The First Challenge on Generating Instructions in Virtual Environments

Alexander Koller,[1] Kristina Striegnitz,[2] Donna Byron,[3] Justine Cassell,[4]
Robert Dale,[5] Johanna Moore,[6] and Jon Oberlander[6]

[1] Saarland University [2] Union College [3] Northeastern University
[4] Northwestern University [5] Macquarie University [6] University of Edinburgh

1 Introduction

The ability to evaluate and compare our algorithms, techniques and systems is fundamental for the progress of research in computational linguistics. Challenges such as the TREC Question-Answering competition[1] and the NIST machine translation competition[2] have helped spawn tremendous interest in their respective subfields. More recently, the Recognizing Textual Entailment challenge[3] has revived broad interest in computational semantics. By making systems comparable and progress measurable, these challenges have improved systems, contributed to our scientific understanding of the research issues, and helped create research communities in their respective areas.

The field of natural language generation (NLG) lags behind other areas in terms of its ability to evaluate system performance, because NLG systems are inherently hard to evaluate. On the one hand, evaluations based on annotated corpora (e.g. [2]) will misjudge some systems because in NLG a mismatch between the gold standard and a system's output does not necessarily indicate that the system's output is inferior (e.g. [4, 24, 15] and also see [6, 8, 17, 5] in this volume). On the other hand, evaluation studies where human subjects interact with a system or compare the outputs of competing NLG systems are time-consuming and expensive to run, and may be completely infeasible outside of large well-established laboratories due to lack of funding or an insufficient pool of potential subjects.

There has recently been a growing consensus that a convincing evaluation method for NLG is needed [3, 12]. But all recent NLG-specific shared tasks that we know about [17, 5] have been limited to a very specific NLG subtask, that of generating referring expressions. Conversely, the large-scale DARPA Communicator challenge [28] and its (telephone-based) evaluation methodology were tied to evaluating end-to-end spoken dialogue systems and too coarse-grained for a convenient evaluation of just the NLG components. Thus the question of what a generic evaluation method for NLG systems should look like is still largely unanswered.

[1] http://trec.nist.gov/

[2] http://www.nist.gov/speech/tests/mt/

[3] http://pascallin.ecs.soton.ac.uk/Challenges/

In this paper, we report on the results of the First Challenge on Generating Instructions in Virtual Environments (GIVE-1). In this shared task, an NLG system must generate natural-language instructions which guide a user towards performing some task in a virtual 3D environment. This makes it possible to explore situated communication in a simulated environment; but perhaps even more importantly, the GIVE Challenge opens up a novel approach to NLG system evaluation, in which human experimental subjects are connected to NLG systems over the Internet. In GIVE-1, this allowed us to collect data from 1143 separate interactions with NLG systems, making GIVE, to our knowledge, the largest ever NLG evaluation effort in terms of the number of experimental subjects taking part.

The paper is structured as follows. We start by reviewing the state of the art in NLG system evaluation, and present the Internet-based evaluation strategy used by GIVE, in Section 2. We then present the specific setup of GIVE-1 in Section 3, and summarize the five NLG systems that participated in GIVE-1 in Section 4. In Section 5, we present the results of GIVE-1. Finally, we validate the Internet-based evaluation methodology we used in GIVE-1 in Section 6, by comparing the results of the Internet evaluation with those obtained in a more traditional laboratory-based setting. Section 7 provides an outlook towards GIVE-2 and concludes.

2 Evaluating NLG systems

The evaluation of NLG systems has recently been the subject of much discussion. We review some of the main trends, and then introduce a new methodology on which the GIVE Challenge is based: evaluating NLG systems over the Internet.

2.1 Previous work

Evaluation efforts for NLG systems typically fall into one of three classes: similarity with respect to a gold standard, effectiveness in terms of task performance, and ratings by human judges. The first class aims at judging the output of an NLG system by comparing it against one or more gold standards produced by human annotators, using comparison metrics such as those proposed by [2]. The advantage of this approach is that evaluations are cheap and easily repeatable once the gold-standard corpus has been built. This is why the first major shared tasks for NLG, the TUNA [17] and GREC [5] challenges on generating referring expressions, both focused on this approach and only added very small studies involving human subjects.

Unfortunately, evaluation against a gold standard is a much less natural evaluation method for NLG than it is, for example, in parsing. One problem is that the same meaning can usually be expressed equally well in many different ways, and a gold standard cannot capture all possible variations. Furthermore, it has been shown on the TUNA data that those gold-standard-based metrics that are currently available do not correlate with task-based evaluation measures [4, 17].

Similarly, gold-standard-based metrics have been found to not always correlate well with the ratings of human judges [9, 7]. In our opinion, gold-standard evaluations may be useful for tracking improvements in a single NLG system, but are less suitable for large-scale evaluation efforts.

Alternatively, human subjects can be asked to to perform some task that depends on the system output (such as identifying the target referent [4]). Their task performance is a measure of the NLG system's effectiveness for this task. The problem with this approach is that such experiments, which involve getting large numbers of subjects into a laboratory, are expensive and time-consuming. As a consequence, the number of subjects tends to be relatively low (for instance, the TUNA 2009 evaluation used sixteen subjects [16]), which can limit the evaluation's ability to detect significant differences. We know of only one evaluation effort that tested systems involving NLG on a task-based evaluation with human subjects on a large scale: the DARPA Communicator evaluation [28]. However, this was an end-to-end evaluation effort for spoken dialogue systems, which makes it too coarse-grained to tell us much about the specific performance of the NLG components.

The third alternative is to present the system outputs to human judges and ask them to rate their quality. Similarly to task-based evaluations, such evaluations involve significant costs and time requirements. These may still be lower than those of a task-based evaluation since judges are typically not naive, but have some training in linguistics and are aware of the study and its purpose, so that they can work independently and can handle larger amounts of data. However, ratings by judges do not produce a direct measure for how "real" users would interact with an NLG system. Interestingly, the TUNA 2009 evaluation [16] found a correlation between task-based evaluation measures and ratings by judges, but since this is, to the best of our knowledge, the only work so far that compares these two approaches, further research is needed.

