# *1001 Paraphrases*: Incenting Responsible Contributions in Collecting Paraphrases from Volunteers

**Timothy Chklovski**

University of Southern California
Information Sciences Institute (USC/ISI)
Marina del Rey, CA, USA
timc@isi.edu

**Abstract**

A variety of applications can benefit from broad and detailed repositories of linguistic and world knowledge. An emerging approach to acquiring such repositories is to collect them from volunteer contributors. To increase the volume of contributions, some deployed systems for collecting volunteer-contributed knowledge offer recognition or prizes to those who provide the highest volume of contributions. However, rewarding for volume alone can encourage irresponsible contributions by unscrupulous participants. In this paper, we present an *approach to collection from volunteers which incents responsible contributions.* Rather than asking contributors to simply enter knowledge, our approach is to collect additional answers by *asking contributors to guess partially obfuscated answers*. To test the approach, we have implemented an online game, *1001 Paraphrases* (http://ai-games.org/paraphrase.html), and deployed it to collect 20,944 entries paraphrasing 400 statements. We present preliminary observations and lessons learned on the success of the approach.

## Introduction

A variety of applications can benefit from repositories of linguistic and world knowledge. An emerging approach to acquiring such knowledge is to collect it from volunteer contributors, allowing anyone to contribute. The systems deployed to date collect various types of knowledge: acceptable paraphrases of a statement in *1001 Paraphrases*, described in detail here, annotations of images in the *ESP game* (von Ahn and Dabbish, 2004), senses of words in given contexts in *Open Mind Word Expert (OMWE)*, (Mihalcea and Chklovski, 2004; Chklovski and Mihalcea, 2002) and common knowledge about everyday objects including in *Open Mind Common Sense* (*OMCS*), (Singh et al. 2002), *Open Mind Indoor Common Sense* (*OMICS*), (Gupta and Kochenderfer, 2004), LEARNER (Chklovski 2003a, 2003b), LEARNER2 (Chklovski, 2005), and potentially the *Fact Entry Tool* (*FET*) by the CYC team (Belasco et al, 2002), which is currently used internally but may be deployed more widely.



Collecting from masses of volunteers can potentially provide very large amounts of corroborated, multi-perspective knowledge. Fully delivering on the potential of collection from volunteers depends on creating systems which can attract a large volume of contributions. To increase the contribution volume, some deployed systems have offered recognition or prizes to those who provide the most contributions. However, rewarding for volume alone can encourage irresponsible contributions by unscrupulous participants. Such participants may opt to provide lots of arbitrary and incorrect, but quick and easy to enter input. Furthermore, automatically rapidly and accurately assessing the quality of the previously unseen input is inherently difficult. The fundamental reason is that the collection systems are relying on contributors to expand the available knowledge; when a novel contribution comes in, it is very difficult to know that it is incorrect without potentially costly additional verification, even if it appears unusual in light of the knowledge already present.

Given the desire to incent high volume of input, and the difficulty of rapidly and inexpensively assessing its quality, it would be beneficial to align the interests of the contributors and the aims of the collection activity. In this paper, we present an *approach to collection from volunteers which incents responsible contributions*.

Rather than asking contributors to simply enter knowledge, our approach is to *ask contributors to guess partially obfuscated answers*. In the process of guessing, contributors provide additional plausible answers, extending the collected knowledge. To test the approach, we have implemented it as an online game, *1001 Paraphrases*, and deployed it to collect paraphrases for a set of 400 sentences needed in a particular target application (Narayanan et al., 2003). We present preliminary observations and lessons learned on the success of the approach.

Specifically, in the rest of the paper, we briefly motivate collecting paraphrases from volunteers, present an approach to incenting responsible contributions of paraphrases, present *1001 Paraphrases* which implements the approach, and discuss some preliminary observations

and lessons learned from using *1001 Paraphrases* to collect 20,944 statements from approximately 1,300 visitors to the site.

