Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006 - 4th Workshop in Information and Human Language Technology (TIL'2006), Ribeirão Preto, Brazil, October 23–28, 2006. CD-ROM. ISBN 85-87837-11-7

A Web-Experiment on Dialogue Classification

Norton Trevisan Roman[1], Paul Piwek[2], and Ariadne Maria Brito Rizzoni Carvalho[1]

[1] Institute of Computing, Unicamp, Brazil.
{norton,ariadne}@ic.unicamp.br
[2] Centre for Research in Computing, The Open University, UK.
p.piwek@open.ac.uk

Abstract. This paper describes some technical details of a web experiment carried out at the State University of Campinas, Brazil, in order to classify a set of dialogues. We discuss the strategies we followed when designing and setting up the experiment, along with the statistical issues we addressed. The experimental results are not presented, for our intention is mainly to draw the community's attention to this new source of experimental data, as well as to describe our own experience with it.

1 Introduction

One of the greatest challenges for Natural Language Processing, and Artificial Intelligence itself, is the approximate reproduction of human behaviour by a computer system. The solution of this problem depends on the solution of yet another problem, namely, the problem of determining how human beings actually behave.

In order to overcome this difficulty, we need to carry out experiments with human beings. For that purpose, experimental psychology presents us with a very rich source of techniques and set-ups. Nevertheless, the obtained results may have their validity threatened by some problems, which come from the experimental methodology itself. These problems range from the presence of an experimenter in the laboratory to the lack of representativity of the selected set of participants, when compared to the target population.

An alternative to this methodology has emerged from the advent of the Internet. Since it became more popular, research has been carried out to determine how useful it might be for the execution of such experiments (cf. [Reips, 1996] [Frick et al., 1999,Gosling et al., 2004]).

Internet experiments have become a very interesting alternative to laboratory research, as they are able to successfully address many issues brought forth by the traditional set-ups. Nevertheless, they also gave birth to a whole new set of issues of their own, which should now be addressed. Such issues range from the possibility of multiple submissions to systematic drop-out rates [Reips, 2002b]. Experimenters who are willing to engage in creating a web-experiment will have to deal with these problems, in order to safeguard the validity of the obtained data.

In this paper we describe our experience with the execution of a web experiment, which was carried out at the State University of Campinas, Brazil, on September 17, 2004. Our main goal is to draw the attention of the community to this new medium for carrying out experiments with human beings. Here we only describe the experiment *per se*, not taking into account its results, hopefully helping new web-researchers in their task of planning a web-experiment[3].

The remainder of this paper is organised as follows. Section 2 describes the experimental set-up, presenting the addressed issues. Section 3 evaluates the usefulness of the adopted strategies to deal with such issues. Finally, Section 4 is the conclusion section of this paper.

2 Methodology

In this experiment, 16 dialogue scripts from movies made up the materials. The dialogues presented either a customer–vendor or a client–servant interaction. Our goal was to have participants classify each dialogue according to a set of categories.

To reduce drop-out and its negative impact, the experiment followed the Hard-Hurdle and Warm-Up techniques, as presented in [Reips, 1996,Reips, 2002a] [Reips, 2002b,Frick et al., 1999], as well as other techniques (*e.g.* prise drawing). Figure 1 illustrates the adopted design.

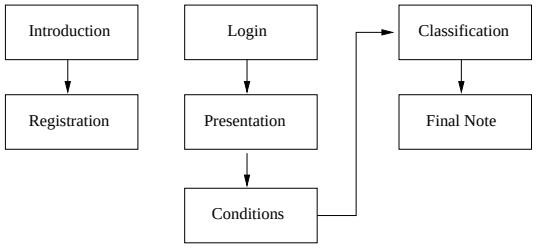

Fig. 1. Experimental Design

Following this design, participants first went through an introductory text explaining the experimental set-up (Introduction), but without giving away its real goal. The text also mentioned how long one might expect the experiment to last. In the next phase (Registration), participants registered for the experiment; they gave some personal information and received by e-mail a password that allowed them to proceed [Gosling et al., 2004].

