Czech Technical University
Faculty of Nuclear Sciences and Physical Engineering

Department of Software Engineering
Field: Engineering Informatics
Specialization: Applied Software Engineering

Navigation in virtual environment
Navigování ve virtuálním světě

MASTER THESIS

Author: Václav Honzík
Supervisor: Doc. Ing. Zdeněk Žabokrtský, Ph.D.
Year: 2013

Insert assignment instead of this page before handing in the thesis!!!!

Declaration

I declare, that I have developed my master thesis independently and used only the materials (literature, projects, SW etc.) listed in attached list.

In Prague on .....

.....
Václav Honzík

Acknowledgment

I would like to thank Doc. Ing. Zdeněk Žabokrtský, Ph.D.for supervising and shaping my master thesis. I would like to also thank Prof. Kristina Striegnitz, Ph.D.for giving me the opportunity and helping me with my master thesis.

Václav Honzík

Title:
Navigation in virtual environment

Author: Václav Honzík

Field: Engineering Informatics
Type of thesis: Master thesis

Supervisor: Doc. Ing. Zdeněk Žabokrtský, Ph.D.
Institute of Formal and Applied Linguistic, Faculty of Mathematics and Physics, Charles University in Prague
Advisor: Prof. Kristina Striegnitz, Ph.D.
Union College, Schenectady USA
Abstract: abstract en

Key words: keywords en

Název práce:
Navigování ve virtuálním světě

Autor: Václav Honzík

Obor: Inženýrská Informatika
Typ práce: Diplomová práce

Vedoucí: Doc. Ing. Zdeněk Žabokrtský, Ph.D.
Institute of Formal and Applied Linguistic, Faculty of Mathematics and Physics, Charles University in Prague
Konzultant: Prof. Kristina Striegnitz, Ph.D.
Union College, Schenectady USA
Abstrakt: abstract cz

Klíčová slova: keywords cz

Contents

Introduction

Chapter 1

GIVE Challenge

Big part of this master thesis revolves around GIVE Challenge. The data I was using to develop the hypothesis where from annotated GIVE experiment. I used GIVE framework to implement and test my hypothesis. Therefore, in this chapter I will describe this academic competition in detail.

First section will answers basic questions such as what is GIVE Challenge, why was it created or what are its interesting properties. In the next section, I will provide a brief history of GIVE Challenge together with some of its results. In the third section, the focus will be a detailed description of the shared task and the virtual world of GIVE Challenge.

1.1 Introduction

GIVE Challenge was a series of Natural Language Generation (NLG) competitions run from November 2008 to March 2012. Participants were developing NLG systems to navigate human-controlled avatars in a 3D virtual environment. The real-time navigation was realized through written instructions displayed on the screen. Goal of the navigation was to finish a treasure-hunt game. In figure 1.1 we can see a client software with virtual world and example of an instruction. More detailed description of the task and the environment is in the section 1.3.

Koller et al. (2010a) state that one of the goals of GIVE Challenge was spawning interest in NLG subfield of computational linguistics (CL), inspired by other competitions in this field such as Recognizing Textual Entailment challenge[1] and NIST machine translation competition[2].

According to Koller et al. (2010a), another important goal was to introduce and explore a new way of evaluating NLG algorithms, techniques and systems in a shared task. More specifically a shared task which was, on one hand, complex enough to encompass multiple NLG subtasks and, on the other hand, was only concerned with