2.2 Internet-based evaluation of NLG systems

In this paper, we propose a new approach to evaluating NLG systems which is meant to provide results that are as meaningful as those of a laboratory-based evaluation with human subjects, but at a much lower cost. The key idea is to provide a software infrastructure that allows us to physically separate the experimental subject and the NLG system; the subject runs a client program on their own computer, which connects to the NLG system over the Internet. Access to subjects is obtained by couching the subject's task in a game-like environment, and making it easy to start the client from a public website.

More concretely, the software architecture we implemented for the GIVE Challenge involves three different software components (see Figure 1):

1. the *client*, which implements the task which the users must perform, and displays the NLG system's outputs;
2. the *NLG servers*, which generate the natural-language instructions; and
3. the *Matchmaker*, which establishes connections between clients and NLG servers and enters game logs into a database for future analysis.

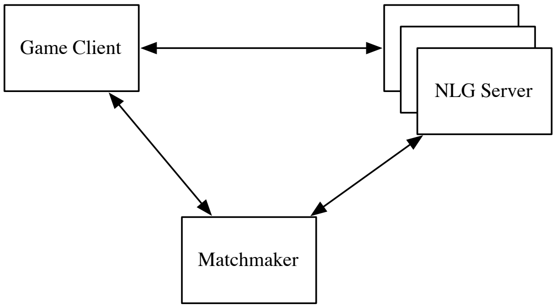


Fig. 1. The architecture of the evaluation software.

These three components run on different machines. The client is downloaded by users from a central 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 provided by the evaluation organizers. When a user starts the client, it connects to the Matchmaker and is randomly assigned an NLG server and a task instance (e.g., one particular 3D world). The client and NLG server then communicate over the course of one task execution. At the end of this run, the client displays a questionnaire to the user, and the log of the user's actions, the NLG system's utterances, and the questionnaire data are uploaded to the Matchmaker and stored in a database.

The Java implementation of our framework is available as an open-source project.[4] While the current implementation is targeted specifically at the GIVE task, it should be relatively straightforward to adapt it to other NLG tasks.

2.3 Access to subjects

One huge advantage of a web-based evaluation is that it provides easy access to the entire population of the Internet as a pool of potential experimental subjects. This resource has been successfully exploited in the past for such tasks as image labelling [1], script learning [22], and corpus creation [10], as well as in psychological and psycholinguistic "web experiments" [19].

From the perspective of the experimental subject, they simply perform a task in which they are guided by natural-language instructions; in principle, they don't even need to know that the instructions come from an automated system. They can be recruited by any means that will get them to visit the website from which they can download the client. As we will see below, even relatively conservative methods of recruiting subjects already provide quite satisfactory numbers.

One complication that must be addressed in web-based evaluations is the lack of control over the subject pool. As we will show below, several demographic factors including language proficiency and gender had measurable effects on the

[4] See `http://www.give-challenge.org/research/page.php?id=software`.

results of the evaluation. We mitigated this concern by logging the IP addresses from which the clients connected (which can be resolved to countries) and asking users to specify their gender and self-rate their language proficiency in a questionnaire. We also made each player complete a tutorial before they started on the game world proper, to ensure at least a basic familiarity with controlling a 3D game.

2.4 Evaluation

Because the complete log of the user's actions and the NLG system's output is stored in the database, it is straightforward to extract a number of objective evaluation measures from the data. This includes measures such as the users' task success rate for each NLG system and each task instance, and the task completion times for successful runs. This data can be collected completely unobtrusively, without requiring any user intervention at all. In addition, experimental subjects can be asked to fill in a questionnaire to provide subjective data such as whether they found the generated texts helpful. Unfortunately, we are not aware of a reasonable way over the Web to force a user to actually provide meaningful answers to all items in a questionnaire.

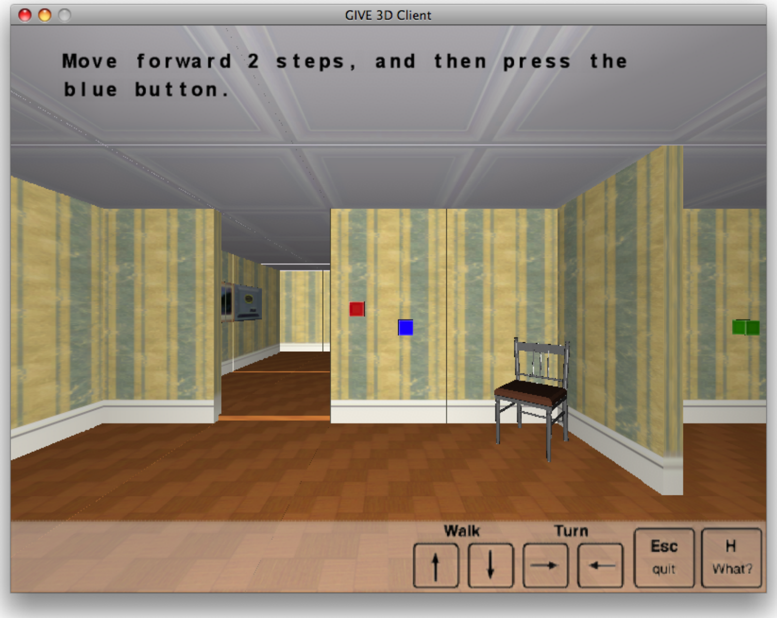
Based on the individual data points with the objective and subjective measures for each subject, it is possible to compute aggregate evaluation measures for each NLG system and determine which differences between these aggregate values are statistically significant. However, more fine-grained analyses are also possible, especially because each NLG system can leave behind timestamped messages in the log at any point, which can later be analyzed by its developers in detail.

Note that the approach to evaluation we have just presented focuses on *gathering* the data on which the evaluation should be based; we are not arguing for a specific set of appropriate measures for evaluating NLG systems, except that it must be possible to collect data for these measures online. In this way, our proposal is orthogonal to frameworks like PARADISE [27], which provide concrete evaluation measures and show how to combine them.

