Human Computation
AAAI Technical Report WS-12-08

# Doodling: A Gaming Paradigm for Generating Language Data

**A. Kumaran**
Microsoft Research
a.kumaran@microsoft.com
Bangalore 560 025, India

**Sujay Kumar Jauhar**[1]
University of Wolverhampton
Sujay.KumarJauhar@wlv.ac.uk
Wolverhampton WV1 1LY, UK

**Sumit Basu**
Microsoft Research
sumitb@microsoft.com
Redmond, WA 98052, USA

## Abstract

With the advent of the increasingly participatory Internet and the growing power of the crowd, "Serious Games" have proven to be a fertile approach for gathering task-specific natural language data at very low cost. In this paper we outline a game we call Doodling, based on the *sketch-and-convey* metaphor used in the popular board game Pictionary®[2], with the goal of generating useful natural language data. We explore whether such a paradigm can be successfully extended for conveying more complex syntactic and semantic constructs than the words or short phrases typically used in the board game. Through a series of user experiments, we show that this is indeed the case, and that valuable parallel language data may be produced as a byproduct. In addition, we explore extensions to this paradigm along two axes – going online (vs. face-to-face) and going cross-lingual. The results in each of the sets of experiments confirm the potential of Doodling game to generate data in large quantities and across languages, and thus provide a new means of developing data sets and technologies for resource-poor languages.

## Introduction

Crowdsourcing has been shown to be an effective paradigm both for solving problems that are computationally hard and for those requiring extensive data creation and labeling (Callison-Burch and Dredze 2010, Ambati and Vogel 2010, Zaidan and Callison-Burch 2011). Such a model holds key advantages for language data generation over the traditional approaches (i.e., using existing parallel corpora or a small set of expert translators) because of the promise of attracting a wide online audience, with immense demographic diversity in terms of languages and interests. Many flavors of crowdsourcing paradigms exist, including the *for-pay* model (Bloodgood and Callison-Burch 2010, Callison-Burch 2009, Irvine and Klementiev 2010) where the contribution is for monetary rewards (*e.g.,* through Amazon's Mechanical Turk), the *for-recognition* model, where the contribution is made for individuals' visibility in a community (*e.g.,* SourceForge), and the *common-good* model (Kumaran et al. 2009, Wikibhasha 2010), where value is produced for the benefit of the community, such as Wikipedia. In this paper, we explore another well-established crowdsourcing paradigm, the *for-fun* model (Chen and Dolan 2011, Cooper, et al. 2010, Duolingo 2011, Law et al. 2009, Von Ahn and Dabbish 2004, Von Ahn, Kedia, and Blum 2006), in which data is a by-product of gameplay. Such games are often referred to as "Serious Games" or "Games with a Purpose" (GWAP) (Von Ahn and Dabbish 2008), and have been shown to be very successful in domains such as photo tagging or linguistic ontological annotation.

We propose Doodling, a *sketch-and-convey* game, which parallels the popular board game, Pictionary®, in which information is conveyed using hand-sketched doodles by one of the players. The other player guesses at the word or phrase represented by the doodle, and this re-surfacing of the sketched concept allows for the production of syntactic variants in the same language or parallel data in another language, depending on whether the game is played between players in the same or different languages. We believe that the Doodling game benefits from a familiar metaphor, real-time human social interactions, the potential for players to rapidly improve their language skills, and above all, the possibility of being fun. Because there are many language pairs for which there is no strong financial incentive for developing language technologies, we believe engaging the online population in volume and on a volunteer basis may be the only way to gather the necessary data. We note that the Duolingo system (Duolingo 2011) is in a similar vein, as it is also designed towards collecting parallel language data from the crowd. However, Duolingo attracts its users by promising an educational benefit, i.e., helping them to learn a language, whereas our system promises entertainment, and is thus a different approach to the problem. To the best of our knowledge, there is not yet any published information on the effectiveness of the Duolingo approach, and thus we cannot compare to it directly.

In this paper, we present the design elements of our online sketch-and-convey game and explore its potential to generate monolingual and multilingual data. We specifically explore the following research questions relating to the sketch-and-convey paradigm in this paper: first and foremost, can the sketch-and-convey paradigm be effective for

[1] Research conducted while an intern at Microsoft Research.
[2] Registered trademark of Hasbro.

complex language structures beyond the simple words and phrases of the board game, while still retaining the fun element? If so, we wish to extend further questions along two independent axes:

- Can an online gameplay scenario be as effective as the face-to-face setting (in terms of objective measures such as accuracy as well as in terms of the user experience)?
- Can the game be as effective and fun in multilingual scenarios in conveying concepts, despite the greater linguistic and cultural complexities and challenges?

