Report on the First NLG Challenge on Generating Instructions in Virtual Environments (GIVE)

Donna Byron
Northeastern University
dbyron@ccs.neu.edu

Alexander Koller
Saarland University
koller@mmci.uni-saarland.de

Kristina Striegnitz
Union College
striegnk@union.edu

Justine Cassell
Northwestern University
justine@northwestern.edu

Robert Dale
Macquarie University
Robert.Dale@mq.edu.au

Johanna Moore
University of Edinburgh
J.Moore@ed.ac.uk

Jon Oberlander
University of Edinburgh
J.Oberlander@ed.ac.uk

Abstract

We describe the first installment of the Challenge on Generating Instructions in Virtual Environments (GIVE), a new shared task for the NLG community. We motivate the design of the challenge, describe how we carried it out, and discuss the results of the system evaluation.

1 Introduction

This paper reports on the methodology and results of the First Challenge on Generating Instructions in Virtual Environments (GIVE-1), which we ran from March 2008 to February 2009. GIVE is a new shared task for the NLG community. It provides an end-to-end evaluation methodology for NLG systems that generate instructions which are meant to help a user solve a treasure-hunt task in a virtual 3D world. The most innovative aspect from an NLG evaluation perspective is that the NLG system and the user are connected over the Internet. This makes it possible to cheaply collect large amounts of evaluation data.

Five NLG systems were evaluated in GIVE-1 over a period of three months from November 2008 to February 2009. During this time, we collected 1143 games that were played by users from 48 countries. As far as we know, this makes GIVE-1 the largest evaluation effort in terms of experimental subjects ever. We have evaluated the five systems both on objective measures (success rate, completion time, etc.) and subjective measures which were collected by asking the users to fill in a questionnaire.

GIVE-1 was intended as a pilot experiment in order to establish the validity of the evaluation methodology and understand the challenges involved in the instruction-giving task. We believe that we have achieved these purposes. At the same time, we provide evaluation results for the five NLG systems which will help their developers improve them for participation in a future challenge, GIVE-2. GIVE-2 will retain the successful aspects of GIVE-1, while refining the task to emphasize aspects that we found to be challenging. We invite the ENLG community to participate in designing GIVE-2.

Plan of the paper. The paper is structured as follows. In Section 2, we will describe and motivate the GIVE Challenge. In Section 3, we will then describe the evaluation method and infrastructure for the challenge. Section 4 reports on the evaluation results. Finally, we conclude and discuss future work in Section 5.

2 The GIVE Challenge

In the GIVE scenario, subjects try to solve a treasure hunt in a virtual 3D world that they have not seen before. The computer has a complete symbolic representation of the virtual world. The challenge for the NLG system is to generate, in real time, natural-language instructions that will guide the users to the successful completion of their task.

Users participating in the GIVE evaluation start the 3D game from our website at `www.give-challenge.org`. They then see a 3D game window as in Fig. 1, which displays instructions and allows them to move around in the world and manipulate objects. The first room is a tutorial room where users learn how to interact with the system; they then enter one of three evaluation worlds, where instructions for solving the treasure hunt are generated by an NLG system. Users can either finish a game successfully, lose it by triggering an alarm, or cancel the game. This result is stored in a database for later analysis, along with a complete log of the game.

Complete maps of the game worlds used in the evaluation are shown in Figs. 3–5: In these worlds, players must pick up a trophy, which is in a wall safe behind a picture. In order to access the tro-

Figure 1: What the user sees when playing with the GIVE Challenge.

phy, they must first push a button to move the picture to the side, and then push another sequence of buttons to open the safe. One floor tile is alarmed, and players lose the game if they step on this tile without deactivating the alarm first. There are also a number of distractor buttons which either do nothing when pressed or set off an alarm. These distractor buttons are intended to make the game harder and, more importantly, to require appropriate reference to objects in the game world. Finally, game worlds contained a number of objects such as chairs and flowers that did not bear on the task, but were available for use as landmarks in spatial descriptions generated by the NLG systems.