3 The GIVE Challenge

We will now describe how we implemented this data-gathering framework in 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 NLG system has access to a complete symbolic representation of the virtual world. The challenge is to generate, in real time, natural-language instructions that will guide the user 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 Figure 2, which displays instructions and allows them to move around in the world and manipulate objects. The first room they visit 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.



Move forward 2 steps, and then press the blue button.

Fig. 2. What the user sees when playing the GIVE game.

3.1 GIVE as a special case of the web-based method

In its reliance on connecting users and NLG systems over the Web, GIVE is a special case of the framework we presented in Section 2. GIVE specializes this general approach in two ways.

First, GIVE is specifically about generating *instructions* in *real time*. This differentiates GIVE from NLG research which focuses on the generation of descriptions or narratives. It also differentiates GIVE from related work which focuses on "batch" instructions, in which an entire discourse of instructions is presented at once, with the result that the instruction giver cannot react dynamically to the instruction follower's behavior.

Second, GIVE focuses on *situated* natural language generation. This makes the NLG task quite different to that found 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 [25]. That is, referring expression generation becomes a very different problem than the classical non-situated Dale and Reiter style referring expression generation [13], which focuses on generating referring expressions that are single noun phrases in the context of an unchanging world.

Both of these characteristics make GIVE a task that fits well with web-based evaluation: It is easy to judge whether and how easily a human user has followed real-time instructions, and by embedding the task in a virtual environment, it becomes possible to present it as an online game. On the other hand, the GIVE task is still open-ended enough to potentially be of interest to a wide range of NLG researchers. This is most obvious for research in sentence planning (i.e., issues such as referring expression generation, aggregation, and lexical choice) and realization, where the real-time nature of the task imposes high demands on the system's efficiency. But there are various ways of extending the task (e.g., to two-way dialog, speech output, or a different scenario) such that it also involves issues of prosody generation (i.e., research on text/concept-to-speech generation) and discourse generation. Finally, the game world can be designed to focus on specific issues in NLG, such as the generation of referring expressions or the generation of navigation instructions.

3.2 Game worlds

Each GIVE game run takes place in a *game world*, consisting of the map of a virtual environment along with descriptions of the objects in the world with their 3D positions and information about relationships between the objects: for example, the fact that pressing a certain button will open a certain door. The teams participating in GIVE-1 were given a *development world* early on in the challenge, against which they tested their systems.

The data gathering then took place on three new *evaluation worlds*. We made these worlds available to the NLG system developers one week before starting the data-gathering phase to ensure the compatibility of all systems with these worlds, but asked the developers not to specifically adapt their systems to these worlds. Figures 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 at a level of complexity similar to that of the development world. World 2 was intended to focus on referring expressions: 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.

All GIVE-1 worlds were split up into square tiles, and only permitted the user to jump from the center of one tile to the center of the next, and to turn in discrete 90-degree increments. The game client enforced the requirement that only such discrete movements could be made.

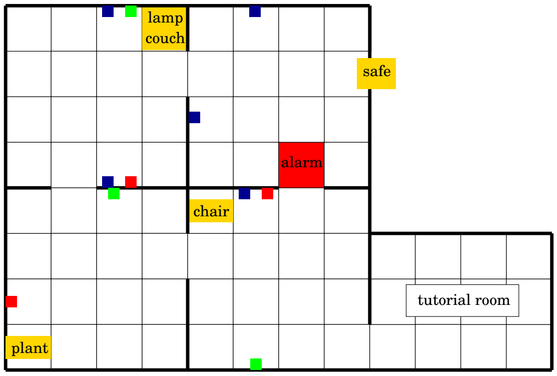


Fig. 3. World 1

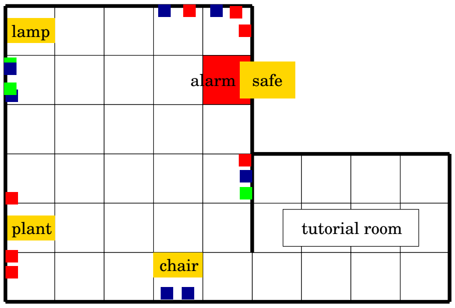


Fig. 4. World 2

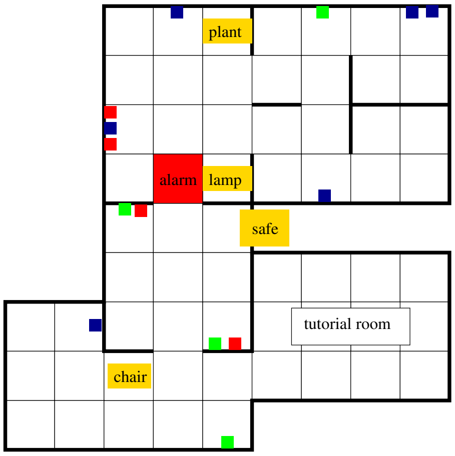


Fig. 5. World 3

3.3 Software infrastructure

The NLG system developers also had access to the GIVE software described in Section 2.2, and used it to develop their own systems.

The client we use in GIVE is responsible for displaying the virtual world, allowing the user to interact with it, and exchanging data with the NLG server over the Internet. To maximize portability, the client, like all other components of the GIVE software, is implemented in Java; users start it directly from the website using Java Web Start. Although the low-level details of drawing OpenGL graphics are handled by the jMonkeyEngine, a free 3D game library for Java, we still spent the majority of our development effort on the 3D graphics. We could have reduced this effort by building upon an existing virtual 3D world system such as Second Life. However, the effort needed to adapt such a system to our needs would have been at least as high (in particular, we would have

```
abstract class NlgSystem:
    void connectionEstablished();
    void connectionDisconnected();
    void handleAction(Atom actionInstance, List⟨Formula⟩ updates);
    void handleMoveTurnAction(Direction direction);
    void handleDidNotUnderstand();
    void handleStatusInformation(Position playerPosition,
                    Orientation playerOrientation, List⟨String⟩ visibleObjects);
    . . .
```

Fig. 6. The interface of an NLG system.

had to ensure that the user could only move according to the rules of the GIVE game and to instrument the virtual world to obtain real-time updates about events), and the result would have been less extensible to future installments of the challenge.

