off by default unlocked). By making some features unlockable, it gives us the power to force the new player to stagger their play of the different parts of the game (instead of exploring it all quickly and growing bored).

Beyond just increasing the player's fun by giving them tangible rewards every time they play for a certain amount of time, this process also helped increase our data integrity. Most players in version one who created before doing any guessing started off writing low quality messages. They just weren't sure exactly how we were asking them to portray the social tag. A common mistake was an attempt to define the tag, or to make a general statement about it. By forcing the user to guess 15 messages first, we make them read examples of what we're looking for, which increases the quality of their creations once they unlock that game play mode.

Malone's (1980) paper on what makes learning fun specifies three categories: challenge, fantasy, and curiosity. This new feature taps into the categories of challenge and curiosity. He says in regards to challenge that it is important to have a goal (an object of the game) that is obvious and compelling, and that people must be able to tell if they're getting closer to the goal. In this case, the goal would be the new unlockable features in version two of WordSleuth, which can be achieved simply through the accumulation of the new activity points score. Additionally, the main game play page tells the user exactly how many more they need before reaching the next goal. This becomes an additional motivator to play when the number becomes small. Unlockable features ties into his curiosity category as well – people want to know what the harder difficulties are like! Their curiosity regarding what will happen in those difficulties drive them to keep playing and unlock the more difficult levels.

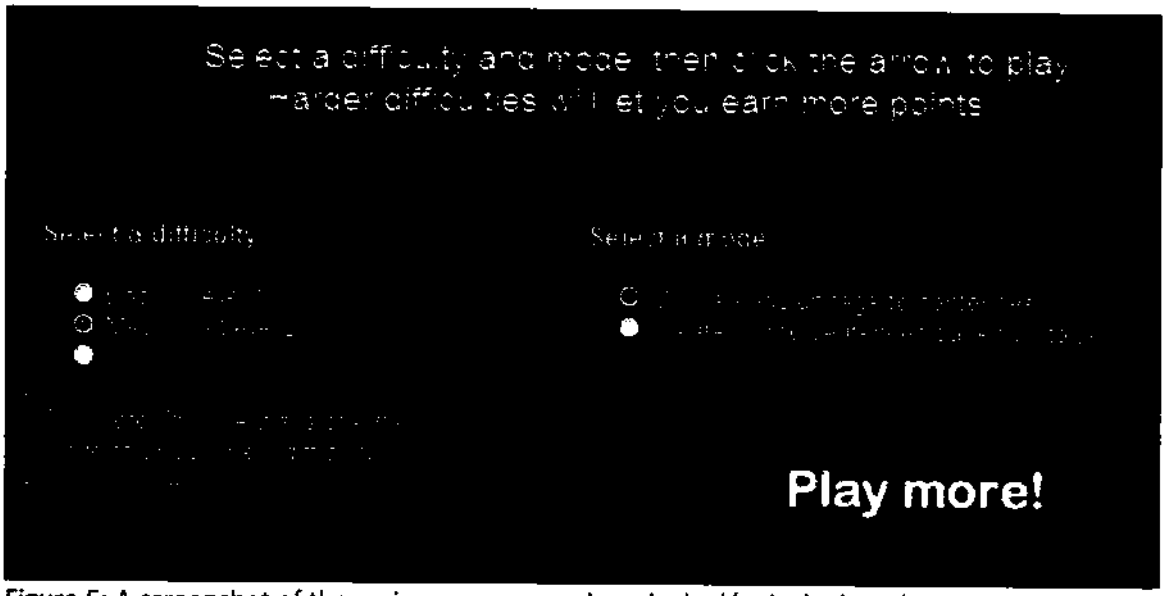

Figure 5: A screenshot of the main game page, where locked/unlocked modes are noted. The user can tell that they have not yet unlocked hard difficulty because the text is grayed out, the associated radio button cannot be clicked, and the highlighted text specifies how many more points are necessary to achieve this. The yellow box was added for emphasis, and does not exist on the actual website.

Large numbers of games have unlockable modes or difficulty levels. One example is Picross 3D for the Nintendo DS (http://www.picross3d.com/), where you can unlock harder methods of play after completing a certain percentage of easier difficulty. Additionally, completing 100% of certain sections unlocks "bonus rounds" which contain slightly different play as a reward. In various Kirby games (throughout many generations of Nintendo systems) (http://nintendo.wikia.com/wiki/List_of_Kirby_games), worlds are unlocked after each one is defeated, providing a new level to play through. In Harvest Moon (http://www.hmfarm.com/), new crops, animals, and people to marry are unlocked as you play through the game. Additionally, you unlock the ability to fish. Many fighting games, such as the Soul Calibur (http://www.soulcalibur.com/) and Super Smash Bros series (http://www.smashbros.com/), let you unlock new "fighters" when you accomplish certain tasks or play for a long enough amount of time. The popularity of this sort of feature in large numbers of games from wildly varying genres demonstrates just how effective of a tool it can be.

Finally, the addition of this feature was not just prudent in application of theory, but a response to a problem initial beta testers expressed having with WordSleuth version 1: "Overall structure is adequate, but non-engaging." This feature works towards resolving this, making the user care more about playing.

3.3 Difficulty levels

Difficulty levels provide a slight variety of game play by providing players the option to perform harder tasks for greater rewards. Three difficulty levels were implemented for each mode. The difficulty levels when guessing are based on the percent of people that have correctly guessed the message. When playing on easy mode, the message is selected from the top third of messages guessed most accurately. Medium mode is associated with the middle third of messages, and the hard mode with the least correctly guessed third.