2.1 Why a new NLG evaluation paradigm?

The GIVE Challenge addresses a need for a new evaluation paradigm for natural language generation (NLG). NLG systems are notoriously hard to evaluate. On the one hand, simply comparing system outputs to a gold standard using automatic comparison algorithms has limited value because there can be multiple generated outputs that are equally good. Finding metrics that account for this variability and produce results consistent with human judgments and task performance measures is difficult (Belz and Gatt, 2008; Stent et al., 2005; Foster, 2008). Human assessments of system outputs are preferred, but lab-based evaluations that allow human subjects to assess each aspect of the system's functionality are expensive and time-consuming, thereby favoring larger labs with adequate resources to conduct human subjects studies. Human assessment studies are also difficult to replicate across sites, so system developers that are geographically separated find it difficult to compare different approaches to the same problem, which in turn leads to an overall difficulty in measuring progress in the field.

The GIVE-1 evaluation was conducted via a client/server architecture which allows any user with an Internet connection to provide system evaluation data. Internet-based studies have been shown to provide generous amounts of data in other areas of AI (von Ahn and Dabbish, 2004; Orkin and Roy, 2007). Our implementation allows smaller teams to develop a system that will participate in the challenge, without taking on the burden of running the human evaluation experiment, and it provides a direct comparison of all participating systems on the same evaluation data.

2.2 Why study instruction-giving?

Next to the Internet-based data collection method, GIVE also differs from other NLG challenges by its emphasis on generating instructions in a virtual environment and in real time. This focus on instruction giving is motivated by a growing interest in dialogue-based agents for situated tasks such as navigation and 3D animations. Due to its appeal to younger students, the task can also be used as a pedagogical exercise to stimulate interest among secondary-school students in the research challenges found in NLG or Computational Linguistics more broadly.

Embedding the NLG task in a virtual world encourages the participating research teams to consider communication in a *situated* setting. This makes the NLG task quite different than in other NLG challenges. For example, experiments have shown that human instruction givers make the instruction follower move to a different location in order to use a simpler referring expression (RE) (Stoia et al., 2006). That is, RE generation becomes a very different problem than the classical non-situated Dale & Reiter style RE generation, which focuses on generating REs that are single noun phrases in the context of an unchanging world.

On the other hand, because the virtual environments scenario is so open-ended, it – and specifically the instruction-giving task – can potentially be of interest to a wide range of NLG researchers. This is most obvious for research in sentence planning (GRE, aggregation, lexical choice) and realization (the real-time nature of the task imposes high demands on the system's efficiency). But if

extended to two-way dialog, the task can also involve issues of prosody generation (i.e., research on text/concept-to-speech generation), discourse generation, and human-robot interaction. Finally, the game world can be scaled to focus on specific issues in NLG, such as the generation of REs or the generation of navigation instructions.

3 Evaluation Method and Logistics

Now we describe the method we applied to obtain experimental data, and sketch the software infrastructure we developed for this purpose.

3.1 Software architecture

A crucial aspect of the GIVE evaluation methodology is that it physically separates the user and the NLG system and connects them over the Internet. To achieve this, the GIVE software infrastructure consists of three components (shown in Fig. 2):

1. the *client*, which displays the 3D world to users and allows them to interact with it;
2. the *NLG servers*, which generate the natural-language instructions; and
3. the *Matchmaker*, which establishes connections between clients and NLG servers.

These three components run on different machines. The client is downloaded by users from our website and run on their local machine; each NLG server is run on a server at the institution that implemented it; and the Matchmaker runs on a central server we provide. When a user starts the client, it connects to the Matchmaker and is randomly assigned an NLG server and a game world. The client and NLG server then communicate over the course of one game. At the end of the game, the client displays a questionnaire to the user, and the game log and questionnaire data are uploaded to the Matchmaker and stored in a database. Note that this division allows the challenge to be conducted without making any assumptions about the internal structure of an NLG system.

The GIVE software is implemented in Java and available as an open-source Google Code project. For more details about the software, see (Koller et al., 2009).

3.2 Subjects

Participants were recruited using email distribution lists and press releases posted on the internet.