To simplify the job of the NLG system developers, we implemented a scaffold for the NLG system servers that handled all the necessary networking. This allowed system developers to focus on the development of their NLG systems. Specifically, they implemented concrete subclasses of the class `NlgSystem`, shown in Figure 6. This involved overriding the six abstract callback methods in this class with concrete implementations in which the NLG system reacts to specific events. The methods `connectionEstablished` and `connectionDisconnected` are called when users enter the game world and when they disconnect from the game. The method `handleAction` gets called whenever the user performs some physical action, such as pushing a button, and specifies what has changed in the world as a consequence of this action; `handleMoveTurnAction` gets called whenever the user moves; `handleDidNotUnderstand` gets called whenever the user presses the H key to signal that they didn't understand the previous instruction; and `handleStatusInformation` gets called once per second and after each user action to inform the server of the player's position and orientation, and the objects which are visible from that position. Ultimately, each of these method calls gets triggered by a message that the client sends over the network in reaction to some event; but this is completely hidden from the NLG system developer.

The NLG system can use the method `send` to send a string to the client to be displayed. It also has access to various methods for querying the state of the game world and to an interface to an external planner which can compute a sequence of actions leading to the goal. The planner that worked best in the GIVE planning domain, and which we provided in Linux and MacOS versions for GIVE-1, was SGPLAN 5.2.2 [18, 20].

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 the GIVE Challenge to the public in order to encourage people to play the GIVE game and serve as experimental subjects. Subjects were recruited via online press releases (in English and German), postings to email lists, and online gaming forums. As an incentive, one Amazon voucher was given away to a randomly-chosen subject each month.

4 Systems participating in GIVE-1

The four participating research teams submitted the following five systems to the evaluation:

System	Research Team
Austin [11]	University of Texas at Austin
Madrid [14]	Universidad Complutense de Madrid
Twente [23]	University of Twente
Union [26]	Union College
Warm-Cold [23]	University of Twente

We provide here a brief comparative overview of the systems; for more details see the papers cited in the above table.

The *Warm-Cold* system stands out as the only system that does not purely focus on generating instructions that are easy to understand and follow, but tries to create a more game-like and entertaining atmosphere. Instead of navigation instructions, it only provides the users with some feedback on whether they are getting closer to the next button they needed to press ("warmer") or are moving away from it ("colder").

The *Austin* team's primary focus was on improving the plans produced by the off-the-shelf planner provided by the organizers, which sometimes lead the user on unintuitive detours. Their system only retains the object manipulation instructions from the plan and uses A* search for path planning. Otherwise, the system uses a relatively simple generation strategy that maps plan actions to instructions almost step-by-step. The only aggregation that is done combines sequences of actions of the same type into one instruction: for example, it generates *move forward three steps* to combine three movement actions, or *turn around* to combine two turn left or turn right actions.

The other three systems, *Madrid*, *Twente*, and *Union*, all have the ability to switch between different levels or modes of instruction giving, which allow the systems, in certain situations, to use higher-level instructions that do not explicitly mention every single action in the plan.

The *Madrid* system directs the user to a door or button by describing the appearance and location of these objects, but without prescribing the individual steps for how to get to there. If this strategy is not possible—for example, because the object is not visible—the system guides the user from reorientation point to reorientation point until it becomes visible. This system produces relatively complex object descriptions. It uses concepts (such as spatial regions like rooms and corners) that are not available in the world representation provided by the GIVE framework, but are computed by the NLG system by analyzing the world map. Another aspect that distinguishes the *Madrid* system is that it pro-actively produces alerts or warnings to keep the user from doing potentially dangerous things.

The *Twente* team focused on making their system adapt to the user's behavior. The system distinguishes between three levels of direction giving: at the first level all instructions contain only one action type (like the *Austin* system), at the second level instructions can combine forward movements with another action (e.g., *walk forward 3 steps and then press the button*), and at the third level, once users can see the object that needs to be manipulated next, they are instructed to do so without prescribing the path to that object (similar to *Madrid*'s strategy). The system starts out in level two and then continuously (re-)estimates the user's success by counting how many actions they perform in 5 seconds. If they perform many actions, they are assumed to be doing well, and the instruction level goes up. If they carry out few actions, they are assumed to have problems and the level goes down. The system also switches to a lower level if the user explicitly asks for help by pressing "H".

Unfortunately, the referring expression generation module of the *Twente* system had a bug which sometimes resulted in the production of referring expressions where left and right were switched; this is fatal in evaluation worlds 2 and 3.

The *Union* team focused on producing landmark-based directions. Their system switches between a landmark mode and a path-based mode. The path-based mode is similar to *Twente*'s second level, in that instructions can combine movement actions with another action type. In the landmark mode, the system checks whether the object that needs to be manipulated next is visible. If so, it is described and the user is instructed to manipulate it without receiving any further instructions on how to reach it. If that object is not visible, the system tries to find another visible object along the way that can be used as a landmark. The path-based mode is only used if the landmark mode is not possible because there are no visible objects that can be used as landmarks, or if users are deemed to have problems because they press "H" or because they are not making enough progress toward their target.

5 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*) will lead the user to complete the task in a minimum number of steps; but it will require more instructions than a system that aggregates these. This tension is intentional, and emphasizes both the exploratory character of GIVE-1 and our desire to make GIVE a friendly comparative challenge rather than a competition with a clear winner. Our goal is to provide as many useful measures as possible in order to establish a framework in which research teams can evaluate and compare their systems along a variety of dimensions.

5.1 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 participants each, and the remaining 2% of participants connected from a further 42 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. Despite this geographical spread, 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 between Linux and Mac OS X.