The difficulty levels for creating messages involve what taboo words the user is presented with. There are two sets of taboo words used for each tag. The first set is static, and made up of words that no message should ever have. These include the social tags themselves, and various forms of the tag that people might try to include that would simply make it too easy to guess. For example, the words "polite," "politely," and "politeness" are all static taboo words for the social tag "politeness." The second set of taboo words are dynamically generated on a weekly basis. All messages are processed, and the most commonly associated words for particular social tags are put in the dynamic taboo list for each tag. This is calculated via mutual information, a process where the mutual dependence of two variables is measured. The formula for this calculation is:

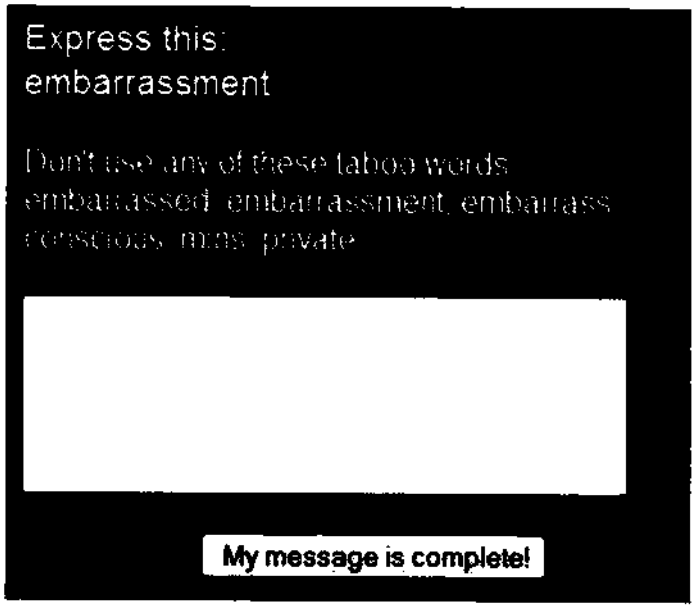

Figure 6: A screenshot of the create mode on medium difficulty. The first three words are the static forms of the tag "embarrassment;" the rest are the three most common words used in items created with this tag.

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p_1(x)p_2(y)}\right)$$

p(x,y) is the joint probability distribution function of X and Y
$p_1(x)$ is the marginal probability distribution function of X
$p_2(y)$ is the marginal probability distribution function of Y

When playing easy mode, the user is given only the static taboo list, giving them a large amount of flexibility in what they can write. When playing medium mode, the user is given the

static taboo list and the three most common words from the dynamic taboo list as well (see Figure 6 above). On hard mode, the user is given the static taboo list and all seven words from the dynamic taboo list.

Each harder difficulty, for both game play modes, provides the user with more points for their actions. Easy difficulty remains the same at 15 expressive or receptive points. Medium difficulty awards 30 expressive or receptive points, and hard difficulty awards 45 expressive or receptive points. Difficulties do not affect activity points or either IQ score. Activity points are supposed to be a cohesive unit, with each single one representing one action on the site, so they do not receive any multipliers. IQs are based on accuracy percentages, regardless of difficulty.

Difficulty levels additionally fall into the challenge category in Malone's (1980) paper, as they relate to "uncertain outcome." This feature allows the users to select how certain they want the outcome of something to be. If they are frustrated with guessing on sentences that seem to fall under multiple social tags (often an issue for hard messages), they can switch to easy difficulty and earn points on those sentences that are more explicit in their social information. For the user that really wishes to be challenged, harder difficulties may be selected which decrease their chances of getting things right and increase the limitations on the messages they create.

Almost all games provide some sort of method to increase their difficulty. Many single player roleplaying games, such as Dragon Age (http://dragonage.bioware.com/) and Mass Effect (http://masseffect.bioware.com/), allow the user to choose from difficulty levels. For example, with Mass Effect you can choose "Casual," "Normal," or "Veteran" (after successfully beating the game on any mode, "Hardcore" is also unlocked, and beating the game on "Hardcore" unlocks "Insanity"). Many first person shooter games provide a similar basic structure. World of Warcraft (http://us.battle.net/wow/en/), a massive multiplayer online roleplaying game, provides two versions of much of its content: "regular" and "hard mode." Regular raids and dungeons are easy to do for casual players. "Hard mode" (sometimes called "heroic") content has additional challenges that can be exceedingly difficult.

Additionally, difficulty levels provide a secondary benefit in terms of the quality of the data we are receiving. When players guess on harder difficulties, we reward them with more points for providing us with more data on difficult sentences. Difficult messages have particular research interest, as they are more ambiguous and provide us with data on how perception of a written message conforms to or deviates from the initial intent of the creator. When players create on harder difficulties, we can also gather a richer variety of messages. For example, say the word "please" is consistently enough for a creator to convey the social tag of politeness.

That is useful, but we want to learn more about what other words or linguistic structures are employed if that specific, common word is forbidden.

3.4 High score tables

High score tables tap into both goal-driven and socially-driven motivation. Five high score tables were introduced in WordSleuth version 2, and more can be added with minimal difficulty. The current five high score tables are basic in nature, corresponding directly to the five scores each user has. Top Receptive Scores is based on the receptive score, Top Expressive Score is based on the expressive score, Most Active Players is based on activity points, Top Receptive IQ is based on receptive IQ, and Top Expressive IQ is based on expressive IQ. A problem encountered during development was that users would guess once correctly, achieving a guess rate of 100%, and shoot instantly and trivially to the top of the high score table. To prevent this, a feature was implemented where users do not unlock the IQ score until they have guessed or had their created messages guessed 15 times.