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

[2] http://www.itl.nist.gov/iad/mig//tests/mt/

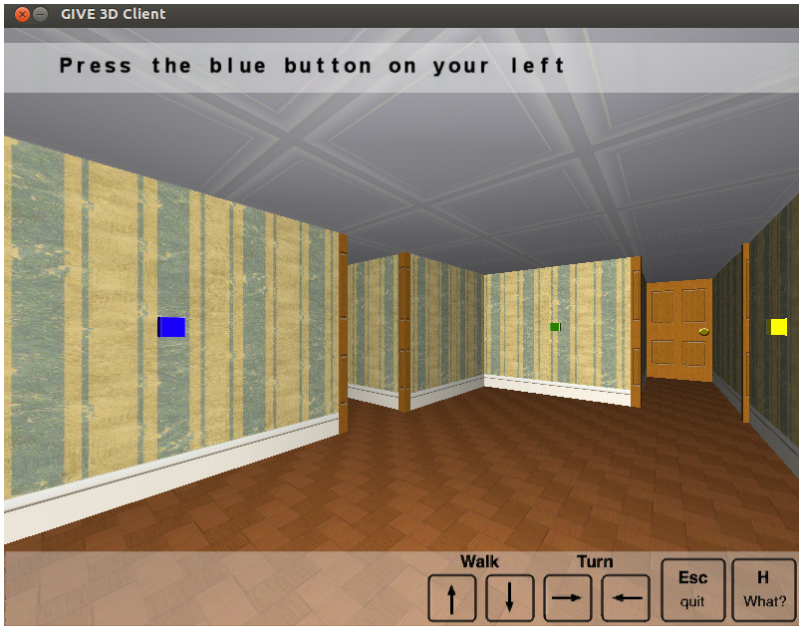

Figure 1.1: A human subject is being navigated through the environment.

NLG and not any other fields of the computational linguistic.

Three basic approaches to evaluation of NLG systems are comparison with annotated corpora, measuring task performance in an experiment and human judges evaluation. Koller et al. (2010a) argue the advantages and disadvantages of these evaluation in more depth, therefore I will only provide a brief summary. The first approach compares output of the NLG system to an annotated corpora, also known as a gold-standard. It is fast and cheap approach, but a problem with it lies in the complexity of the natural language. We can often express concepts in many different ways and there is often no telling which way is a better one. The second approach conducts an experiment and measures task performance on human subjects. Measuring task performance avoids the problems of gold-standard, but it is expensive and time consuming. Lastly, trained human judges are used to evaluate the system. It is less demanding than the second approach, but for the cost of certainty, that the results correspond with results one might achieve with non-expert subjects.

GIVE Challenge proposes and successfully implements a new, in a sense that it wasn't used for NLG before, approach through Internet-based evaluation. The basic premise is using a client-server software methodology. The client is a program installed on test subject computer, which is easily downloadable from a public website. The client connects through the Internet to a matchmaker server and random evaluation world is selected. Matchmaker also connects the client to a randomly selected NLG system, which itself, can be hosted on a different server. Client and NLG systems then communicate back and forth until the task is finished. Matchmaker finally logs the entire sessions to database. Figure 1.2 shows that architecture in a simple diagram.

This approach immediately presents several advantages. It does not require physical presence of the test subject in a laboratory. The subject simply downloads the client

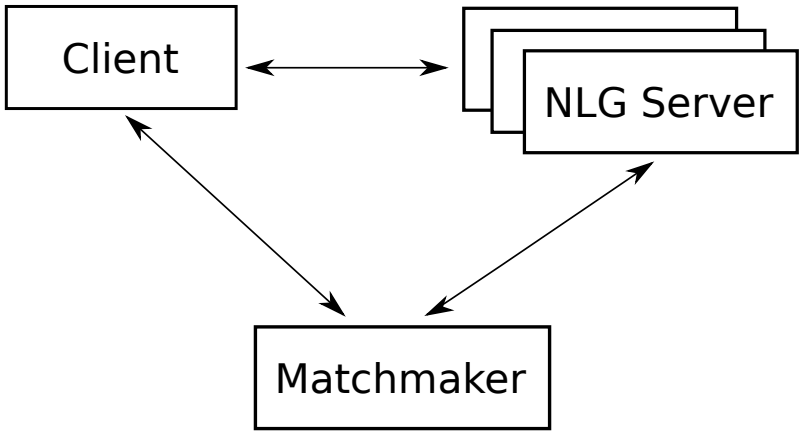

Figure 1.2: Software architecture of GIVE Challenge.

from a website and is able to do the experiment at his/her convenience. Second obvious advantage is a scalability. The number of individuals which can parallelly undertake the experiment is only limited by the servers' load. Thanks to the low costs, advertisement becomes the decisive limitation on the number of subjects. Take for example second instalment of GIVE Challenge which had up to 1800 participants.