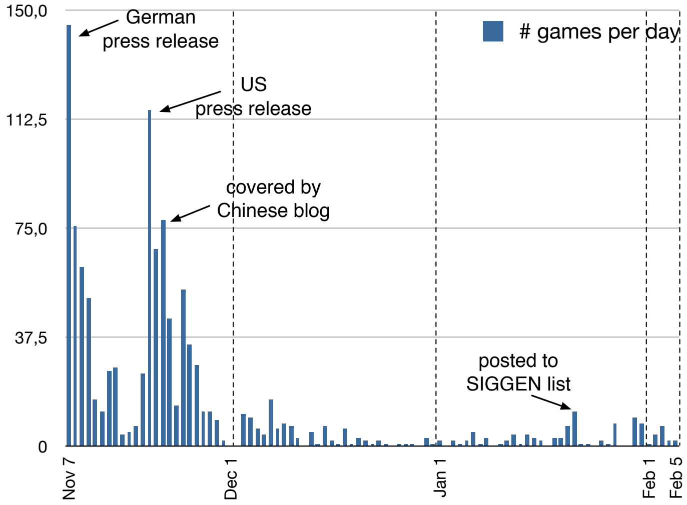


Fig. 7. Histogram of the connections per day.

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, 2008 to February 5, 2009 (Figure 7). 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. By comparison, posting an invitation to connect to the GIVE game to the mailing list of SIGGEN (the ACL's Special Interest Group on NLG) yielded a much weaker response. This illustrates the potential for recruiting experimental subjects over the Internet, compared to recruitment within the scientific community.

There is a risk that a user who plays twice will perform better the second time, because they may be more familiar with the task. We did not actively control this because we wanted to keep access to the online game as simple as possible. In the end, users from about 20% of the participating IP addresses connected multiple times. We believe this an acceptably low number, particularly because the duplicates are distributed evenly over the NLG systems.

5.2 Objective measures

We extracted objective and subjective measurements from the valid games. The objective measures are summarized in Figure 8. For each system and game world, we measured the percentage of games which were completed successfully. Furthermore, for each game we counted the number 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). 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.

- 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 actions.*)
- second (Time in seconds.*)

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

Fig. 8. Objective measurements

Figure 9 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.

	Austin	Madrid	Twente	Union	Warm-Cold
task success	40%	71%	35%	73%	18%
		A		A	
	B		B		
					C
instructions	83.2	58.3	121.2	80.3	214.1
		A			
	B			B	
			C		
					D
steps	103.6	124.3	160.9	117.5	307.4
	A			A	
		B		B	
			C		
					D
actions	11.2	8.7	14.8	9.0	15.1
		A		A	
	B				
			C		C
seconds	129.3	174.8	207.0	175.2	320.7
	A				
		B		B	
			C		
					D

Fig. 9. *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 $p < 0.05$.

Overall, there is a top group consisting of the *Austin*, *Madrid*, and *Union* systems: while *Madrid* and *Union* outperform *Austin* on task success (with 70–80% 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 cat-

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

Fig. 10. Questionnaire items.

egories. This confirms the ability of the approach described here to distinguish the performances of different systems.

5.3 Subjective measures

The subjective measures, which were obtained by asking the users to fill in a questionnaire after each game, are shown in Figure 10. All of the questions were answered on 5-point Likert scales, with the exception of "overall", which used a 7-point scale, and the "informativity" and "timing" questions, which had nominal answers. For each question, the user could choose not to answer.

The results of the subjective measurements are summarized in Figure 11, in the same format as for the objective measurements. 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.

	Austin	Madrid	Twente	Union	Warm-Cold
overall	4.9 A	4.9 A	4.3 B	4.6 A B	3.6 C
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
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
friendliness	3.4 A B	3.8 A	3.1 B	3.6 A	3.1 B
informativity	46% B	68% A	51% B	56% B	51% B
timing	78% A	62% B	60% B C	62% B	49% C

Fig. 11. *Subjective* measures by system. Informativity and timing are reported as the percentage of successfully completed games in which users chose "just right". The other measures are the mean ratings reported by the players. Letters group indistinguishable systems; systems that don't share a letter were found to be significantly different with $p < 0.05$.

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.

5.4 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 Figures 12 and 13 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. World 2 was especially 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 no longer have an effect on the task success rate for any of the systems.

We also asked the participants to rate their computer expertise and how many hours of video games they played per week. Neither of these showed an effect on our measures.

6 Validating the experimental approach

We have just seen that the web-based data gathering method provided informative results for the GIVE Challenge in that we managed to access a large number

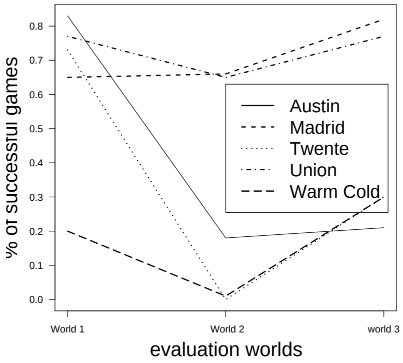


Fig. 12. Effect of the evaluation worlds on the success rate of the NLG systems.

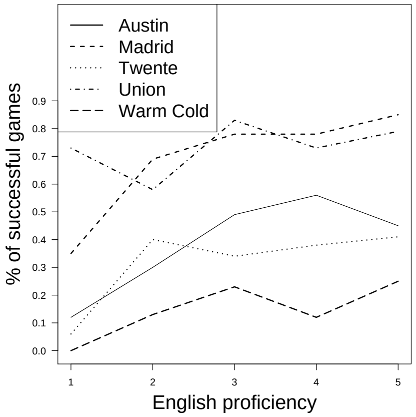


Fig. 13. Effect of the players' English skills on the success rate of the NLG systems.

of experimental subjects and detect a number of significant differences between the different NLG systems. However, one could argue that, because of the lack of control we have over the selection of subjects over the Internet, these results may be artificially skewed or less precise than the results we would have obtained with a more traditional laboratory-based experiment. To allay such concerns, we repeated part of the GIVE-1 Challenge evaluation in a laboratory setting and compared the results to those of the web-based setting.