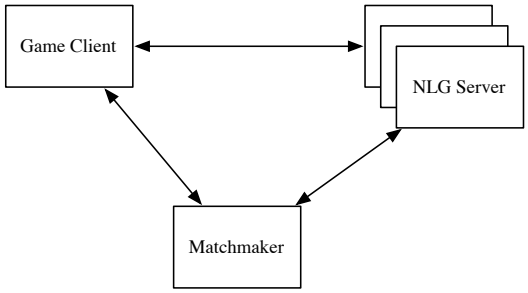

Figure 2: The GIVE architecture.

Collecting data from anonymous users over the Internet presents a variety of issues that a lab-based experiment does not. An Internet-based evaluation skews the demographic of the subject pool toward people who use the Internet, but probably no more so than if recruiting on a college campus. More worrisome is that, without a face-to-face meeting, the researcher has less confidence in the veracity of self-reported demographic data collected from the subject. For the purposes of NLG software, the most important demographic question is the subject's fluency in English. Players of the GIVE 2009 challenge were asked to self-report their command of English, age, and computer experience. English proficiency did interact with task completion, which leads us to conclude that users were honest about their level of English proficiency. See section 4.4 below for a discussion of this interaction. All-in-all, we feel that the advantage gained from the large increase in the size of the subject pool offsets any disadvantage accrued from the lack of accurate demographic information.

3.3 Materials

Figs. 3–5 show the layout of the three evaluation worlds. The worlds were intended to provide varying levels of difficulty for the direction-giving systems and to focus on different aspects of the problem. World 1 is very similar to the development world that the research teams were given to test their system on. World 2 was intended to focus on object descriptions - the world has only one room which is full of objects and buttons, many of which cannot be distinguished by simple descriptions. World 3, on the other hand, puts more emphasis on navigation directions as the world has many interconnected rooms and hallways.

The difference between the worlds clearly bears out in the task completion rates reported below.

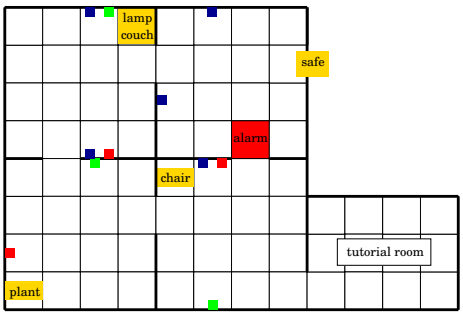

Figure 3: World 1

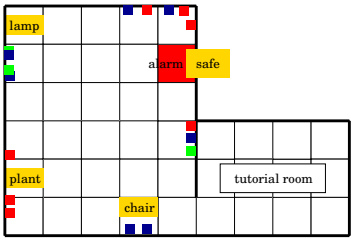

Figure 4: World 2

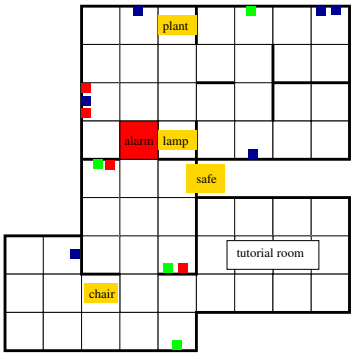

Figure 5: World 3

3.4 Timeline

After the GIVE Challenge was publicized in March 2008, eight research teams signed up for participation. We distributed an initial version of the GIVE software and a development world to these teams. In the end, four teams submitted NLG systems. These were connected to a central Matchmaker instance that ran for about three months, from 7 November 2008 to 5 February 2009. During this time, we advertised participation in the GIVE Challenge to the public in order to obtain experimental subjects.

3.5 NLG systems

Five NLG systems were evaluated in GIVE-1:

1. one system from the University of Texas at Austin ("Austin" in the graphics below);
2. one system from Union College in Schenectady, NY ("Union");
3. one system from the Universidad Complutense de Madrid ("Madrid");
4. two systems from the University of Twente: one serious contribution ("Twente") and one more playful one ("Warm-Cold").

Of these systems, "Austin" can serve as a baseline: It computes a plan consisting of the actions the user should take to achieve the goal, and at each point in the game, it realizes the first step in this plan as a single instruction. The "Warm-Cold" system generates very vague instructions that only tell the user if they are getting closer ("warmer") to their next objective or if they are moving away from it ("colder"). We included this system in the evaluation to verify whether the evaluation methodology would be able to distinguish such an obviously suboptimal instruction-giving strategy from the others.