## The Doodling Game

We present here the design elements and the game flow of the Doodling game as well as metrics for its effectiveness.

### Doodling as a GWAP game

Pictionary® is a popular game played and enjoyed around the world in which players must communicate words (or very short phrases) to one another using only hand drawn sketches. In Doodling, our primary intuition is that the user-sketches provide a (largely) language-independent means of communication of concepts between players, which could be effectively employed for the generation of paraphrase data (in a given language) or parallel data (between different languages), especially for sentences and complex concepts. In a crowd-sourcing scenario, it renders possible a way for re-surfacing textual elements in different syntactic forms or languages, via a semantic equivalent inferred from the user sketches.

(Von Ahn and Dabbish 2008) describe three base templates upon which GWAPs may be designed. We note that the Doodling game subscribes to the Inversion Problem Template, where Player 1, given some input produces an output, from which Player 2 must guess the original input. In addition, we note that the game we propose fulfills the traditional requirements for a successful GWAP, namely:

- It promotes the resolution of the underlying computational problem (i.e., the generation of language data).
- Game rounds are solvable in a relatively short time.
- It has the potential to be fun.

### Doodling: Game Flow and Design

Doodling is played as a set of game rounds between two players, each of whom alternate between the role of a Drawer *D* and a Guesser *G*. A sketch of the interface for the proposed game is in Figure 1.

In a given game round, *D* receives a text element *Q* (a word, a phrase, or a sentence) that must be conveyed to *G*, using only sketches for communication. In the particular example in Figure 1, the *Q* given to *D* is "*How do I get to the international airport?*", as shown in the box on the top. *D* sketches a series of elements in the canvas, a plane, a runway, a globe, etc. to convey the partial concept "*international airport.*" For her/his part, *G* re-surfaces the concepts she/he guesses, in the box below the sketch pane. The input *Q* is gradually built up by *G*, which could be critiqued non-verbally by *D*, using the meta-information icons shown in the top right corner of the Figure 1. These icons represent some additional information ('*wrong direction*', '*right direction*', '*abstract*', '*specialize*', '*similar concept*', '*opposite concept*', etc.) that are chosen by *D* and conveyed to *G* to guide her in refining the guesses to match the input. When *D* determines that a semantically equivalent guess is produced by *G*, the round ends. The game continues to the next round with the roles of *D* and *G* switched.


*Figure 1: Sketch of the Doodling interface*

In essence, the game produces two surface forms of a single semantic intent (the input *Q* to *D* and the guess by *G*) that have a relationship similar to that of the input-output pair in the "noisy-channel" model used often in Machine Translation. Effectively, in this scenario, the text element *Q* is passed through a noisy channel – a sketch created by the *D* – which possibly results in some interference and causes it to be re-surfaced by the *G* as a paraphrase (in a monolingual setting) or a translation (in a multilingual setting).

### Domain and Data Used

For the purposes of our experiments, we restricted all text elements to the travel domain, one we felt most players would relate to and would also be a high value domain for multilingual data. We used the travel phrase book from the popular site WikiTravel,[3] which contains around 400 templates, as a seed to produce a standard set of common words, phrases, and sentences that are used in the travel domain. We produced surface forms for all templates by

[3]http://wikitravel.org/en/wikitravel:phrasebook_template

filling in the blanks with an appropriately generic noun, taking care not to use language or culture specific nouns. For example, a template, "*I would like to order ____.*" was completed with "*a cup of cold coffee*", rather than "*Frappuccino.*" We chose text elements from the corpus that were evenly distributed in two different criteria: difficulty (simple, medium and hard), and granularity (word, phrase or sentence). This corpus was used for all of our experiments. Table 1 provides a sample of the corpus used for our experiments.