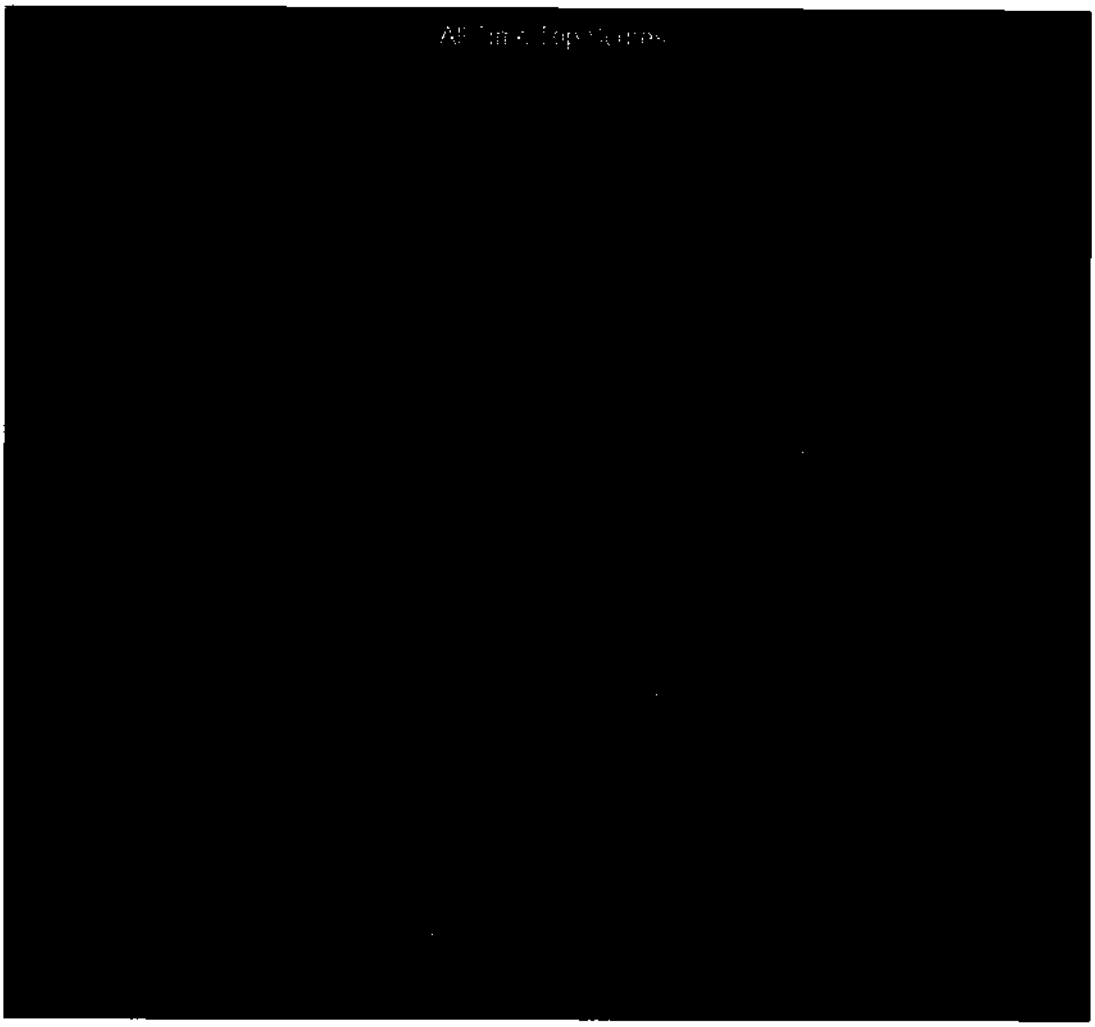

Figure 7: A screenshot of the high score table page

Peter Vorderer (2011) calls this aspect of a game "social competition" (to distinguish it from the competition inherent in a person challenging a computer). He claims that doing well in social competition can cause a user to have increased self-esteem and can result in a better mood, causing positive feelings towards the game. Additionally, "participation in challenging and competitive situations appears to be an important reason for the enjoyment felt by computer game players". The combination of these factors provides motivation for users to partake in social competition. Our newly working high score tables provide some amount of social competition to the general user of WordSleuth. They are finally able to compare their scores to other peoples, and to see how close they come to being the best in a specific area of the game.

Many enjoyable games employ high score tables to address social competition. Arcade games are the classic example. These are played by one or two people, then afterwards if the score is sufficiently high they are able to enter their name and anyone who plays in that arcade later will see it. A culture of fierce competition was common around some of these games before home systems became more popular. Bejeweled Blitz is a Facebook game (http://www.facebook.com/bejeweledblitz) centered on being the highest scorer out of your friends. The high score table is reset every week, forcing a person to play frequently if they want to remain on the top. They are then encouraged to post this achievement on their Facebook profile for all their friends to see. Audiosurf (http://www.audio-surf.com/) is a computer game downloadable from Steam (http://store.steampowered.com/) where levels are created from songs in the player's music library. Each song has a high score table associated with it.

All of the above features were added based on the existing research, an examination of popular games, and the expressed desires of the survey takers of our first round of beta testers. They marked a fairly significant upgrade for the game overall, and it was decided at this point to release this second version of WordSleuth.

3.5 Second round beta testers

We once again decided not to release this version of the game to the public. Our second group of beta testers, however, was a little broader in scope. It consisted of those who had played the first round, as well as friends, family members, and students who Professor Pearl and I were less close to. These students were contacted with the following request:

The second launch of WordSleuth is now ready and open for beta testers!