Detailed descriptions of these systems as well as each team's own analysis of the evaluation results can be found at `http://www.give-challenge.org/research/give-1`.

4 Results

We now report on the results of GIVE-1. We start with some basic demographics; then we discuss objective and subjective evaluation measures.

Notice that some of our evaluation measures are in tension with each other: For instance, a system which gives very low-level instructions ("move forward"; "ok, now move forward"; "ok, now turn left"), such as the "Austin" baseline, will lead the user to completing the task in a minimum number of steps; but it will require more instructions than a system that aggregates these. This is intentional, and emphasizes both the pilot experiment character of GIVE-1 and our desire to make GIVE a friendly comparative challenge rather than a competition with a clear winner.

4.1 Demographics

Over the course of three months, we collected 1143 valid games. A game counted as valid if the game client didn't crash, the game wasn't marked as a test game by the developers, and the player completed the tutorial.

Of these games, 80.1% were played by males and 9.9% by females; a further 10% didn't specify their gender. The players were widely distributed over countries: 37% connected from an IP address in the US, 33% from an IP address in Germany, and 17% from China; Canada, the UK, and Austria also accounted for more than 2% of the partic-

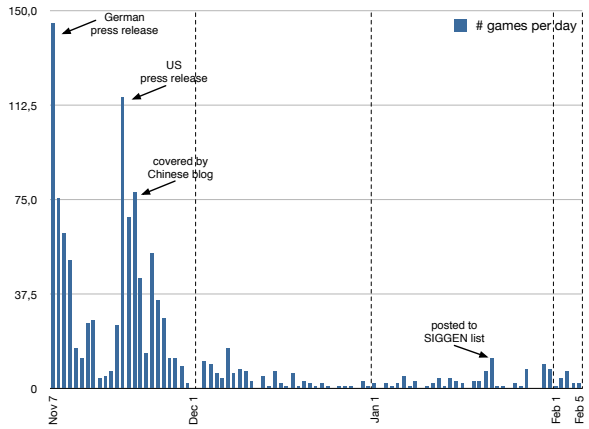

Figure 6: Histogram of the connections per day.

ipants each, and the remaining 2% of participants connected from 42 further countries. This imbalance stems from very successful press releases that were issued in Germany and the US and which were further picked up by blogs, including one in China. Nevertheless, over 90% of the participants who answered this question self-rated their English proficiency as "good" or better. About 75% of users connected with a client running on Windows, with the rest split about evenly among Linux and Mac OS X.

The effect of the press releases is also plainly visible if we look at the distribution of the valid games over the days from November 7 to February 5 (Fig. 6). There are huge peaks at the very beginning of the evaluation period, coinciding with press releases through Saarland University in Germany and Northwestern University in the US, which were picked up by science and technology blogs on the Web. The US peak contains a smaller peak of connections from China, which were sparked by coverage in a Chinese blog.

4.2 Objective measures

We then extracted objective and subjective measurements from the valid games. The objective measures are summarized in Fig. 7. For each system and game world, we measured the percentage of games which the users completed successfully. Furthermore, we counted the numbers of instructions the system sent to the user, measured the time until task completion, and counted the number of low-level steps executed by the user (any key press, to either move or manipulate an object) as well as the number of task-relevant actions (such as pushing a button to open a door).

- task success (Did the player get the trophy?)
- instructions (Number of instructions produced by the NLG system.*)
- steps (Number of all player actions.*)
- actions (Number of object manipulation action.*)
- second (Time in seconds.*)

* Measured from the end of the tutorial until the end of the game.

Figure 7: Objective measurements

	Austin	Madrid	Twente	Union	Warm-Cold
task success	40% B	71% A	35% B	73% A	18% C
instructions	83.2 B	58.3 A	121.2 C	80.3 B	190.0 D
steps	103.6 A	124.3 B	160.9 C	117.5 A B	307.4 D
actions	11.2 B	8.7 A	14.3 C	9.0 A	14.3 C
seconds	129.3 A	174.8 B	207.0 C	175.2 B	312.2 D

Figure 8: *Objective* measures by system. Task success is reported as the percentage of successfully completed games. The other measures are reported as the mean number of instructions/steps/actions/seconds, respectively. Letters group indistinguishable systems; systems that don't share a letter were found to be significantly different with $p < 0.05$.