| **Text Element** | **Diff.** | **Gran.** |
|---|---|---|
| *Taxi* | Easy | Word |
| *Cuisine* | Medium | Word |
| *Ethnicity* | Hard | Word |
| *Cheese Omelette* | Easy | Phrase |
| *Museum of Modern Art* | Medium | Phrase |
| *I would like a bowl of soup.* | Easy | Sentence |
| *What time does the beach close?* | Medium | Sentence |
| *I am sorry, Officer!* | Hard | Sentence |

*Table 1: Example text elements from the travel corpus*

## Initial Pilot Experiment

The initial experiments for Doodling were conducted among the authors, primarily to test the feasibility of conveying complex text elements, and to design a simple game interface for scaled-up experiments. We selected about 10 random elements of different granularity and hardness, and the completion criterion was that the guesser produced a guess judged to be correct by the drawer. Between the three authors, 30 rounds of games were played, producing results that were exact or near-equivalent word/phrase or sentence in games that were under 2 minutes in length in 95% of the cases. This pilot encouraged us to undertake experiments with a larger selection of text and a larger population of users as described in the next section.

## Experiment 1: Monolingual, Face-to-Face

As with the pilot, our primary focus in this set of experiments was to measure the effectiveness of the game when using more complex data (words, phrases, and sentences) than that found in the board game, this time with a larger number of players and trials.

### Experimental Setup

Our experimental procedure followed the process described in the game design section, with players participating face-to-face using paper-and-pencil. We randomly chose 45 text elements from our travel corpus distributed evenly in the two dimensions of *difficulty* and *granularity* (5 for every combination). Inputs to drawer *D* were chosen randomly from this set of 45 elements. *D* and guesser *G* were also given a card with meta-information symbols using traffic metaphors (stop, wrong way, etc.) that they could use to communicate information about an element or progress of the game. The game players were 14 volunteers, mostly undergraduate students, well-versed in English, and for many it was their first exposure to the Pictionary game. The volunteers were randomly paired up and asked to play as many rounds of the game as possible in approximately one hour, alternating between the roles of Drawer and Guesser. The *D* was only allowed to use sketches (and not written text), and the *G* could write in any text on the paper, next to the sketches, and such guesses could be critiqued by *D* using the icons from the meta-information palette; however, no verbal communications were allowed. Once *G* guessed the entire text element correctly (as judged by *D*) the round ended; the next round began with the two players alternating their roles. Players were permitted to forfeit a round if they felt that they were not progressing towards convergence, and such rounds were considered failures. At the beginning of every game round, *D* recorded the perceived hardness of the text element (in a scale of 1-3), and at the end of the round, if it completed successfully, the time taken, accuracy of the guess, the perceived hardness of the text element by *G*, and the accuracy of the guess as judged by the *D* (on a scale of 1-3) were recorded. At the end of the session (lasting several rounds), we conducted a qualitative interview in which the player quantified their level of engagement and perceived fun during the play duration (dubbed as the Fun Factor), on a scale of 1-5.

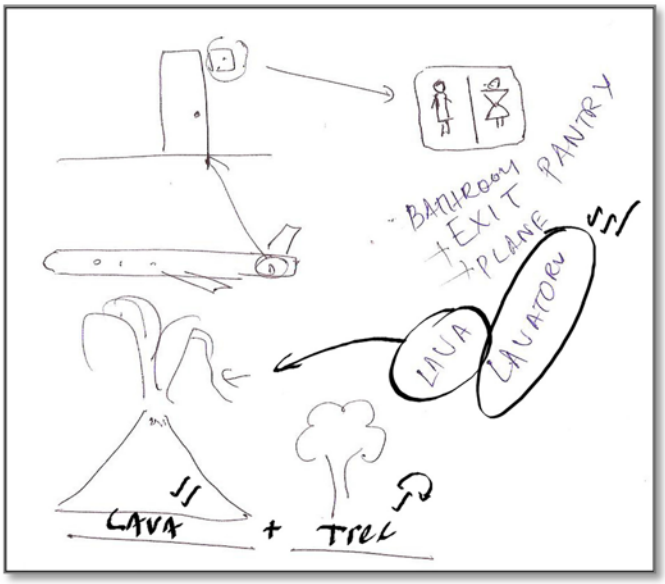


*Figure 2: Sample game from monolingual/face-to-face condition*

A sample of a completed game (on paper) is shown in Figure 3 (conveying the word, "*Lavatory*"). This example illustrates several interesting aspects of the flow of the game. First, *D* has taken the route of conveying by pictures the lavatory at the back of an airplane, perhaps due to the common usage of the word in airplanes. Note that an earlier guess of "*Bathroom*" was not accepted by the *D* (thought it would have been a perfectly acceptable alternative). Finally, the *D* has split the text element into two parts to convey, respectively, "*Lava*" and "*Tree*", which are put together. Note that the word "*Tree*" was used in conjunction with the "*Sounds like*" icon from the meta information palette, to yield the final accepted guess, "*Lavatory.*"