## Approach

An important motivation of need for incenting responsible contributions is evidence of irresponsible contributions when such an incentive is not present. In deployed systems for collecting knowledge from volunteers which reward contributors solely based on the volume of data contributed, there have been instances of contributors "gaming" the system. In *Open Mind Word Expert* (*OMWE*), a collection site on which contributors annotate instances of words with their senses (Mihalcea and Chklovski, 2004), a concerned contributor has contacted the organizers asking that statements contributed from his account be deleted. The contributor indicated that his little brother used his account to enter a lot of low-quality data to be rewarded for the high volume of contribution, disregarding the notice that entries by winning contributors will be spot-checked for quality. Another experimental web-browser-based knowledge collection activity presented pages of multiple choice questions. Some contributors quickly discovered that they could rapidly increase their scores by using the browser's back button and re-submitting many times the same form filled out once, further garnering high "agreement" with themselves (Stork, 2003).

Furthermore, the more engaging and game-like the knowledge collection interfaces become, the greater may be the temptation for contributors to cut corners to get ahead.

At the same time, quickly and reliably *identifying invalid contributions is made challenging by the very nature of collecting new knowledge*, because the collection system is constantly receiving statements which it did not already have. One approach to judging quality of statements is to wait until other contributors also enter it. While a statement having been entered several times may be indicative of its reliability, a statement which has been entered only once may still be acceptable and useful, just rarely contributed. Another approach is to use "validating contributors" who are presented with previously entered statements and express whether they agree or disagree with them. Yet, if a validating contributor disagrees with a statement, further investigation is needed to establish whether the original or validating contributor is to be trusted. Despite these issues, assessing quality of statements contributed by volunteers is important and should be used in conjunction with the approach we investigate here.

Rather than struggle with the contributors, we note that it would be helpful to structure the collection activity so that contributors try to provide responsible contributions in the first place. To that end, we propose a simple approach to incentivising high quality contributions. The approach is to collect additional answers by *asking contributors to guess partially obfuscated seed or previously known answers*. The approach is applicable when a question has a number of valid answers, as in the case of paraphrases for a given statement or question. For each "question" or a prompting item (in our case, the expression to be paraphrased), the approach requires at least one valid answer to seed the collection. Such seed answers may be solicited from contributors separately or perhaps be the highest precision answers automatically extracted from text corpora. Our approach then allows collection of additional answers.

To allow collection of novel answers completely unrelated to the already collected answers, in our approach as we have deployed it, each user has to make an initial guess with the answers completely obfuscated. Since the contributor is playing the "game" of guessing the target answers, the contributor still has an incentive to make plausible guesses even when the answers are completely obfuscated.

By incorporating elements of guessing and immediate feedback on success and failure, the approach may also be more engaging than simply entering knowledge, although we have not evaluated such a claim.

The approach can also be used to collect entries other than paraphrases. For instance, it can be used to collect answers to questions about everyday objects. To collect such entries, a question such as "`computers are used to _____`" would be presented, collecting as answers such phrases as "`compose email`," "`send email`," "`compose documents`," "`view images`" and so on. Previously established answers can be used as targets to be guessed. For situations in which answers are easier to guess, the awarded score may be tied to the number of target answers guessed correctly.

## Description of *1001 Paraphrases*

Natural language permits us to say nearly the same thing in a great many ways. This variability of the surface form without significant impact on the meaning can present difficulties for speech recognition systems identifying what is being said, even if only small set of commands or queries is expected. Such variability also complicates the task of machine translation. To address such variability for a given expression, it can be useful to collect different ways to paraphrase it.