On the other hand, part of the control over the experiment is lost in this approach; for example the control over subject pool. Another problem which rises with this approach is that individuals can repeat the experiment.

In addition to Internet-based evaluation, the GIVE Challenge utilizes variety of evaluation measures of both objective and subjective nature. Among objective measure are task success rate, number of instructions or time required to finish the task. For subjective measures a questionnaire was used at the end of the session. The questionnaire mostly used a 5 point scale with question such as how clear where the instruction or how friendly was the system. Some of the measures intentionally collided with each other, putting emphasize on a certain characteristic of the system.

Having presented the basic concept of GIVE Challenge and reasons for its creation, I will now move onto brief history of this competition.

1.2 History of GIVE Challenge

The first instalment of GIVE Challenge (GIVE-1) was publicized in March 2008. Koller et al. (2010a) report on this instalment and are the source of following information. For more details please refer to their paper. The data collection period was from November 2008 to February 2009. Four teams participated in this challenge, namely from these universities: University of Texas at Austin, Universidad Complutense de Madrid, University of Twente and Union College. The team from University of Twente submitted two systems, making the final number of systems five.

What is important to note about GIVE-1 is a different world representation from the

following instalments. GIVE-1 used discrete square grid for player movement. Player was able to rotate only by 90Â° and walk forward and backwards by one square of the grid. That had a major impact on the design of NLG systems. Participating teams atleast occasionally used this grid in their references (eg. *move forward three steps*). Afterwards organizers realized that the grid and the discrete movement made the task easier than intended and they were after GIVE-1 removed.

Altogether, 1143 valid games were recorded. The demographics featured a majority of males (over 80%) and wide spread over different countries in the world. For the actual results, the system from Austin significantly outperformed all other systems in task completion time. At the same time systems from Union and Madrid outperformed other systems in success rate. That shows the significance of different measures for the evaluation. Similar interesting conclusion in both objective and subjective measures can be found in previously mentioned paper. Apart from objective and subjective measures, the report examined influence of English language proficiency and differences between evaluation worlds. The English proficiency had an impact on the task success rate but solely for the least proficient category. The evaluation world also had a significant influence on the task success rate.

Finally, the first instalment also compared the Internet-based evaluation with more standard laboratory evaluation. The conclusion was that Internet-based evaluation provides meaningful results comparable and even more precise in some areas to the laboratory setting.

The second instalment (GIVE-2) run from August 2009 (data collection starting in February 2010) to May 2010 and is thoroughly described by Koller et al. (2010b). Following information are based on this paper. Biggest difference to the GIVE-1, which was mentioned previously, is that players were now able to move freely. This made the instruction generation considerably harder. Additionally, the questionnaire was revised and a few new objectives measures were introduced. Evaluation worlds used in GIVE-2 were considerably harder than in GIVE-1. Number of distracting buttons was increased and same-colored buttons were in some cases next to each other. Also number of alarm tiles was increased. Otherwise, the architecture and the rest of the details stayed the same as in GIVE-1.

This time 1825 games were played over seven NLG systems developed by six teams from: Dublin Institute of Technology, Trinity College Dublin, Universidad Complutense de Madrid, University of Heidelberg, Saarland University and INRIA Grand-Est in Nancy (2 systems).

There was a big drop in success rate, most likely linked to the free movement and the increase of difficulty in the evaluation worlds. Similarly to results in GIVE-1, there was an influence of English proficiency and game world on the task success rate. Additionally, age of the subject played a role in the time required to finish the task and number of actions to finish the task (younger subjects being faster and requiring less actions). The difference between genders in time required to finish the task disappeared in GIVE-2.