Some of the exciting new features include:

- High score tables
- Unlockable difficulty levels
- New "activity points" score
- Improved IQ calculations

Please go to http://gwap.ss.uci.edu/ to make an account and start playing. **(NOTE: please create a new account if you played previously, so you can experience the new version of the beginner game).** Play as frequently as you wish over the next few days, then fill out the new beta tester survey. If you don't have a lot of time to commit and want to do this all in one go, even playing 15 minutes of each mode (or for the goal oriented: guess 30 sentences and create 15) and taking the survey would help significantly.

[...information on how to find the survey...]

If you encounter any errors, please note the details of them. You may either save these notes and submit them to me in the survey, or email the issues to me (gtucker@uci.edu) as you encounter them.

The results of the beta testers are as follows (purple bars are for version 2):

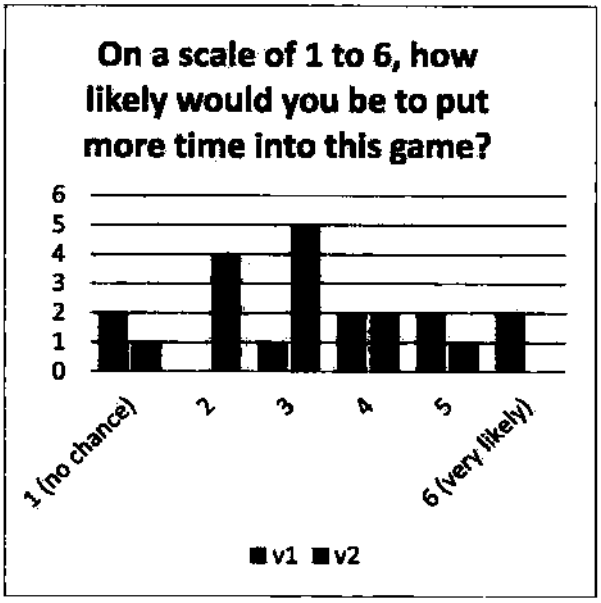

Figure 8: Further time commitment

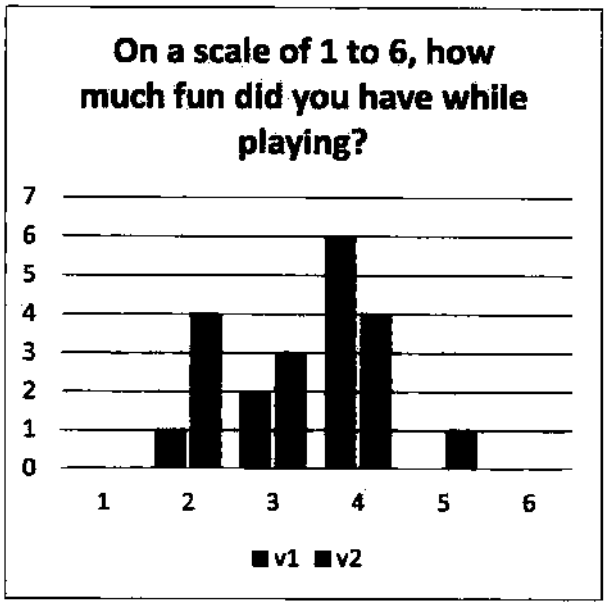

Figure 9: Enjoyment

These graphs show the results of three key questions on the survey. In total, 12 people took the survey. 6 of these people had played the previous version of WordSleuth, and 6 of these people had not. At this point, the players did not necessarily have an investment in linguistic research or a close friendship with those working on the game. Thus, there was less of a positive bias towards the game in this survey. This is a likely cause of the drop in average scores from the first survey. The average number selected for putting more time into the game (on a scale of 1 to 6) dropped from 3.88 to 2.83. The average amount of fun (also on a scale of 1 to 6) dropped from 3.55 to 3.16. One positive item worth noting is that the most common answer to "How Often Do You See Yourself Playing" is now "a few times per year" instead of "never again."

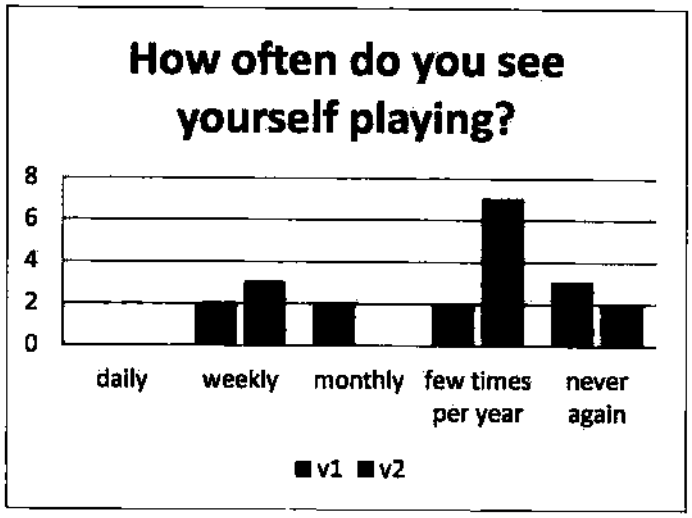

Figure 10: Anticipated frequency of play

We believe that the drop in numbers is strongly tied towards the lesser bias of our new participants, and not that our improvements made the game "less fun." The freeform comments from the survey support this belief. The six players who had played the previous version of WordSleuth were asked what their favorite and least favorite differences were. Some 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."

As far as least favorite differences went, most people stated that they had none, or erroneously reported differences that did not exist (such as disliking a change of pictures, which had not occurred). One person did note that: "Frustrating higher difficulties were annoying." However, this is not a complaint about the existence of higher difficulties. Instead, it is a request for an alteration of how we determine what is "hard."