To ensure comparability, we only counted successfully completed games for all these measures, and only started counting when the user left the tutorial room. Crucially, all objective measures were collected completely unobtrusively, without requiring any action on the user's part.

Fig. 8 shows the results of these objective measures. This figure assigns systems to groups A, B, etc. for each evaluation measure. Systems in group A are better than systems in group B, etc.; if two systems don't share the same letter, the difference between these two systems is significant with $p < 0.05$. Significance was tested using a χ^2-test for task success and ANOVAs for instructions, steps, actions, and seconds. These were followed by post-hoc tests (pairwise χ^2 and Tukey) to compare the NLG systems pairwise.

Overall, there is a top group consisting of the Austin, Madrid, and Union systems: While Madrid and Union outperform Austin on task success (with 70 to 80% of successfully completed games, depending on the world), Austin significantly outperforms all other systems in terms of task completion time. As expected, the Warm-Cold system performs significantly worse than all others in almost all categories. This confirms the ability of the GIVE evaluation method to distinguish between systems of very different qualities.

4.3 Subjective measures

The subjective measures, which were obtained by asking the users to fill in a questionnaire after each game, are shown in Fig. 9. Most of the questions were answered on 5-point Likert scales ("overall" on a 7-point scale); the "informativity" and "timing" questions had nominal answers. For each question, the user could choose not to answer.

The results of the subjective measurements are summarized in Fig. 10, in the same format as above. We ran χ^2-tests for the nominal variables informativity and timing, and ANOVAs for the scale data. Again, we used post-hoc pairwise χ^2- and Tukey-tests to compare the NLG systems to each other one by one.

Here there are fewer significant differences between different groups than for the objective measures: For the "play again" category, there is no significant difference at all. Nevertheless, "Austin" is shown to be particularly good at navigation instructions and timing, whereas "Madrid" outperforms the rest of the field in "informativity". In the overall subjective evaluation, the earlier top group of Austin, Madrid, and Union is confirmed, although the difference between Union and Twente is not significant. However, "Warm-Cold" again performs significantly worse than all other systems in most measures. Furthermore, although most systems perform similarly on "informativity" and "timing" in terms of the number of users who judged them as "just right", there are differences in the tendencies: Twente and Union tend to be overinformative, whereas Austin and Warm-Cold tend to be underinformative; Twente and Union tend to give their instructions too late, whereas Madrid and Warm-Cold tend to give them too early.

7-point scale items:

overall: What is your overall evaluation of the quality of the direction-giving system? (very bad 1 . . . 7 very good)

5-point scale items:

task difficulty: How easy or difficult was the task for you to solve? (very difficult 1 2 3 4 5 very easy)

goal clarity: How easy was it to understand what you were supposed to do? (very difficult 1 2 3 4 5 very easy)

play again: Would you want to play this game again? (no way! 1 2 3 4 5 yes please!)

instruction clarity: How clear were the directions? (totally unclear 1 2 3 4 5 very clear)

instruction helpfulness: How effective were the directions at helping you complete the task? (not effective 1 2 3 4 5 very effective)

choice of words: How easy to understand was the system's choice of wording in its directions to you? (totally unclear 1 2 3 4 5 very clear)

referring expressions: How easy was it to pick out which object in the world the system was referring to? (very hard 1 2 3 4 5 very easy)

navigation instructions: How easy was it to navigate to a particular spot, based on the system's directions? (very hard 1 2 3 4 5 very easy)

friendliness: How would you rate the friendliness of the system? (very unfriendly 1 2 3 4 5 very friendly)

Nominal items:

informativity: Did you feel the amount of information you were given was: too little / just right / too much

timing: Did the directions come ... too early / just at the right time / too late