Following GIVE-2 was so called Second Second instalment (GIVE-2.5), which kept almost the same settings as GIVE-2. There was just a small addition to objective

measures and a reduce in the number of subjective questions. The data collection took place between July 2011 and Mar 2012. Striegnitz et al. (2011) reports on the partial results of 536 valid games from July and August 2011, which however constitute a majority of the final number of 650 valid games.

Eight NLG systems participated from 7 teams: University of Aberdeen, University of Bremen, Universidad Nacional de CÃłrdoba, Universidad Nacional de CÃłrdoba and LORIA/CNRS, LORIA/CNRS, University of Potsdam (2 systems) and University of Twente. In this instalment the teams employed more broad spectrum of approaches. Team from University of Bremen used decision trees learned from GIVE-2 corpus. Universidad Nacional de CÃłrdoba and LORIA/CNRS, LORIA/CNRS selected instructions from a corpus of human to human interactions. The teams also often included algorithms from existing NLG and CL literature.

Apart from comparing the systems through objective and subjective measures, Striegnitz et al. (2011) again examined effects of evaluation worlds and demographics factors on task success rate. The evaluation worlds and the English proficiency had an effect. Additionally computer expertise and familiarity with computer games significantly influenced the task performance. The difference between male and female subject wasn't significant.

Following section describes the shared task in more detail and lists possible contents of GIVE virtual worlds.

1.3 Task and GIVE world

GIVE world is a 3D virtual world. The world is indoor in a sense, that it comprises rooms connected by doors. It's defined in a human-readable format and stored in text file. Following objects can be places in a world:

- Alarm tile
- Button
- Door
- Landmark
  - Bed
  - Chair
  - Couch
  - Dresser
  - Flower
  - Lamp
  - Table
  - Window

- Picture
- Safe
- Trophy
- Wall

In addition, some of these objects can have attributes, states or can operate other objects. Buttons have colors as an example of attribute. Doors and safes can be in a closed or an open state. Buttons can operate doors, safes or pictures.

Walls are actually created automatically by defining shapes of rooms. Rooms can have rectangular shape or can be defined by a polygon. I will sometimes use a term "corridor" which is a connecting room, usually not containing any button.

The landmarks serve as a decoration but they can be used in an expression generation. Picture is technically a landmark as well, but in GIVE Challenge it often serves another purpose. It covers the safe and needs to be put aside by a button press.

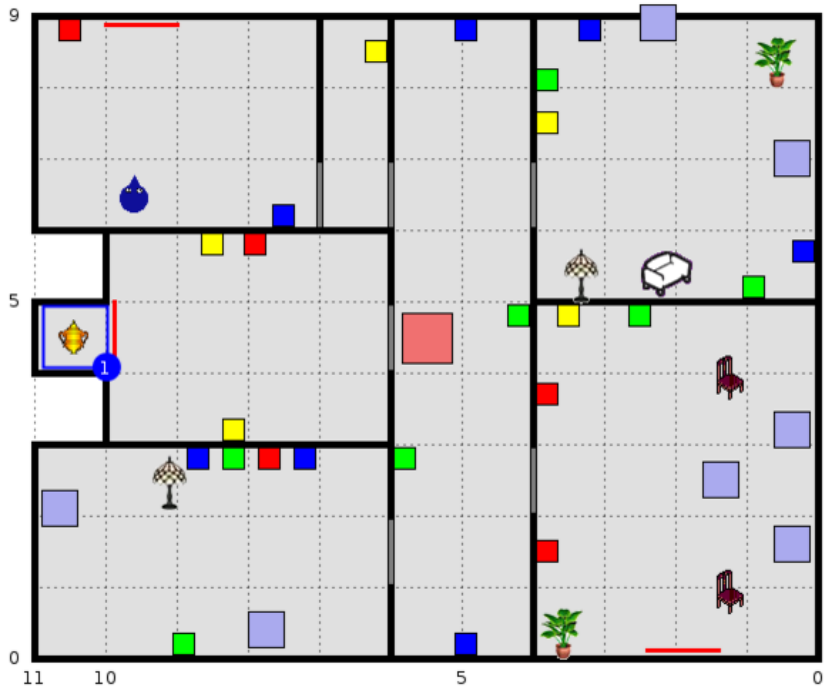

Figure 1.3: Example GIVE world viewed in GIVE map viewer utility.