### Analysis & Observations

The results from 103 rounds of the game are shown in Table 3. We also present some qualitative observations about the game play below.

| | |
|---|---|
| Players | 14 |
| Rounds | 103 |
| Success Ratio (completed/played) | 96% |
| Accuracy (in a scale 1-3) | 2.65 (σ=0.50) |
| Average Time (in Min:Sec) | 2:52 (σ=1:33) |
| Fun Factor (in a scale of 1-5) | 4.65 |

*Table 2: Monolingual, face-to-face (Exp. 1) results*

Overall, most (96%) of the rounds completed successfully, and with an accuracy of 2.65 out of 3 – confirming our results during Pilot phase – though the games took longer to complete than in the Pilot. An examination of a sampling of drawings and outcomes revealed that the users were consistent in their judgments of guesses, but in many cases more demanding than necessary in accepting a guess; that is, in most cases, only perfect matches were accepted, though we found many cases in which equivalents were produced by the *G*, but were not accepted by *D*. Compared with words, as expected, the sentences were conveyed less accurately and took more time; however, phrases were completed faster (in 2:31 for phrases *vs.* 2:43 for words), and more accurately (2.91 for phrases, *vs.* 2.73 for words). Upon closer examination, we found that the drawer first conveyed the easier word, letting the whole phrase be guessed given that context. Most importantly, the fun factor rated by the players averaged at 4.65, even after one hour of continuous play.

We found that the dynamics between the players played a very important role in their productivity; some teams were nearly twice as productive as others, and productive teams specifically mentioned the partner dynamics as a factor for productivity. Furthermore, we found the productivity and the fun factor seemed correlated across teams.

In summary, our scaled-up experiments showed qualitatively and quantitatively that the Doodling game has good potential to generate high quality data in volume, and is perceived to be fun by most players. Encouraged by the positive feedback from users, we continued our experiments to find the effects of modifying the game to involve 1) an online setting and 2) cross-lingual gameplay.

## Experiment 2: Monolingual, Online

In the second round of our experiments, we focused primarily on a simple online version of the game in order to explore whether the game is as effective without face-to-face interaction, which we had observed to be a major social factor in the first experiment.

### Experimental Setup

In order to isolate the key focal point of this round, as well as provide continuity and comparability, we designed an experiment that followed the format of the previous one, though this time it was played over the network on tablet PCs and players had no visual contact with each other. We used the shared whiteboard feature of Microsoft Lync as the drawing surface, which includes a text input tool and multiple (per-user) pointers. For the experiment we instructed *D* to use the pen, and *G* to input text using the text input tool; both players used different colored pointer tools to focus the other's attention onto a particular sketch or word on the whiteboard. The metrics recorded by the players were exactly the same as before: hardness (before the round, by *D*, and after the round, by *G*), accuracy, time taken, and Fun Factor.

For this round we had 14 volunteers, interestingly with 12 of them from the previous set of volunteers. In addition, 10 of the returning players chose to pair up with the same partner as before, reinforcing our observation about the importance of partner dynamics.

| | |
|---|---|
| Players | 14 |
| Rounds | 63 |
| Success Ratio (completed/played) | 78% |
| Accuracy (in a scale 1-3) | 2.73 (σ=0.32) |
| Average Time (in Min:Sec) | 4:29 (σ=3:21) |
| Fun Factor (in a scale of 1-5) | 4.20 |