Figure 9: Questionnaire items

	Austin	Madrid	Twente	Union	Warm-Cold
task difficulty	4.3 A	4.3 A	4.0 A	4.3 A	3.5 B
goal clarity	4.0 A	3.7 A	3.9 A	3.7 A	3.3 B
play again	2.8 A	2.6 A	2.4 A	2.9 A	2.5 A
instruction clarity	4.0 A	3.6 A B	3.8 A B	3.6 B	3.0 C
instruction helpfulness	3.8 A	3.9 A	3.6 A	3.7 A	2.9 B
informativity	46% B	68% A	51% B	56% B	51% B
overall	4.9 A	4.9 A	4.3 B	4.6 A B	3.6 C
choice of words	4.2 A	3.8 B C	4.1 A B	3.7 C	3.5 C
referring expressions	3.4 B	3.9 A	3.7 A B	3.7 A B	3.5 B
navigation instructions	4.6 A	4.0 B	4.0 B	3.7 B	3.2 C
timing	78% A	62% B	60% B C	62% B	49% C
friendliness	3.4 A B	3.8 A	3.1 B	3.6 A	3.1 B

Figure 10: *Subjective* measures by system. Informativity and timing are reported as the percentage of successfully completed games. The other measures are reported as the mean rating received by the players. Letters group indistinguishable systems; systems that don't share a letter were found to be significantly different with $p < 0.05$.

4.4 Further analysis

In addition to the differences between NLG systems, there may be other factors which also influence the outcome of our objective and subjective measures. We tested the following five factors: evaluation world, gender, age, computer expertise, and English proficiency (as reported by the users on the questionnaire). We found that there is a significant difference in task success rate for different evaluation worlds and between users with different levels of English proficiency.

The interaction graphs in Figs. 11 and 12 also suggest that the NLG systems differ in their robustness with respect to these factors. χ^2-tests that compare the success rate of each system in the three evaluation worlds show that while the instructions of Union and Madrid seem to work equally well in all three worlds, the performance of the other three systems differs dramatically between the different worlds. Especially World 2 was challenging for some systems as it required relational object descriptions, such as *the blue button on the left of another blue button*.

The players' English skills also affected the systems in different ways. Union and Twente seem to communicate well with players on all levels of proficiency (χ^2-tests do not find a significant difference). Austin, Madrid and Warm Cold, on the other hand, don't manage to lead players with only basic English skills to success as often as other players. However, if we remove the players with the lowest level of English proficiency, language skills do not have an effect on the task success rate anymore for any of the systems.

5 Conclusion

In this document, we have described the first installment of the GIVE Challenge, our experimental methodology, and the results. Altogether, we collected 1143 valid games for five NLG systems over a period of three months. Given that this was the first time we organized the challenge, that it was meant as a pilot experiment from the beginning, and that the number of games was sufficient to get significant differences between systems on a number of measures, we feel that GIVE-1 was a success. We are in the process of preparing several diagnostic utilities, such as heat maps and a tool that lets the system developer replay an individual game, which will help the participants gain further insight into their NLG systems.

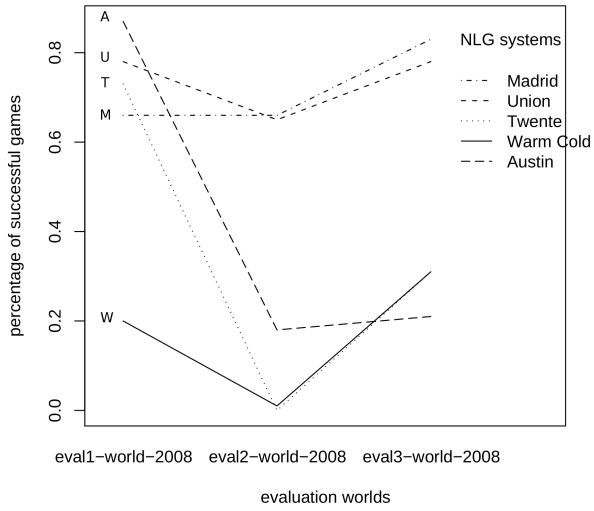

Figure 11: Effect of the evaluation worlds on the success rate of the NLG systems.