6.1 The laboratory experiment

For this laboratory-based evaluation, we recruited 91 participants from a college campus. Each participant played the GIVE game once with each of the five NLG systems, in different orders. To avoid learning effects, we only used the first game run from each subject in the comparison with the web experiment; as a consequence, subjects were distributed evenly over the NLG systems. To accommodate for the much lower number of participants, the laboratory experiment only used a single game world: World 1, which was known from the online version to be the easiest world.

Among this group of subjects, 93% self-rated their English proficiency as "expert" or better; 81% were native speakers. In contrast to the online experiment, 31% of participants were male and 65% were female (4% did not specify their gender).

6.2 Results

Figure 14 shows the results of the laboratory experiment for the objective measures collected. The table also includes the results for the Internet experiment on World 1; this makes the data comparable between the two experiments, and is in contrast to the data in Figure 9, which reports aggregate data for all three worlds. The task success rate is only evaluated on games that were completed successfully or lost, not cancelled, as laboratory subjects were asked not to cancel. This brings the number of Internet subjects to 322 for the success rate, and to 227 (only successful games) for the other measures. Significance was tested and is reported as in Section 5.2.

	Web experiment					Laboratory experiment				
	Austin	Madrid	Twente	Union	Warm-Cold	Austin	Madrid	Twente	Union	Warm-Cold
	91%	76%	85%	93%	24%	100%	95%	93%	100%	17%
task	A		A	A		A	A	A	A	
success		B	B	B						B
					C					
	83.4	68.1	97.8	99.8	159.7	78.2	66.3	107.2	88.8	134.5
		A				A	A			
instructions	B					B			B	
			C	C				C	C	
					D			D		D
	99.8	145.1	142.1	142.6	256.0	93.4	141.8	134.6	128.8	213.5
steps	A					A				
		B	B	B			B	B	B	
					C					C
	9.4	10.0	9.7	10.3	9.6	9.9	10.5	9.6	9.8	10.0
actions	A	A	A		A	A	A	A	A	A
		B	B	B	B					
	123.9	195.4	174.4	194.0	234.1	143.9	211.8	205.6	195.1	252.5
seconds	A					A			A	
		B	B	B			B	B	B	B
		C		C	C					

Fig. 14. Comparison of the five NLG systems in terms of the *objective* measures collected in the *web* (left) and the *laboratory* (right) experiments on World 1.

The results for the subjective measures are summarized in Figure 15, in the same format and using the same analysis method as for the objective measures. Also as above, the table is based only on games in World 1, and on games that were completed successfully. We justify this latter choice below. One consequence, however, is that the results for the *Warm-Cold* system in the lab experiment are based on only two subjects (all others lost when working with *Warm-Cold*) and may therefore not be meaningful.

	Austin	Madrid	Twente	Union	Warm-Cold	Austin	Madrid	Twente	Union	Warm-Cold
overall	5.6	5.0	4.5	4.5	4.8	5.7	5.4	4.9	5.7	5.0
	A	A			A	A	A	A	A	A
		B	B	B	B					
task difficulty	4.9	4.3	4.3	4.4	4.0	4.9	4.0	4.4	4.3	4.0
	A					A		A	A	A
		B	B	B	B		B	B	B	B
goal clarity	4.9	4.1	4.2	4.0	4.2	4.9	3.7	4.3	4.6	3.0
	A				A	A		A	A	A
		B	B	B	B		B	B	B	B
play again	2.2	2.7	2.2	2.8	3.3	2.9	2.7	1.9	1.9	1.0
	A	A	A	A	A	A	A	A	A	A
instruction clarity	4.9	3.9	4.1	3.9	3.5	4.8	3.8	4.4	4.8	2.0
	A					A		A	A	
		B	B	B	B		B	B		
							C			C
instruction helpfulness	4.6	4.1	3.8	3.6	3.24	5.0	4.4	3.3	4.5	4.0
	A	A				A	A		A	A
		B	B	B	B			B		B
choice of words	4.7	3.8	4.4	4.0	3.8	4.7	3.8	4.5	4.7	4.5
	A		A			A		A	A	A
		B	B	B	B		B	B		B
referring expressions	4.7	4.0	4.3	4.0	4.2	4.8	4.3	4.4	4.3	4.0
	A		A		A	A	A	A	A	A
		B	B	B	B					
navigation instructions	4.6	4.0	4.1	3.8	3.4	4.7	3.7	4.1	4.3	4.0
	A		A			A	A	A	A	A
		B	B	B	B					
friendliness	3.8	3.9	3.3	3.7	3.5	3.9	4.3	3.5	4.1	3.0
	A	A	A	A	A	A	A	A	A	A
informativity	63%	67%	48%	62%	59%	77%	84%	43%	75%	100%
	A	A	A	A	A	A	A	A	A	A
timing	81%	70%	73%	51%	50%	92%	95%	64%	100%	100%
	A	A	A			A	A	A	A	
		B	B		B	B	B	B		B
		C		C	C					

Web experiment **Laboratory experiment**

Fig. 15. Comparison of the five NLG systems in terms of *subjective* measures collected in the *web* (left) and the *laboratory* (right) experiments on World 1.

6.3 Discussion

The primary question that interests us in a comparative evaluation is which NLG systems performed significantly better or worse on any given evaluation measure. In the experiments above, we find that of the 170 possible significant differences (= 17 measures × 10 pairs of NLG systems), the laboratory experiment only found six that the web-based experiment didn't find. Conversely, there are 26 significant differences that only the Internet-based experiment found. But even more importantly, all pairwise rankings are consistent across the two evaluations: Where both systems found a significant difference between two systems, they always ranked them in the same order. We conclude that the Internet experiment provides significance judgments that are comparable to, and in fact more precise than, the laboratory experiment.