One goal of ours is still to improve these average numbers. However, given how small our set of sample players is, we feel that the feedback given in comments regarding specific features is more important than these numeric results.

4. How to make it more fun – WordSleuth version 3

There were further improvements that could be simply implemented for substantial benefits. These included a slight change to a previously added feature, new features to address a part of the game that people found frustrating, and two more features to directly increase enjoyment.

4.1 A change to point allocation

One thing added to WordSleuth version 2 was activity points, partially for the benefit of providing people with instant feedback when creating sentences. However, as was noted previously, the decision was made for activity points to not multiply when playing on higher difficulty levels. This meant that users had no immediate reward for creating messages on harder difficulties, despite the fact that they may eventually earn more expressive points when people guess on them.

The solution that was implemented in version 3 was to give users a small amount of expressive points directly for creating messages on harder difficulties. On easy, a user receives no instant points, and earns 15 when people later guess their message correctly. On medium, a user now receives 5 instant points, then 30 when people later guess their message correctly. On hard, a user receives 10 instant points and 45 when people later guess their message correctly. This keeps the basic structure of scoring intact while still giving users enough incentive to play harder difficulties while creating.

4.2 Ambiguous tags

One problem commonly identified by beta testers of both versions is the fact that some sentences were very ambiguous, or even viewed as outright wrong by the guesser. Here are some answers to the "What part of the game did you enjoy least?" question:

- "The distinction between 'formality' and 'politeness' is perhaps too subtle."
- "The fact that some of the messages weren't very fitting for their tags/were ambiguous."
- "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."

In fact, 57% of all responses to this question in the first and second survey combined touched on this issue in some way. This was clearly a problem of high priority. Three features in an attempt to address this problem: letting people skip messages, tweaking the create message process, and flagging.

4.2.1 Skipping messages

The first is a very straightforward solution. If people are faced with a sentence that seems too difficult or they just don't feel comfortable guessing, they may click a "skip" link and proceed to the next question. This did not technically add new functionality, as the user was able to refresh the guessing page at any time and would have been presented with a new question. The addition of the link made the functionality more transparent, however, and available to users that would not think to try refreshing. This may end up having the effect of slightly decreasing the amount of guesses on the hardest sentences of each difficulty bracket; however, the additional amount of data we should receive from users playing longer due to now having direct control over their questions seems well worth the trade-off. This simply allows people to ignore the problem, however, and doesn't seek to really solve it. The other two solutions attack the cause and effect of the problem more directly.

4.2.2 Modifying the create message process

In order to address the cause, the create page was altered. Specifically, it was changed so that the user creating a message is shown all tags, and not just the one they are creating a message for. They are advised to be careful about ambiguity, and to do what they can to make sure their message applies more to their given tag than to any of the other options. This approach was suggested by a beta tester that wrote the following:

> "I think that when creating a caption, the user should be able to view all 8 possible tags and see the one they have to use highlighted. This way they can make sure that their caption is most like the one they are aiming for rather than another one. Otherwise they can make a caption that works for both, which can make the game very frustrating."

This solution of redesigning the create page was particularly strong because it attacked the root of the problem, instead of just providing ways for a guesser to get around ambiguous messages.

Figure 11: The new create page. Red and yellow transparent boxes are not visible on actual website. The three items boxed in red were added in WordSleuth version 3. The one item boxed in yellow was reworded to more clearly address the issue of ambiguity.

4.2.3 Flagging messages

One final feature was implemented to attack the questionable messages themselves. While some messages are legitimately ambiguous, where their social information varies person to person, there are some messages that are just plain bad. Specifically, a message counts as "bad" if it falls under any of the following categories:

- **The message is nonsensical or very poorly spelled.**
- **The message gives away the tag in an explicit fashion. (Example: there was a message for "deception" that included a claim followed by "(he is lying)")**
- **The message very strongly does not match the given tag at all.**

As our content is user-created, it is not feasible for moderators to individually examine every message sent to the database. Instead, we implemented a way for users themselves to report "bad" messages through a flagging feature. After guessing an item, on the results page (where the user is shown what they guessed as well as the correct answer), the user may click a new link: "Click here to flag this sentence." This takes them to a page where the above conditions for a bad sentence are listed. They are told emphatically **not** to flag messages with less serious problems. Specifically, they are told to ignore messages with minor spelling or grammar errors, and messages that they think match another tag better, but may be ambiguous.

If a user chooses to flag a message, this is saved in the database. Additionally, their guess is removed from the database and from their score (although this is not noted explicitly on the website itself, to minimize the potential of abuse by users flagging every message they guess wrong). If a message reaches a threshold of flags (the raw number of flags must be above a certain number – currently 10, and the number of flags must then be a certain percentage of total guesses for that item – currently 33%), then it is changed to inactive in the database. This means it is taken instantly out of circulation on the website, but the message itself and all corresponding guesses still exist. Any moderator may view the bad messages, and then choose to either reactivate some manually or run a stored procedure to remove all and any associated guesses.

4.3 Achievements

The motivation contained in unlocking new game play modes can only last so long. Even the hardest game play aspect to unlock must be achievable with only moderate difficulty, or people will get frustrated and give up before reaching it. However, there is a subset of users that enjoy having incredibly difficult goals to work for. The average player would also benefit from having random goals that they may choose whether or not to pursue with no gameplay repercussions.