*Table 3: Monolingual, online (Exp. 2) results*

### Analysis & Observations

In total, 63 games were played in the online experiments; the results are summarized in Table 4. The results showed several interesting trends. First, the success ratio dropped to 78% (from 96% in Round 1), but the mean accuracy

improved to 2.73 (from 2.65 in Round 1), though this was not significant ($p$=0.23, $df$=135, $t$=1.2 in an independent samples, two-tailed, unequal sample size and unequal variances t-test. Note all future p-values reported in this paper are for this type of test unless otherwise specified; also note that the degrees of freedom measure is not $n-1$ for this case, but a more complex relation of the variances and counts of the two samples (Ross 2009)). It should be noted here that the paraphrase data collected were judged (by hand verification, post-experiment) to be of very good quality. The average time taken for completion increased by nearly 50%, to 4m:29s, a statistically significant increase ($p$=0.002, $df$=59, $t$=3.20). However, on closer inspection we observed that one specific hard sentence ("*I was told that you could buy Persian carpets here*.") was completed only by 3 of the 7 teams, averaging more than 15 minutes to complete, and thus inflating the overall average substantially. Excluding just that text element brought the average time significantly, to 3m:46s (and improving the accuracy marginally, and reducing the completion ratio marginally).

The fun factor was 4.2, still a relatively high value, but a drop from the 4.65 we observed in face-to-face experiments. A contributing factor to this drop, as reported by the users, was the latency of the Lync virtual whiteboard. Small delays therein resulted in many false starts and stops, reducing the perceived fun factor.

The qualitative feedback also indicated that the players missed the general social interactivity of the first round, suggesting a need for the introduction of some social elements (voice/video) to enhance the social experience. We hope in further iterations of the UI design to better retain the effectiveness and fun factor of the game in an online environment; this is an avenue we are currently exploring.

## Experiment 3: Going Multilingual

We also conducted a series of experiments to explore the extension of the gaming paradigm to a multilingual gameplay scenario, in which the players are multilingual and the guesses are in a language other than the original provided to the drawer. However, there are several flavors of multilingualism (linguistic families of the two languages), the language-commonality between two players (how multilingual they are, and the language in common). Given our practical constraints on data, potential volunteers, etc., we narrowed down our experiments to a key subset of possible multilingual extensions:

- All users were multilingual with one common language – English – in which the guesses were made. This meant that an independent verification mechanism could be employed for all experiments.
- The non-English language could be from the same linguistic family as English (e.g., French, from the Indo-European family) or a different one (e.g., Tamil – from Dravidian family).

### Experimental Setup

The sessions in multilingual rounds followed much the same format as the first round of experiments. For consistency, we constrained the guessing language for all sessions to be English. We augmented the previous input set by about 30% with hand-crafted tourism phrases that represented concepts or constructs that were culture-centric or idiomatic to the source language. All rounds were played face-to-face and hence the results of Round 1 are taken as the baseline for these Round 3 experiments.

For the English-French cross-lingual experiments, a new set of 20 volunteers, all bilingual in both English and French, were recruited and paired up randomly. We carefully selected 24 text elements, evenly distributed in Granularity and Hardness, both in English and French, including some French text elements that represented concepts that were very specific to the culture (for example, the French "*Pourboire*", which literally means "*For Drink*" alluding to the root for this word, but translates as "*Tip*" or "*Gratuity*" in English), where a simple translation would be considered inelegant.

Similarly for English-Indic experiments, we had a set of 8 volunteers, all bilingual in English and an Indic language, Hindi or Tamil. As in the French case, culture and language specific text elements were added to the game (for example, the Tamil ஜல்லிக்கட்டு – pronounced *jallikattu*), which refers to a Tamil cultural game close to the Spanish bull-running). Three pairs of players played the Hindi–English version of the game, and one pair played the Tamil–English counterpart. On an average, the teams played for an hour, completing as many game rounds as possible.

### Results & Analysis

The experimental results of games in Round 3 played between English and French as well English and Indic languages are given in Table 4.

| | En-French | En-Indic |
|---|---|---|
| Players | 20 | 8 |
| Rounds | 233 | 50 |
| Success Ratio | 97% | 98% |
| Accuracy (Scale: 1-3) | 2.81 (σ=0.24) | 2.79 (σ=0.32) |
| Average Time (in Min:Sec) | 2:54 (σ=1:31) | 3:10 (σ=2:19) |
| Fun Factor (Scale: 1-5) | 4.12 | 4.24 |

*Table 4: Multilingual, face-to-face (Exp. 3) results*

The first observation is that the game results – in terms of Success Ratio and Average Time – are almost identical with the monolingual experiments of Round 1 (compare Tables 2 and 4). There is also a significant improvement on accuracy over monolingual experiments ($p$=0.004, $df$=112, $t$=2.92). We attribute this phenomenon to the more tolerant evaluation of a guess by the players in cross-lingual settings. In the monolingual experiments, we observed that some players played longer for an exact match rejecting acceptable variations (say, for "*toilet*"), whereas in French-English they accepted a variation (say, "*washroom*" for "*toilettes*"). A sampling of source phrases and the corresponding accepted guesses are shown in Table 5.