Nevertheless, there are important differences between the laboratory and Internet-based results. For instance, the success rates in the laboratory tend to be higher than on the Internet, but so are the completion times. We believe that these differences can be attributed to the demographic characteristics of the participants in the two experiments. To substantiate this claim, we looked in some detail at three of these differences: gender, language proficiency, and questionnaire response rates.

First, the gender distribution differed greatly between the Internet experiment (10% female) and the laboratory experiment (65% female). This is relevant because gender had a significant effect on task completion time (women took longer) and on six subjective measures including "overall evaluation" in the laboratory. We speculate that the difference in task completion time may be related to reported gender differences in processing navigation instructions [21].

Second, the two experiments collected data from subjects with different language proficiencies. While 93% of the participants in the laboratory experiment self-rated their English proficiency as "expert" or better, only 62% of the Internet participants did. This partially explains the lower task success rates on the Internet, as Internet subjects with English proficiencies of 3–5 performed significantly better on "task success" than the group with proficiencies 1–2. If we only look at the results of high-English-proficiency subjects on the Internet, the success rates for all NLG systems except *Warm-Cold* rise to at least 86%, and are thus close to the laboratory results.

Finally, the Internet data are skewed by the tendency of unsuccessful participants to not fill in the questionnaire. Figure 16 summarizes some data about the "overall evaluation" question. Users who didn't complete the task successfully tended to judge the systems much lower than successful users, but at the same time tended not to answer the question at all. This skew causes the mean subjective judgments across all Internet subjects to be artificially high. It is to make the data more comparable with respect to this that Figure 15 includes only judgments from successful games in both the laboratory and Internet experiments.

In summary, we find that while the two experiments made consistent significance judgments, and the Internet-based evaluation methodology thus produces

Web

	# of games	reported	mean
success	227 = 61%	93%	4.9
lost	92 = 24%	48%	3.4
cancelled	55 = 15%	16%	3.3

Lab

	# of games	reported	mean
success	73 = 80%	100%	5.4
lost	18 = 20%	94%	3.3
cancelled	0	–	–

Fig. 16. Skewed results for "overall evaluation". "Reported" is the percentage of subjects who answered the "overall evaluation" question after having succeeded/lost/cancelled the game. "Mean" is the mean "overall evaluation" score.

meaningful results, the absolute values they find for the individual evaluation measures differ due to the demographic characteristics of the participants in the two studies. This could be taken as a possible disadvantage of the Internet-based evaluation. However, we believe the opposite to be the case. In many ways, an online user is in a much more natural communicative situation than a laboratory subject who is being discouraged from cancelling a frustrating task. In addition, every experiment, whether in the laboratory or on the Web, suffers from some skew in the subject population due to sampling bias; for instance, one could argue that an evaluation that is based almost exclusively on native speakers in universities leads to overly benign judgments about the quality of NLG systems.

One advantage of the Internet-based approach to data collection over the laboratory-based one is that, due to the sheer number of subjects, we can detect such skews and deal with them appropriately. For instance, we might decide that we are only interested in the results from proficient English speakers and ignore the rest of the data; but we retain the option to run the analysis over all participants, and to analyze how much each system relies on the user's language proficiency. The amount of data also means that we can obtain much more fine-grained comparisons between NLG systems. For instance, the second and third evaluation worlds specifically exercised an NLG system's abilities in generating referring expressions and navigation instructions respectively, and there were significant differences in the performance of some systems across different worlds. Such data, which is highly valuable for pinpointing specific weaknesses of a system, would have been prohibitively costly and time-consuming to collect using laboratory subjects.

7 Conclusion

In this paper, we have described GIVE-1, the first installment of the GIVE Challenge. GIVE uses a novel evaluation methodology for NLG systems: It connects

NLG systems to human subjects over the Internet and evaluates them according to several objective measures related to task success, as well as a variety of subjective measures in a questionnaire. GIVE-1 collected data from 1143 valid games over a period of three months, and found a number of significant differences between the five NLG systems under evaluation. We established that these results are comparable to, but more precise than, those gained from a laboratory-based version of the evaluation, and thus validated the web-based data-gathering strategy for NLG systems.

We are currently running GIVE-2, which differs from GIVE-1 in that it allows users to move and turn freely, as opposed to the discrete moves and turns they could execute in the GIVE-1 client. This makes the NLG task much harder. For instance, the *Austin* system exclusively used navigation instructions of the form *walk three steps forward* and *turn left*; but *three steps* is not a distance measure that a GIVE-2 user will be able to interpret, and it is an open research question how far a user will turn after the instruction *turn left*.

Beyond this, there are many directions in which one could take GIVE in the future. In particular, it would be interesting to try GIVE with spoken rather than written language generation; to extend GIVE to a dialogue challenge by allowing the instruction follower to speak; and to reverse GIVE into a challenge for instruction *understanding* systems. More generally, we plan to generalize the GIVE software into a generic platform for the Internet-based evaluation of NLG systems. This would then allow evaluators to replace the specific implementations of the GIVE client and NLG system interface by tools that are suitable to their domains, while retaining the networking, database, and world management backbone that the GIVE software provides.

Acknowledgments. We are grateful to the participants of the 2007 NSF/SIGGEN Workshop on Shared Tasks and Evaluation in NLG and the ENLG 2009 workshop, and to 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. We are indebted to Sara Dalzel-Job for running the laboratory-based experiment at the University of Edinburgh. Particular thanks go to Mariet Theune for her insightful and detailed comments on this paper. The creation of the GIVE infrastructure was supported in part by a Small Projects grant from the University of Edinburgh and by the Cluster of Excellence "Multimodal Computing and Interaction" at Saarland University.