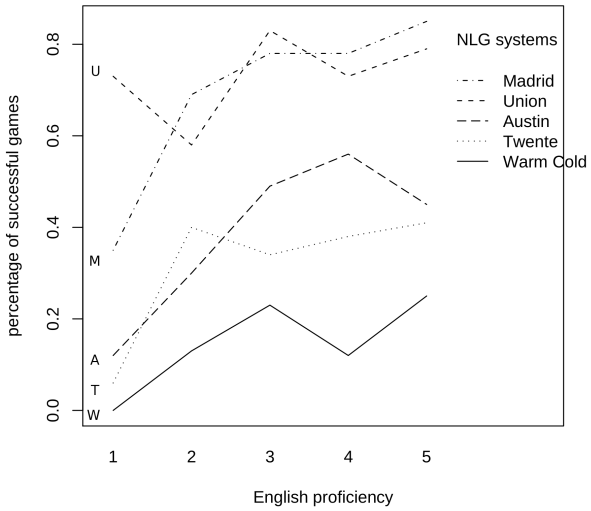

Figure 12: Effect of the players' English skills on the success rate of the NLG systems.

Nevertheless, there are a number of improvements we will make to GIVE for future installments. For one thing, the timing of the challenge was not optimal: A number of colleagues would have been interested in participating, but the call for participation came too late for them to acquire funding or interest students in time for summer projects or MSc theses. Secondly, although the software performed very well in handling thousands of user connections, there were still game-invalidating issues with the 3D graphics and the networking code that were individually rare, but probably cost us several hundred games. These should be fixed for GIVE-2. At the same time, we are investigating ways in which the networking and matchmaking core of GIVE can be factored out into a separate, challenge-independent system on which other Internet-based challenges can be built. Among other things, it would be straightforward to use the GIVE platform to connect two human users and observe their dialogue while solving a problem. Judicious variation of parameters (such as the familiarity of users or the visibility of an instruction giving avatar) would allow the construction of new dialogue corpora along such lines.

Finally, GIVE-1 focused on the generation of navigation instructions and referring expressions, in a relatively simple world, without giving the user a chance to talk back. The high success rate of some systems in this challenge suggests that we need to widen the focus for a future GIVE-2 – by allowing dialogue, by making the world more complex (e.g., allowing continuous rather than discrete movements and turns), by making the communication multi-modal, etc. Such extensions would require only rather limited changes to the GIVE software infrastructure. We plan to come to a decision about such future directions for GIVE soon, and are looking forward to many fruitful discussions about this at ENLG.

Acknowledgments. We are grateful to the participants of the 2007 NSF/SIGGEN Workshop on Shared Tasks and Evaluation in NLG and many other colleagues for fruitful discussions while we were designing the GIVE Challenge, and to the organizers of Generation Challenges 2009 and ENLG 2009 for their support and the opportunity to present the results at ENLG. We also thank the four participating research teams for their contributions and their patience while we were working out bugs in the GIVE software. The creation of the GIVE infrastructure was supported in part by a Small Projects grant from the University of Edinburgh.

References

A. Belz and A. Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation.

In *Proceedings of ACL-08:HLT, Short Papers*, pages 197–200, Columbus, Ohio.

M. E. Foster. 2008. Automated metrics that agree with human judgements on generated output for an embodied conversational agent. In *Proceedings of INLG 2008*, pages 95–103, Salt Fork, OH.

A. Koller, D. Byron, J. Cassell, R. Dale, J. Moore, J. Oberlander, and K. Striegnitz. 2009. The software architecture for the first challenge on generating instructions in virtual environments. In *Proceedings of the EACL-09 Demo Session*.

J. Orkin and D. Roy. 2007. The restaurant game: Learning social behavior and language from thousands of players online. *Journal of Game Development*, 3(1):39–60.

A. Stent, M. Marge, and M. Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *Proceedings of CICLing 2005*.

L. Stoia, D. M. Shockley, D. K. Byron, and E. Fosler-Lussier. 2006. Noun phrase generation for situated dialogs. In *Proceedings of INLG*, Sydney.

L. von Ahn and L. Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the ACM CHI Conference*.