Although *1001 Paraphrases* is a general platform for collecting paraphrases, it has been deployed for collecting paraphrases for a specific research project. We introduce this specific project, and then describe the interface, interaction, and scoring in *1001 Paraphrases.*

### Specific Application of Paraphrase Collection

*1001 Paraphrases* has been deployed to collect training data for a machine translation system which needs to recognize paraphrase variants of specific target expressions (Narayanan et al., 2003). The initial objective

of the translation system is to allow an English-speaking doctor in a foreign country to communicate a limited number of statements and questions to a non-English (Persian) speaking patient. The supported communication is limited to four hundred recognized statements, such as "do you have a fever?" and "how long have you had these symptoms?" The goal is to allow the doctor to say, in English, any paraphrase of any recognized statement into a hand-held device. The paraphrase then is to be matched to the correct statement of the four hundred, and the pre-stored translation of that statement is to be output in the output language. Due to the potential variability of the statements, a challenging stage is to map the statement made to the closest matching recognized statement. To assist with this step, it is helpful to have a large corpus of paraphrases of the allowed statements. *1001 Paraphrases* has been deployed to collect a large number of possible paraphrases for the four hundred recognized statements. Although the site has not been actively promoted, it has been visited by approximately 1,300 contributors (not all of whom played the game), and collected 20,944 distinct paraphrases over 15 months. The recognized statements initially came with one or two paraphrases each, a total of 400 statements and 638 additional paraphrases. These statements and their paraphrases were used as the seeds.

## Interface of *1001 Paraphrases*[1]

The "game" consists of a contributor making multiple attempts at guessing any of the several partially obscured paraphrases for a displayed expression. Figure 1 shows a screen shot of the *1001 Paraphrases* interface. Displayed at the top is the expression to be paraphrased, in this case "*this can help you*". The partially obscured target expressions to guess are shown in the hints box. The "..." in the hints indicates that one or more words have been obscured.

The contributor enters a paraphrase in the box titled "Another way to say it". If the contributor correctly guesses one of the paraphrases, he is awarded the amount of points specified next to "you can win," and the game proceeds to the next item for paraphrasing. Otherwise, the entered expression is added to the list of expressions already tried by this contributor, and a larger fraction of the words for the paraphrases to be guessed are revealed in the hints box. The number of points you can win is also decreased. The contributor may also choose to ask for a hint. Just as an unsuccessful guess, this reveals more words in the target expression, but decreases the number of points you can win. The hint functionality has been added to allow contributors to get more information when they cannot think of a guess. At the same time, number of points you can win is decreased to encourage guessing earlier, when little information available, increasing diversity of collected paraphrases and reflecting the difficulty of guessing. The contributor can also give up; the full answers are then revealed and a new item to paraphrase is chosen at random from the target set. No points are awarded or subtracted.


**Figure 1. A Screenshot of the Paraphrase Game**

When a new item to paraphrase is presented, no hints words are provided at all for the first guess. This is done to allow collection of paraphrases which share no words with the known paraphrases.

## Selection and Obfuscation of Hints

As mentioned earlier, our approach relies on presenting obfuscated answers. In the deployed version, how much and exactly how the answers are obfuscated is important, both for whether novel answers will be entered and for the user experience. In designing the obfuscation, we aimed to provide the setting to enter new spontaneous contributions and to prompt contributors' thinking rather than to constrain it. While the hints may potentially bias new contributions towards the ones already collected, as the example in Figure 1 illustrates, the partial hints such as "this could …" and "… help" for paraphrases of "*this can help you*" still allow much variation in the input.

The mechanism we have selected is to show a certain percentage of the words in the expression, replacing the runs of omitted words with "…". For the first guess, we do not reveal any text of the hint, calling for a complete guess from the contributor. For the second guess, 66% of all words (regardless of part of speech etc) are obfuscated. For third and subsequent guesses, 33% of all words are obfuscated. We do not indicate the actual number of words in the expression by design. If we did, contributors may not enter some paraphrases because they would not "fit," and this self-censorship would run counter to the goal of the collection.