References

1. von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proceedings of the ACM CHI Conference (2004)

2. Bangalore, S., Rambow, O., Whittaker, S.: Evaluation metrics for generation. In: Proceedings of the First International Natural Language Generation Conference (INLG2000), Mitzpe Ramon. pp. 1–8 (2000)
3. Belz, A.: That's nice ... what can you do with it? Computational Linguistics 35(1), 111–118 (2009)
4. Belz, A., Gatt, A.: Intrinsic vs. extrinsic evaluation measures for referring expression generation. In: Proceedings of ACL-08:HLT, Short Papers. pp. 197–200. Columbus, Ohio (2008)
5. Belz, A., Eric Kow, J.V., Gatt, A.: Generating referring expressions in context: The GREC task evaluation challenges. In: Krahmer, E., Theune, M. (eds.) Empirical Methods in Natural Language Generation, LNCS, vol. 5980. Springer, Berlin / Heidelberg (2010)
6. Belz, A., Kow, E.: System building cost vs. output quality in data-to-text generation. In: Krahmer, E., Theune, M. (eds.) Empirical Methods in Natural Language Generation, LNCS, vol. 5980. Springer, Berlin / Heidelberg (2010)
7. Belz, A., Reiter, E.: Comparing automatic and human evaluation of NLG systems. In: Proceedings of EACL-06. pp. 249–256. Trento, Italy (2006)
8. Cahill, A., Forst, M.: Human evaluation of a german surface realisation ranker. In: Krahmer, E., Theune, M. (eds.) Empirical Methods in Natural Language Generation, LNCS, vol. 5980. Springer, Berlin / Heidelberg (2010)
9. Callison-Burch, C., Osborne, M., Koehn, P.: Re-evaluating the role of Bleu in machine translation research. In: Proceedings of EACL-06. pp. 249–256. Trento, Italy (2006)
10. Chamberlain, J., Poesio, M., Kruschwitz, U.: Addressing the Resource Bottleneck to Create Large-Scale Annotated Texts. In: Bos, J., Delmonte, R. (eds.) Proceedings of the Symposium on Semantics in Text Processing (STEP), pp. 375–380 (2008)
11. Chen, D., Karpov, I.: The GIVE-1 Austin system. In: Proceedings of the First NLG Challenge on Generating Instructions in Virtual Environments (2009), available at `http://www.give-challenge.org/research`
12. Dale, R., White, M. (eds.): Proceedings of the NSF/SIGGEN Workshop for Shared Tasks and Comparative Evaluation in NLG. Arlington, VA (2007)
13. Dale, R., Reiter, E.: Computational interpretations of the Gricean maxims in the generation of referring expressions. Cognitive Science 19(2), 233–263 (1995)
14. Dionne, D., de la Puente, S., n, C.L., Hervás, R., Gervás, P.: Guide. In: Proceedings of the First NLG Challenge on Generating Instructions in Virtual Environments (2009), available at `http://www.give-challenge.org/research`
15. Foster, M.E.: Automated metrics that agree with human judgements on generated output for an embodied conversational agent. In: Proceedings of INLG 2008. pp. 95–103. Salt Fork, OH (2008)
16. Gatt, A., Belz, A., Kow, E.: The TUNA-REG challenge 2009: Overview and evaluation results. In: Proceedings of the 12th European Workshop on Natural Language Generation. pp. 174–182 (2009)
17. Gatt, A., Belz, A., Kow, E.: Comparative evaluation of referring expression generation: The TUNA shared task evaluation challenges. In: Krahmer, E., Theune, M. (eds.) Empirical Methods in Natural Language Generation, LNCS, vol. 5980. Springer, Berlin / Heidelberg (2010)
18. Hsu, C.W., Wah, B.W., Huang, R., Chen, Y.X.: New features in SGPlan for handling soft constraints and goal preferences in PDDL 3.0. In: Proceedings of the Fifth International Planning Competition, 16th International Conference on Automated Planning and Scheduling. pp. 39–41 (2006)

19. Keller, F., Gunasekharan, S., Mayo, N., Corley, M.: Timing accuracy of web experiments: A case study using the WebExp software package. Behavior Research Methods 41(1), 1–12 (2009)
20. Koller, A., Petrick, R.: Experiences with planning for natural language generation. In: Proceedings of SPARK-08: The ICAPS-08 Scheduling and Planning Applications Workshop. Sydney, Australia (2008)
21. Moffat, S., Hampson, E., Hatzipantelis, M.: Navigation in a "virtual" maze: Sex differences and correlation with psychometric measures of spatial ability in humans. Evolution and Human Behavior 19(2), 73–87 (1998)
22. Orkin, J., Roy, D.: The restaurant game: Learning social behavior and language from thousands of players online. Journal of Game Development 3(1), 39–60 (2007)
23. Rookhuiszen, R.B., Obbink, M., Theune, M.: Two approaches to GIVE: dynamic level adaptation versus playfulness. In: Proceedings of the First NLG Challenge on Generating Instructions in Virtual Environments (2009), available at `http://www.give-challenge.org/research`
24. Stent, A., Marge, M., Singhai, M.: Evaluating evaluation methods for generation in the presence of variation. In: Proceedings of CICLing 2005 (2005)
25. Stoia, L., Shockley, D.M., Byron, D.K., Fosler-Lussier, E.: Noun phrase generation for situated dialogs. In: Proceedings of INLG. Sydney (2006)
26. Striegnitz, K., Majda, F.: Landmarks in navigation instructions for a virtual environment. In: Proceedings of the First NLG Challenge on Generating Instructions in Virtual Environments (2009), available at `http://www.give-challenge.org/research`
27. Walker, M., Litman, D., Kamm, C., Abella, A.: PARADISE: A framework for evaluating spoken dialogue agents. In: Proceedings of ACL97. pp. 271–280. Madrid, Spain (1997)
28. Walker, M., Rudnicky, A., Prasad, R., Aberdeen, J., Bratt, E.O., Garofolo, J., Hastie, H., Le, A., Pellom, B., Potamianos, A., Passonneau, R., Roukos, S., Sanders, G., Seneff, S., Stallard, D..: DARPA communicator: Cross-system results for the 2001 evaluation. In: ICSLP-2002:Inter. Conf. on Spoken Language Processing. vol. 1, pp. 273–276. Denver, CO USA (2002)