These goals were implemented in the form of achievements: pictorial awards earned through accomplishing anything from a trivial to incredibly difficult goal. An example of an easy achievement would be "guess 25 messages." A difficult one might be as time-consuming and skill-dependent as "guess 100 messages in a row correctly." To account for the differences in difficulties between achievements, each achievement additionally has a certain amount of "achievement points" associated. The easy achievement noted above might be worth 5 of these, where the difficult one could be worth as much as 50. This score is visible not on the score bar, but can be seen on profile pages (to be discussed shortly). Additionally, it may be used for a new high score table in the future.

Mikael Jakobsson (2011) argues that adding achievements to a game is actually adding a second level MMO (massively multiplayer online game) on top of the original game. He discusses this in the context of the Xbox360, a latest generation gaming console that was wildly successful in part because it required all games released to include a certain amount of achievements that a player can unlock. Any Xbox360 player can unlock one of these achievements through average play of the game, and view what they have to do to earn more of them. Other users of the Xbox360's online system can view what achievements their friends have earned (or strangers, if they know the person's gamer tag). He writes specifically about the benefits he has seen from achievements. For the average, casual player, he has observed the following:

> "It is not until I have finished a game but want to continue playing it the achievements come into play in a significant way. ...achievements provide a [strong] sense of optional unfinishedness. I can convince myself that further engagement with the game is reasonable and worthwhile although I have reached the formal end, because the achievement scaffolding stretches further and provides a direction. ... The achievements need to matter enough that we can use them as motivations to continue as long as we enjoy playing a game, but not so much that we feel forced to continue playing when the pleasure is gone."

Many games successfully implement achievements in a way that drives users to play ridiculous amounts. Other than the Xbox360 online system, two popular examples spring to mind. The first is Farmville on Facebook (http://www.facebook.com/FarmVille), which rewards diligent play with ribbons and occasionally little banners you can put up on your farm. The second is the game World of Warcraft (http://us.battle.net/wow/en/), which incorporated achievements two expansions ago with the release of The Wrath of the Lich King.

4.4 Profile Pages

To improve the community aspect of WordSleuth, and to provide an easy way for users to view the achievements they have collected, profile pages were created for each individual user. The information displayed on this page is all known scores, how many items have been guessed / created, achievement points, and a visible representation of all earned achievements in an aesthetically pleasing manner.

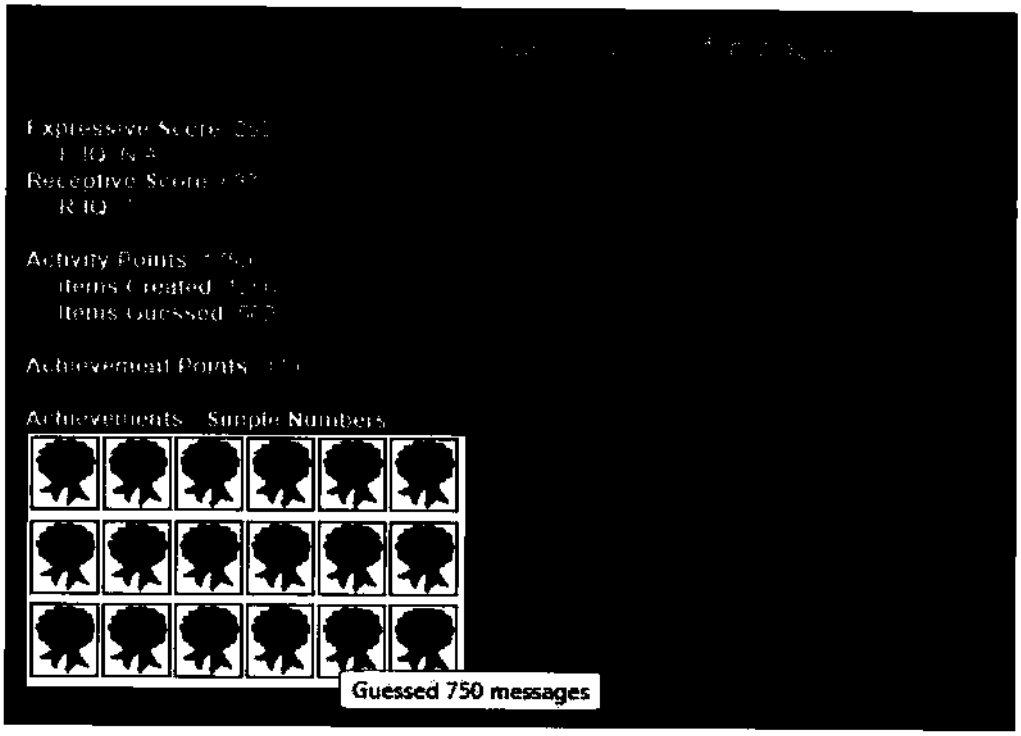

Figure 12: When a ribbon is hovered over, you can see why you earned the achievement (the mouse was hovering over the second to last purple ribbon when screenshot was taken). This user has earned all 18 currently earnable achievements. Missing achievements are denoted by a ribbon outline, which, when hovered over, informs the user what they need to do to earn it.

The average user greatly enjoys being able to view their statistics, or those of their friends. This ties in with Vorderer (2010)'s social competition drive previously discussed in regards to high score tables. People begin to feel more attached to a website once they have a page that belongs to them. In fact, the principle of being able to maintain your own profile page and view those of your friends is the foundation for a website as popular as Facebook!

4.5 Third Round Beta Testers

In this version, we decided to release the game to the public. There has been no strong effort to advertise, but a few placements allowed us to get some testers who did not necessary know the creators of the game. Specifically, a link to the game was put on Professor Pearl's research website, and a short description requesting players was played in UCI's honors program newsletter.