[1] *1001 Paraphrases* is available online at http://ai-games.org/paraphrase.html

### Support for Entering Several Variants of a Paraphrase at Once

In designing *1001 Paraphrases*, we anticipated that it may be frustrating to contributors to think of many variants of paraphrases but to be able to enter only one as their guess. Restricting the number of paraphrases a contributor can enter is also counterproductive to the goal of the collection. To address this issue, *1001 Paraphrases* supports compact entry of many variants of a paraphrase at once. All the entered variants count as one guess; the contributor wins the round if any variant matches. The variants are entered by using special symbols, such as "/" (for alternation), "_" (for keeping entities together in alternations), and "(", and ")" for enclosing optional parts in the paraphrases. For example, entering "hello/hi there" expands to two phrases: "hello there" and "hi there." In designing the language for entering multiple guesses at once, we aimed to balance simplicity and expressiveness. Here is a more elaborate example exercising all of the allowed constructs:

```
how are_you_doing/is_it_going (today/this_morning)?
```

This expands to the following 6 questions:

```
how are you doing?
how are you doing today?
how are you doing this morning?
how is it going?
how is it going today?
how is it going this morning?
```

The contributed multi-paraphrase expressions are not used as hints. Only instances of expansions are used as hints.

In practice, 3.8% of the contributed paraphrases used this functionality. We speculate that many contributors may have been unaware of this functionality, because one had to visit the "help" page to learn about it, and server logs indicate that few visitors did. Also, perhaps more contributors would use this functionality to gain an edge in their scores if the activity was made more engaging or competitive (for example, by introducing into the game competition against time and maintaining a high-scores list).

## Preliminary Observations on Collected Paraphrases

Examining the collected paraphrases suggests that contributors indeed tended to make responsible contributions. Of the contributed 20,944 paraphrases, only 0.23% contained swear words, and 0.12% were nonsense entries. By contrast, a much larger fraction of the statements, approximately 5%, contained misspellings.

Although the seed set contained 1,038 statements (400 distinct recognized expressions and 638 paraphrases for them), Contributors provided many more by playing the game. Of the 20,944 contributions, some were identical, because different contributors made the same guesses. In all, 14,850 distinct entries were collected. By way of illustration, Table 1 shows all the statements entered for the statement "this will help you." The seed paraphrases were "this'll help" and "this will be of help." One contribution was irrelevant. It contained swear words and was accordingly discarded. Of 22 remaining new contributions, the majority were good although some, such as "try this" or "it's healthy" were further off in meaning, and one, "nice going" seems not relevant.

For longer statements, such as "we will begin the operation as soon as we can," majority of paraphrases were still accurate, but some paraphrases omitted some of the information. For example, paraphrases such as "we will be ready soon" or "will start as soon as possible." Although such omissions may be undesirable in other circumstances, they were mostly acceptable for our target application.

| Paraphrase | # times contri-buted | Paraphrase | # times contri-buted |
|---|---|---|---|
| try this | 6 | its healthy | 2 |
| this should do the trick | 5 | this could be of help | 1 |
| **this will help you** | 3 | this will make it better | 1 |
| this will help | 3 | this could be better | 1 |
| this should help | 3 | this makes you feel better | 1 |
| this will do the trick | 3 | this should do | 1 |
| **this'll help** | 2 | this will fix the problem | 1 |
| **this will be of help** | 2 | this should fix the problem | 1 |
| this should make it better | 2 | good for you | 1 |
| this should be better | 2 | this may help | 1 |
| this can help you | 2 | that should do the trick | 1 |
| this should help you | 2 | nice going | 1 |
| this is better | 2 | [irrelevant contribution] | 1 |

**Table 1. All 26 distinct statements collected for "this will help you."**

## Related Work

Instead of collecting paraphrases from contributors, paraphrases can also be extracted from texts by identifying and aligning multiple versions of the same text (e.g. Dolan et al., 2004) and translations of the same text (e.g., Barzilay and Lee, 2003). While text extraction can potentially extract a large volume of paraphrases, turning to contributors allows focusing deeply on several selected expressions, collective very many variants and ensuring that many volunteers are able to generate them. Such focus and selectable degree of validation are useful in

applications such as the limited-phrasebook machine translation for which *1001 Paraphrases* has been deployed. At the same time, the approaches may complement each other, with the text extraction providing several high-confidence seed items, and/or with paraphrases collected from volunteers being used to bootstrap extraction from text corpora.