Figure 1.3 shows an evaluation world number one from GIVE-2.5. In top-left room we can see player starting position. Buttons are colored squares on the walls. Grey bars on the walls are closed doors. Trophy in a safe is in the middle-left room. There are also landmarks (like lamp or chair) and one big read square marking an alarm tile.

The flexibility of GIVE world creation allows relatively broad range of scenarios for the task. On the other hand, all the GIVE Challenge instalments consisted of similar sequence of steps.

The goal of all the GIVE Challenge worlds is to pick up a trophy. The trophy is hidden in a closed safe. In order to open the safe a sequence of buttons, usually counting somewhere around 6 buttons, has to be pressed. The safe can be also hidden by a picture, which needs to be put aside. The buttons in a safe-opening series are often in different rooms. Rooms can be also closed off, requiring another button press to open the door. While moving around the world, player has to avoid alarm tiles. Stepping on an activated alarm tile causes an immediate loss. Alarm tiles can also block the path and need to deactivated by a button press. Some buttons also cause an alarm and an immediate loss.

Depending on the number of rooms, complexity of buttons arrangement and length of safe-opening sequence, the task can range from short and trivial problem easily handled by a few instruction templates, to a long and hard case, where it's impossible to capture every possible scenario.

To summarize, navigation system of a player in the GIVE world has to deal with following steps. Note that their order depends on the world definition and they can be thought of as layers of behaviours the system must enforce on player.

- Avoid alarm tiles
- Avoid pressing alarm-causing buttons
- Deactivate path-blocking alarm tiles by a button press
- Open closed-off rooms by pressing correct buttons
- Press a sequence of buttons to open the safe
- Reveal safe behind a picture by pressing correct button
- Take the trophy

After the safe was opened and possibly revealed from behind the picture player can pick up the trophy and therefore win the game.

Chapter 2

Dataset

Some teams participating in GIVE Challenge tried to learn either language expression or decision-making process from a human-human interactions in GIVE scenario. They were however relying on a small self-collected datasets. In a light of this, organizers of GIVE Challenge decided they would collect and provide dataset for future use. This chapter serves as an introduction and analysis of this dataset.

As a side note, Striegnitz et al. (2012) report on a smaller German dataset, which is similar to the one I will be talking about.

In the first section, I will introduce the dataset and provide technical details of how it was created. Second section, will, after the fashion of GIVE Challenge look how world and demographic factors influenced the task performance.

2.1 General overview

The data-collection started in July and finished November of 2012. Through that period 21 interactions between two human subjects were recorded. Originally, 22 pairs participated, but one of the pair failed to finish the tasks and is excluded from the dataset. The subjects were asked to bring someone they know and they were financially compensated for the effort.

The set-up for the experiment is in the figure 2.1. One human subject was an instruction giver (IG). He is on the right in the figure 2.1. His role was essentially the role of NLG system in GIVE Challenge. He was able to see a map of the world, which was updated real-time and informed of all necessary steps to finish the task. In addition, he was able to see the other's person client screen. He communicated with the other person through a microphone and his goal was to navigate the other person through the world and make him finish the treasure-hunt.

The other person was an instruction follower (IF). He is on the left in the figure 2.1. He interacted with the client and listened to IG's instructions through a headset.