We also found that the Fun factor dropped to 4.12 (En-French) and 4.24 (En-Indic), as compared to 4.65 for monolingual experiments. We believe that this drop may be due to the long session time, or cultural idioms that were hard to guess. When considering only the subset of culture-specific idioms, the average time for completion increased to 4:16 (a marginally significant difference with $p$=0.12, $df$=13, $t$=1.69), without a statistically significant change in accuracy ($p$=0.74, $df$=19, $t$=0.33) and a drop in completion rate to 84%. Though we cannot draw strong conclusions here, it seems that it may be harder to convey culture-specific concepts across languages.

## Analysis & Observations

### Monolingual vs. Multilingual Game Dynamics

The most important finding from our multilingual experiments (by comparing the results in Tables 2 and 4) is that the game results are almost identical in terms of Success Ratio and Average Time, irrespective of how closely related or diverse the languages are.

| | |
|---|---|
| ஜல்லிக்கட்டு | Bull-tying/Bull-fighting Festival |
| நான் செருப்போடு வீட்டிற்க்குள்ளே வரவா? | Can I come into the house with my shoes on? |
| இந்த சட்டையை தேய்க்கவேண்டும் | I need to iron my shirt. |
| J'ai besoin de repasser ma chemise. | |
| मुझे ये कमीज़ इस्तरी करनी है। | |
| Haute Couture | Designer Clothes |
| Mes Valises sont perdues. | I have lost my luggage. |
| | I lost my baggage. |
| मेरा सामान खो गया है। | I lost my belongings. |
| | My luggage is lost. |
| मेरी संध्या का समय हो गया। | It's time for my evening prayers. |

*Table 5: Sample multilingual data gathered*

### Gathered Data & Quality

A hand-verification of the results from the experiments indicated that the accuracy metric captures the quality of the data both in monolingual and multilingual scenarios with high fidelity; in monolingual scenarios several players tended to be more conservative in evaluation. The gathered data appeared to be of sufficient quality for training modern machine translation systems.

### Player Dynamics

Our post interviews also revealed that the player dynamics played a very significant role in the productivity as well as the fun-factor of the game. In several cases, we found as in the monolingual case that the player pairs who were matched evenly on skill, interest, or interaction style were the most productive and perceived the game to be the most fun. An important component for the final game design will thus be to match the players evenly based on their backgrounds and abilities. This is another area that we plan to pursue in our future research.

Another interesting outcome from these interviews was the discovery that the game could potentially serve as a medium for online collaborative language and cultural learning. Several participants revealed that they spent significant time discussing a language and various culture-specific words. Though such words were difficult to convey in the sketch-and-convey paradigm, and often ended in failures, the game provided an exciting way of discovering ideas and concepts from each other. This aspect could be further developed for an online version of the game, to enable greater cross-cultural interactions and the potential for some language learning occurring simultaneously with the entertainment of the gameplay.

## Summary of Experiments

In Figures 3 and 4 below, we show a summary of the three experimental conditions – monolingual/face-to-face, monolingual/online, and multilingual/face-to-face – in terms of accuracy, completion times, and fun factor.

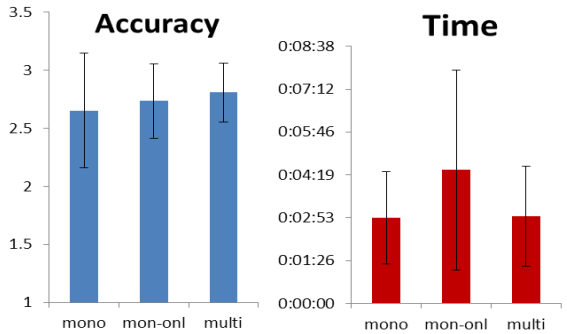


*Figure 3: Accuracy and completion times for all experiments: monolingual (face-to-face), monolingual (online), and all multilingual (face-to-face)*

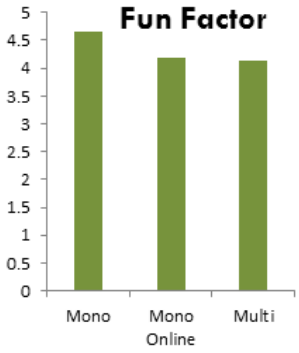


*Figure 4: Average Fun Factor for all experiments: monolingual (face-to-face), monolingual (online), and multilingual (face-to-face)*

Overall, other than the substantially longer completion times for the online case, we note that the statistics of gameplay were quite similar despite the substantial changes in conditions.