Additionally, all previous beta testers were contacted with the following request:

The third launch of WordSleuth is now ready and open for beta testers!

Some of the exciting new features include:

- The ability to flag bad messages
- The ability to "skip" guesses / creations that you don't want to answer
- Achievements!!!
- Profile pages

Please go to http://gwap.ss.uci.edu/ to make an account and start playing. **(NOTE: please create a new account if you played previously, so you can experience the new version of the beginner game)**. Play as frequently as you wish over the next few days, then fill out the new beta tester survey. If you don't have a lot of time to commit and want to do this all in one go, even playing 15 minutes of each mode (or for the goal oriented: guess 30 sentences and create 15) and taking the survey would help significantly.

[...information on how to find the survey...]

If you encounter any errors, please note the details of them. You may either save these notes and submit them to me in the survey, or email the issues to me (gtucker@uci.edu) as you encounter them.

The results of the beta testers are as follows (light blue bars are for version 3):

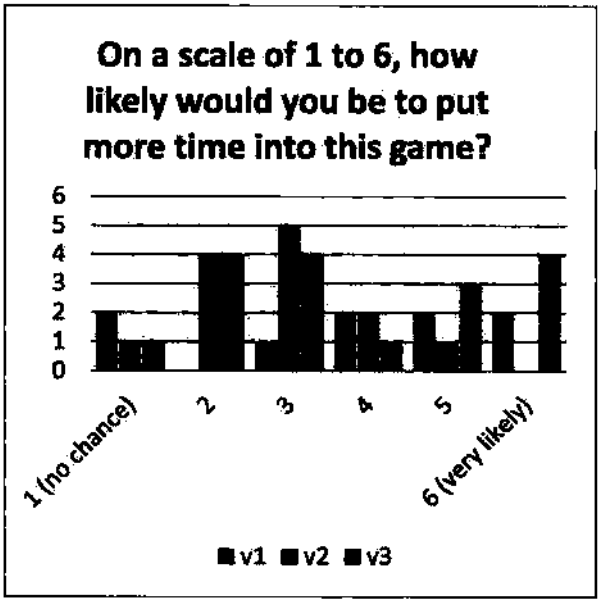

Figure 13: Further time commitment

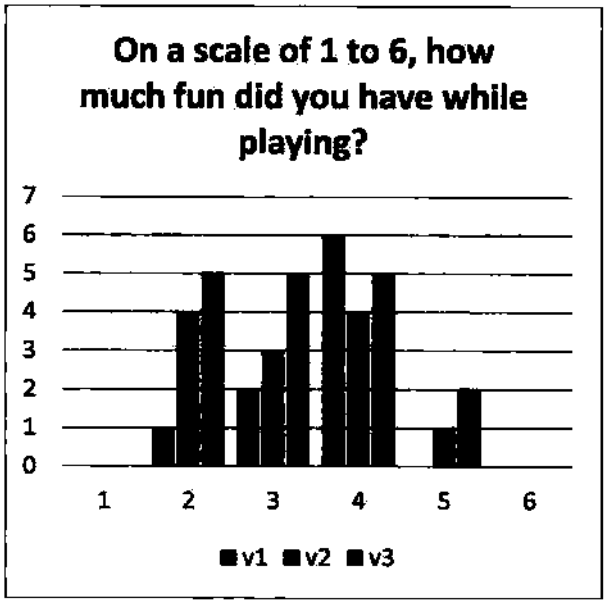

Figure 14: Enjoyment

These graphs show the results of three key questions on the survey. In total, 17 people took the survey. 6 of these people had played the previous version of WordSleuth, and 11 of them had not. Once again, there was likely less of a positive bias in this case than in the first survey. The average number selected for putting more time into the game (on a scale of 1 to 6)

rose from v2's score of 2.83 to a 3.76, almost returning to the biased number from survey 1 of 3.88. The average amount of fun (also on a scale of 1 to 6) rose as well from v2's score of 3.16 to a 3.24. It did not quite reach version 1's 3.55, but was still an improvement. dropped from 3.55 to 3.16. Another improvement worth noting is that more people selected "weekly" and "monthly" in "How often do you see yourself playing?" than ever before.

The freeform comments from the survey support the numeric evidence that v3 was an improvement. Some notable comments were:

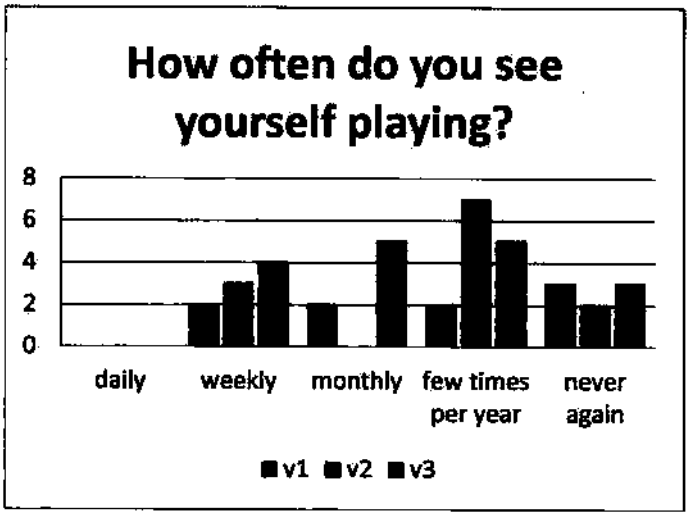

Figure 15: Anticipated frequency of play

> "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."