Each pair did one short tutorial world and after that, they switched roles of instruction giver and instruction follower. Following was one world randomly chosen from

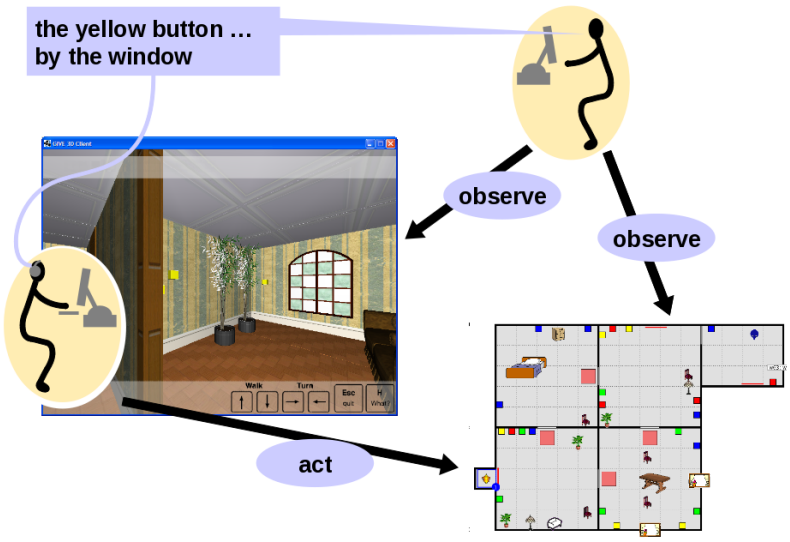

Figure 2.1: Experiment set-up of data collection

two variants (marked with 1 or 3). Finally they did a difficult version of the second variant (1-d or 3-d). A difficult version of the world had an increased number of distracting buttons and landmarks compared to the "normal" version. If not present in the report or not stated otherwise, the short tutorial worlds are normally excluded from the statistics.

Similarly to the GIVE Challenge, after all 3 rounds subjects were asked to fill a questionnaire. Its purpose was to get demographic information on subjects. Age, gender, gaming experience, previous cooperation between IG and IF and ability to navigate in the world were part of the questionnaire. To measure the ability to navigate in the world the Santa Barbara Sense of Direction (SBSOD) scores were used Hegarty et al. (2002).

As was mentioned in chapter 1, the entire session is logged to the database. The player's position, orientation and all visible objects are logged at fixed rate. Moreover, other information as buttons presses or an end of a session are also stored in the logs. Because the worlds are static, distances and angles between player and other game's objects are easily computed from there.

Apart from logs, there are of course sound files of IG giving directions. These were transcribed and together with the logs transformed into an ELAN files. ELAN is an annotation software (Sloetjes and Wittenburg, 2008). I will use a term automatic annotations for these files.

Building on top of these automatic annotations are manual annotations. They are primarily concerned with referring expressions and also stored in ELAN format. Most referring expressions in GIVE aim to locate a button, which needs to be pressed. I will call the button, which is a goal of an referring expression, the target button. Which button is the target button of a reference is among the manual annotations. Next layer of the annotations is some basic grouping of the references. Whether it is a reference to a single button, group of button, to a landmark and so on. Third

layer looks deeper into the contents of the reference. It notes whether the reference contains for example the color of a button, whether distractors or landmarks are part of the reference or whether the reference points out that a button was already pressed before.

Previously mentioned logs, automatic annotations and manual annotations together form the dataset this chapter is dealing with.

An example of what can be extracted from the data is in the following text. Spatial information are transcribed in parentheses for a sake of clearness.

(IF enters a room)
IG: Go towards the red buttons.
(IF turns right and start walking, but he turns too much)
IG: No the ones next to the lamp...
(IF corrects his direction)
IG: Yeah that lamp. On the right.
(IF is facing three buttons.)
IG: Press the button on the wall you are looking at, that's far from the lamp and on the left.
(IF goes towards the correct button and stops close to him)
IG: Press it.

2.2 World and Demographic factors

As was noted repeatedly in the chapter 1, the world had major influence on the task success rate in GIVE Challenge. However the dataset was created in a different manner, so the question about influence of worlds must be reformulated. First of all, since all the sessions were successful and the one, which wasn't, was discarded, the task success rate no longer makes sense. Instead, I will measure task performance by time required to finish the task (duration). Secondly, the normal worlds were designed to have the same or similar difficulty in order to minimize effects outside of navigation strategy. Same idea goes with the difficult worlds.