## Conclusion

In this paper we investigated the potential of the sketch-and-convey paradigm to generate language data. We designed our game, Doodling, using the Pictionary® metaphor, to help us explore this question. Through a set of user experiments we explored this paradigm with pen/paper and online prototypes; the results show the substantial promise of such a game in gathering natural language data.

Our primary conclusions are as follows:

- The sketch-and-convey paradigm may be used successfully for conveying complex language elements – not just words, but phrases and sentences as well.
- The game rounds typically completed in ~3 minutes, and were perceived as fun, even after players had been at it for a long time – typically an hour or more.
- The online version of the game can be equally productive and almost as fun; the drop in fun factor highlights the need for careful UI design in this case.
- The game can produce highly accurate data in different language settings – paraphrases (in monolingual settings) and parallel data (in multilingual settings).

Our clear next step is to develop an online version of the full game with both monolingual and multilingual options; we are pursuing this direction in the hopes of making it broadly available to the online population.

## References


Ambati, V. and Vogel, S. (2010). Can crowds build parallel corpora for machine translation systems? In *Proc. NAACL HLT 2010 - Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Bloodgood, M. and Callison-Burch, C. (2010). Using Mechanical Turk to build machine translation evaluation sets. In *Proc. NAACL HLT 2010 - Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Callison-Burch, C. (2009). Fast, cheap, and creative: evaluating translation quality using amazon's Mechanical Turk. In *Proc EMNLP'09*.

Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Chen, D.L. and Dolan, W.B. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proc. of the 49th ACL*.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fey, A., Baker, D., Popovic, Z. and Foldit Players. (2010). Predicting protein structures with a multiplayer online game. *Nature* (466), Aug 2010.

Deng, J., Dong,W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proc CVPR'09*.

DuoLingo (2011). *http://duolingo.com*. launched November 2011.

Foldit (2008). *http://fold.it*. launched May 2008.

Grady, C. and Lease, M. (2010). Crowdsourcing document relevance assessment with Mechanical Turk. In *NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Ho, C.J., Chang, T.H., Lee, J.C., Hsu, J.Y.j. and Chen, K.T. (2009). Kisskissban: a competitive human computation game for image annotation. In *Proc. ACM SIGKDD Workshop on Human Computation*.

Hsueh, P.Y., Melville, P. and Sindhwani, V. (2009). Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 - Workshop on Active Learning for Natural Language Processing*.

Irvine, A. and Klementiev, A. (2010). Using Mechanical Turk to annotate lexicons for less commonly used languages. In *Proc NAACL-HLT 2010 Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk*.

Kumaran, A., Saravanan, K., Datha, N., Ashok, B. and Dendi, V. (2009). WikiBABEL: a wiki-style platform for creation of parallel data. In *Proc. ACL-IJCNLP 2009 Software Demonstrations*.

Law, E.L.M., von Ahn, L., Dannenberg, R.B. and Crawford, M. (2007). Tagatune: A game for music and sound annotation. In *Proc ISMIR'07*.

McGraw, I., Lee, C.Y., Hetherington, L. and Glass, J. (2010). Collecting voices from the crowd. In *Proceedings of LREC'10*.

Ross, S. (2009). *Introduction to Probability and Statistics for Engineers and Scientists, Fourth Edition*. Academic Press.

Von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *Proc CHI'04*.

Von Ahn, L. and Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, Vol 51.

Von Ahn, L., Kedia, M. and Blum, M. (2006). Verbosity: a game for collecting common-sense facts. In *Proc CHI'06*.

Wikibhasha (2010). *http://wikibhasha.org*. launched October 2010.

Wikipedia. *http://Wikipedia.org*. launched January 2001.

Zaidan, O.F. and Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from nonprofessionals. In *Proc. 49th ACL*.