Having the password, participants were able to move to the login phase, where they entered the experiment. After having logged in, participants had

[3] The reader will find a full description of the technicalities involved in the experiment in [Roman et al., 2004]

to go through yet another text giving some historical background to the research (Presentation). These measures where necessary to try to move the drop-out rates out of the experimental phase, that is, participants who might drop-out because they were initially in doubt about taking part in the study would be expected to make up their minds when going through these steps (Warm-Up and Hard-Hurdle techniques). Besides, it has been noticed that asking for personal information in the beginning of the experiment can reduce drop-out [Frick et al., 1999].

When participants finished the last text, they entered the real experiment. Firstly, the random distribution of participants amongst the experimental conditions took place (Conditions phase). After this phase, participants were provided with the experimental materials and were asked to classify the dialogues (Classification phase). Finally, when all the dialogues were classified, participants were presented with a final thank you note, as suggested in [Reips, 1996].

Participants were instructed to complete the experiment in one go. To guarantee this, we took some steps to prevent that once a participant had logged in, s/he could not leave and log in again. Other steps that we had to take were to prevent the pages, apart from the initial one and the login page, to be directly accessed. These pages had to be accessed in a very specific order.

Also, participants were not allowed to change their previous answer, *i.e.*, once they had classified a dialogue, they could not change it. Had they tried to do so, the system would have dropped them out. To avoid people being dropped out by the system due to their lack of knowledge on this fact, we stressed in all the instructional texts that they should not go back to previous pages and resubmit any information.

As a way of giving an incentive, we announced that any participant who successfully completed the experiment, following the instructions, would be eligible for one out of three prises of R$100,00. This measure was found to help reducing drop-out rates [Frick et al., 1999]. To avoid multiple submissions, we also stressed the fact that taking part in the experiment more than once would not increase their chances of winning the prise [Birnbaum, 2004a].

Finally, since we wanted to "save" some participants from our total population for a second, more important and future experiment, we limited the amount of participants for this experiment to 30. The script stopped accepting participants as soon as it detected that more than 30 had already finished the experiment. Nevertheless, by the time the script stopped accepting new participants, there might still be some going through the experiment. That is the reason why we ended up with more than 30 participants.

2.1 Participants

Our participants are graduate and postgraduate students from the State University of Campinas, Brazil. We chose to consider only these students because it meant that (1) it was easy to control for false identity issues, (2) there were no cross-cultural issues (cf. [Reips, 1996,Lang, 2002]), for almost all of them were Brazilians, (3) they were quite well distributed according to gender, (4) we could

find people from almost all Brazilian states in the university, and (5) the big set of possible participants available – in 2003, according to the university website[4] there were 13,777 graduate, 4,563 MPhil and 4,779 DPhil students, being 20,165 students in total.

The main drawback about this participants pool is that a set of university students may not be representative of the whole Brazilian population. Therefore, all of our claims are restricted to the university's subset, i.e., to the population with a high education, predominantly young, and mostly from the state of São Paulo.

In order to reach the participants, we broadcasted an e-mail to all university students, asking them to take part in the experiment and providing the address of the experiment's website. We only mentioned that we were doing an experiment on dialogues, that it was a voluntary work, that it would not take too long, and that those who finished the experiment would be eligible for a cash prise draw.

2.2 Variables

In this experiment, we had no between-subjects variable, for we did not separate the participants in different groups. The only within-subjects variable, *i.e.*, the variable affecting all the participants, is the order of the dialogues. To avoid any confounding about the dialogues' classification and their order of appearance, *i.e.*, to avoid order effects [Reips, 2002b], when a participant started the experiment, a dialogue order was randomly generated and assigned to her/him.

The dependent variable, *i.e.*, the variable to be measured, was the classification of the dialogues, according to a set of categories, by the participants.

2.3 Addressed Issues

In the experimental design, we addressed the following threats to the internal validity of the experiment.

- **History and Maturation** To minimise the effects of outside events and the passage of time *per se*, the participants had to complete the experiment in one go.
- **Instrumentation** To avoid the participants improving over repeated classifications, they were asked to classify the dialogues only once. Unfortunately, as they had to classify 16 dialogues, some of them might have experienced some improvement in the course of the experiment. In this case, the randomisation of the dialogue order should randomly distribute those participants who were in such a situation amongst the dialogue orders, distributing also the error across the dialogues.