In the *ESP game* (von Ahn & Dabbish, 2004), randomly paired contributors simultaneously enter annotations of a selected images from the Web, with a label being acquired when the two contributors enter the same label. Since the goal of the game is for two contributors to guess the same label the *ESP game* also encourages responsible contributions. In the *ESP game*, the contributors have to agree without additional hints. The *ESP game* has been designed to generate mostly single-word labels for images. Applying the approach to sentence-long paraphrases without additional hints may be difficult on sentence-long paraphrases, due to variety of paraphrases. However, it may be interesting to explore an approach which is a hybrid of the one we have described and of the approach taken in ESP game, with several contributors compete on being the first to correctly guess the paraphrase of a given item, or with contributors guessing each others' paraphrases by being given hints about them.

## Conclusions

We have presented an approach to incent potentially unscrupulous volunteers to make responsible contributions. Rather than reward for pure volume, the approach rewards for correctly guessing obfuscated previously known answers. Preliminary analysis of deploying the approach to collect paraphrases indicates that a large number of novel, useful paraphrases were indeed collected with the approach, and that there was little evidence of non-compliant contributions.

## References


Luis von Ahn and Laura Dabbish. 2004. Labeling Images with a Computer Game. In *ACM CHI* 2004

Barzilay R. and Lee, L. 2003. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. In *Proc. of NAACL-HLT*, 2003.

Belasco, A., Curtis, J., Kahlert, R., Klein, C., Mayans, C., Reagan, P. 2002. Representing Knowledge Gaps Effectively. In *Practical Aspects of Knowledge Management*, (*PAKM*), Vienna, Austria, December 2-3.

Chklovski, T. and Mihalcea, R. 2002. Building a Sense Tagged Corpus with Open Mind Word Expert. In *Proceedings of the Workshop on "Word Sense Disambiguation: Recent Successes and Future Directions"*, ACL 2002. pp. 116-122

Chklovski, T. 2003a. *Using Analogy to Acquire Commonsense Knowledge from Human Contributors*, PhD thesis. MIT Artificial Intelligence Laboratory technical report AITR-2003-002

Chklovski, T. 2003b. LEARNER: A System for Acquiring Commonsense Knowledge by Analogy. In *Proceedings of Second International Conference on Knowledge Capture (K-CAP 2003).*

Chklovski, T. 2005. Designing Interfaces for Guided Collection of Knowledge about Everyday Objects from Volunteers. In *Proceedings of Conference on Intelligent User Interfaces (IUI05)* San Diego, CA

Dolan, W. B., Quirk, C., and Brockett, C. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *Proceedings of COLING 2004*, Geneva, Switzerland.

Gupta, R., and Kochenderfer, M. 2004. Common sense data acquisition for indoor mobile robots. In *Nineteenth National Conference on Artificial Intelligence (AAAI-04).*

Mihalcea, R., and Chklovski, T. 2004. Building Sense Tagged Corpora with Volunteer Contributions over the Web. In *Current Issues in Linguistic Theory: Recent Advances in Natural Language Processing*, Nicolas Nicolov and Ruslan Mitkov (eds), John Benjamins Publishers.

Narayanan, S., Ananthakrishnan, S., Belvin, R., Ettelaie, E. Ganjavi, S. Georgiou, P., Hein, C., Kadambe, S., Knight, K., Marcu, D., Neely, H., Srinivasamurthy, N., Traum, D. and Wang, D. 2003. Transonics: A Speech to Speech System for English-Persian Interactions. In *Proc. IEEE ASRU*.

Stork, D. 2003. Invited talk at the *Workshop on Distributed and Collaborative Knowledge Capture* (DC-KCAP), held in conjunction with the *International conference on Knowledge Capture* (*K-CAP 2003*).