Also found in the comments was an interesting backlash against the new feature of skipping. When asked about their least favorite part of the game (or their least favorite difference, if they played previously), some responses like the following were observed:

> "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"

It seems as if there were people who felt that the skip feature made it too easy for them – that they were just able to breeze over everything they weren't sure about. Players that were familiar with the overall purpose of the game additionally didn't understand how their work could possibly be beneficial to research when they're able to avoid answering anything that strikes them as even a little hard. However, the fact remains that there will always be someone willing to answer a question, and the research benefit from having *more* data far outweighs the

consequences of losing a few answered questions on the hardest messages of each difficulty category. Also, while the hardest messages may be informative for human perception data, they are not ideal for a database to train machine learning algorithms on – in that case, we would only want messages whose social information is clear, and therefore much less likely to be skipped.

5. Future Extensions

There are many features that we have brainstormed and would like to add to future versions of WordSleuth to further enhance the amount of enjoyment players get. In order to increase the social competition aspect, we would like to add basic social networking (allowing people to add friends, send messages, compare profiles). We would also like to provide users with trivia related to their guesses and creations. For example, if a user guesses incorrectly on a message, they might be presented with a fact saying "Although you were wrong, 65% of people selected the same answer as you." This would support Malone's curiosity motivation (1980). He also discussed a third motivation that we have not explored much yet: fantasy. This is something we would be interested in trying to integrate in the future through visual rewards for playing, and perhaps some story-mode elements. A specific, easy place to start might be allowing users to unlock various titles that can appear next to their username for reaching certain goals in the game. We would also like to expand the content of some features integrated over the last few versions, adding more interesting and complex high score tables and more diverse achievements.

6. Conclusion

The amount of data we have generated so far makes the GWAP method of data collection seem very successful so far. At the time of writing, WordSleuth has been playable now for six months. Much of this was in a limited capacity and restricted to people we had hand-selected. Our 56 seed members are now 171 registered members (although some of this growth is due to asking users to create new accounts to test new versions). Our 2,873 seed guesses have grown to 8,569 total guesses. Our 1,060 seed messages have grown to 2,191 total messages. If this sort of growth is possible in the process of simply getting the game to the point where we feel comfortable advertising it heavily to the public, it seems very promising that the rate of data collection will increase significantly over the lifetime of WordSleuth.

We believe that the addition of new features discussed above has had a substantial effect on how quickly we've been acquiring new data. With our rapidly growing database, the applications of this social language information repository begin to seem much more plausible in the near future. With a large amount of data, certain interesting practical and theoretical applications may be investigated. Specifically, cognitive science models on human perception

and transmission of social information can be tested. Machine learning algorithms can be trained on this data, and then be able to automatically recognize the tone of a message. This could be useful in a variety of programs, ranging from automatic tone analyzers of pre-written blocks of texts (such as webpages or articles) to learning programs created to help people who find it difficult to understand tone (such as people suffering from Asperger's syndrome) improve the quality of their life. All of this is all possible because people now enjoy playing a simple internet game like WordSleuth.

7. References

Diab, M., Dorr, B., Levin, L., Mitamura, T., Passonneau, R., Rambow, O. and Ramshaw, L. 2009. Language Understanding Annotation Corpus. LDC, Philadelphia.

Jakobsson, M. 2011. The Achievement Machine: Understanding Xbox 360 Achievements in Gaming Practices. *Game Studies – the international journal of computer game research*, 11(1)

Malone, T.W. 1980. What Makes Things Fun to Learn? Heuristics for Designing Instructional Computer Games." *SIGSMALL '80 Proceedings of the 3rd ACM SIGSMALL Symposium and the First SIGPC Symposium on Small Systems*: 162-169. ACM Digital Library.

Mason, W. and Suri, S. 2010. Conducting Behavioral Research on Amazon's Mechanical Turk. Social Science Research Network.

Pearl, L. and Steyvers, M. 2010. Identifying Emotions, Intentions, & Attitudes in Text Using a Game with a Purpose. *Proceedings of NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.* Los Angeles, CA: NAACL.

Thompson, C. 2011. "For Certain Tasks, the Cortex Still Beats the CPU." *Wired.com*. 25 June 2007. Web. <http://www.wired.com/techbiz/it/magazine/15-07/ff_humancomp>.

von Ahn, L. 2006. Games with a Purpose. *IEEE Computer Magazine*, June 2006: 96-98.

von Ahn, L., Ginosar, S., Kedia, M. and Blum, M. 2006a. Improving Accessibility of the Web with a Computer Game. *ACM Conference on Human Factors in Computing Systems, CHI Notes*: 79-82.

von Ahn, L., Kedia, M. and Blum, M. 2006b. Verbosity: A Game for Collecting Common-Sense Knowledge. *ACM Conference on Human Factors in Computing Systems, CHI Notes*: 75-78.

von Ahn, L., Liu, R. and Blum, M. 2006c. Peekaboom: A Game for Locating Objects in Images. *ACM Conference on Human Factors in Computing Systems, CHI*: 55-64.

von Ahn, L. and Dabbish, L. 2008. General Techniques for Designing Games with a Purpose. *Communications of the ACM*, August 2008: 58-67.

Vorderer, P., Hartmann, T. and Klimmt, C. 2003. Explaining the Enjoyment of Playing Video Games: the Role of Competition. *ICEC '03 Proceedings of the Second International Conference on Entertainment Computing*: 1-9. ACM Digital Library.

Zhou, L. and Sung, Y. 2008. Cues to the deception in online Chinese groups. *Proceedings of the 41st Annual Hawaii international Conference on System Sciences*, 146. Washington, DC: IEEE Computer Society.