I found out that the normal worlds, 1 and 3, had a different mean duration (p-value 0.0473 for two-sample t-test). The difficult world did not have significant difference between their mean duration (p-value 0.6195 for two-sample t-test).

Another thing I was interested in was influence of gaming experience of both participants on the duration and also on the average speed of IF movement and the time IF spent moving. I found correlations between gaming experiences and these variables. Not surprisingly, these correlation are especially high for IF, since he/she is the one who is actually playing the world. The past gaming experience are more important than contemporary playing. Most prominent are the hours per week spent playing of IF at the past peak gaming period, the same variable for IG and hours spent gaming per week for IF at present. For the difficult worlds some correlations change slightly. In general, gamers take less time to finish the world, they spent more time moving and they have higher speed.

The influence of SBSOD scores on the task proficiency was another thing I have looked at. Correlation matrix revealed weak or almost no correlation between SBSOD scores and time needed to complete the world.

The data suggest that there is positive correlation between male gender and task proficiency measured in the duration. The effect of male IG diminishes in the difficult worlds but the effect of IF is even stronger in the difficult worlds. However there are several facts to take in consideration here. First of all, we don't have enough data to have a statistically significant conclusion. This correlation might have also been caused by having more male gamers than female gamers. Lastly, there has been research about influence of gender on spatial cognition and mental rotation; an example of more recent one is (Geary et al., 2000). They conclude that males are more proficient in tasks requiring mental rotation. Since IG have to do mental rotation while giving direction in GIVE scenario, this might be a source of correlation. Another paper worth considering on this topic is (Moffat et al., 1998), which found a difference in time required to finish a virtual maze.

The age of IF have positive correlation with task proficiency and in difficult worlds this correlation is one the strongest ones: 0.6. Older IF are also moving less and are generally slower. For IG the correlations have the same sign, however they are much weaker.

Lastly I was interested how familiarity of participants with each other influence the task efficiency. The knowledge of the partner had an impact on the task efficiency. What was interesting is that, previous cooperation with the partner became much more important for the difficult worlds, while not having much impact in normal worlds.

Conclusion

Bibliography

David C Geary, Scott J Saults, Fan Liu, and Mary K Hoard. Sex differences in spatial cognition, computational fluency, and arithmetical reasoning. *Journal of Experimental Child Psychology*, 77(4):337–353, 2000.

Mary Hegarty, Anthony E Richardson, Daniel R Montello, Kristin Lovelace, and Ilavanil Subbiah. Development of a self-report measure of environmental spatial ability. *Intelligence*, 30(5):425–447, 2002.

Alexander Koller, Kristina Striegnitz, Donna Byron, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander. The first challenge on generating instructions in virtual environments. In *Empirical Methods in Natural Language Generation*, pages 328–352. Springer, 2010a.

Alexander Koller, Kristina Striegnitz, Andrew Gargett, Donna Byron, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander. Report on the second nlg challenge on generating instructions in virtual environments (give-2). In *Proceedings of the 6th international natural language generation conference*, pages 243–250. Association for Computational Linguistics, 2010b.

Scott D Moffat, Elizabeth Hampson, and Maria Hatzipantelis. 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.

Han Sloetjes and Peter Wittenburg. Annotation by category: Elan and iso dcr. In *LREC*, 2008.

Kristina Striegnitz, Alexandre Denis, Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Mariët Theune. Report on the second second challenge on generating instructions in virtual environments (give-2.5). In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 270–279. Association for Computational Linguistics, 2011.

Kristina Striegnitz, Hendrik Buschmeier, and Stefan Kopp. Referring in installments: a corpus study of spoken object references in an interactive virtual environment. In *Proceedings of the Seventh International Natural Language Generation Conference*, pages 12–16. Association for Computational Linguistics, 2012.

Appendix