[4] *http* : *//www.aeplan.unicamp.br/anuario_estatistico_*2004/*arquivos_html/*4_2_3_%20*grad_graf_geral.html* and *http* : *//www.aeplan.unicamp.br/anuario_estatistico_*2004/*arquivos_html/*4_3_9_*pos_graficos_*93_03.*html*

- **Experimenter Bias** To avoid any unconscious bias from the experimenters, interactions with participants were kept at a technical level only, answering only technical questions. The fact that there is actually no experimenter physically present during the web experiment also reduces this source of error (cf. [Reips, 1996,Reips, 2002b,Birnbaum, 2004b,Hewson, 2003]).
- **Generalisation** Threats to the generalisation of results range from intercultural issues to the dependency on computers. In order to determine intercultural issues, participants were asked to tell us which Brazilian State they came from. They can also choose the option "foreigner". This allowed us to direct the research towards the Brazilian population only.
  The fact that we had students for participants reduces the problem with the dependency on computers and networks, which could limit generalisation (cf. [Reips, 1996,Reips, 2002b,Mueller et al., 2000,Hewson, 2003]), since the university provides the students with enough computers, as well as a quick network connection. Nevertheless, the set of participants is still very limited according to some aspects, such as age and state of origin.
- **Misunderstandings** To reduce the effect of the participants misunderstanding instructions, due to a lack of clarity, we tested the system with some volunteers in the university, before releasing it, as suggested in [Reips, 2002b] [Birnbaum, 2004b,Birnbaum, 2004a].
- **Technical Variance** People using different hardware and software could contribute to the error variance, although this error variance would be randomly distributed [Reips, 1996,Reips, 2002a]. In order to reduce browser compatibility problems, we used only server-side programming, which is less susceptible to platform-dependent issues [Reips, 2002b,Birnbaum, 2004b], while delivering pure html code, without requiring any further plugins or other non-standard technologies. Also, we used the default font types and sizes defined in the user's browser, expecting that the participants had actually set them up to better suit their needs.
- **Multiple Submissions** In a web experiment, there is always the possibility of a participant taking part in the experience more than once. To avoid multiple submissions, the participants had to provide some personal details [Reips, 2002b,Birnbaum, 2004b], like the e-mail address and university id number. The script, then, did not allow the same id to be registered more than once.
- **Selection** Participants truly volunteered to take part in the experiment. The fact that the call for participation was done through an e-mail broadcast to the entire university guaranteed impersonality, meaning that all data came from self-selected participants, instead of having someone participate in the experiment due to any kind of pressure. This measure was found to help increasing the completeness of the answers [Gosling et al., 2004], even though it might affect generalisation, for the self-selected sample might not generalise to the target population [Birnbaum, 2004b].
- **Form Bias** According to Reips [Reips, 2002b], it is potentially biasing to use scripts that do not allow the participants to leave any items unanswered. This might create a confounding by which there is no way of telling whether

a participant deliberately refused to answer some question or just forgot it. In our experiment, this issue was addressed by always having a default neutral answer pre-selected in every single form, along with a "I don't want to answer" choice.

- **Ecological Validity** The fact that it is a web-experiment, instead of a laboratory one, increases its ecological validity, for participants were in their usual surrounding, reducing the effects of being in an unfamiliar setting [Reips, 1996,Reips, 2002b]. Also, the participants should not worry about when and where to go to take part in the experiment, picking the moment and place that suit them better, which is highly convenient for both participant and experimenter [Birnbaum, 2004a].
- **Drop-out** In web experiments, participants are free to drop out, meaning that those who finish the experiment are the ones who were actually willing to finish it. This increases the reliability of data, for those participants that might have stayed because they felt compelled to [Frick et al., 1999] most probably had dropped out.
  On the other hand, a considerable drop-out rate may raise serious concerns about the validity of the data, mainly if participants drop out in a selective way. In that case, the explanatory power of the experiment could be compromised, if drop-out varies systematically amongst the experimental conditions [Reips, 2002b].
  One measure implemented to reduce drop-out was a financial incentive [Reips, 1996,Frick et al., 1999,Reips, 2002b], through which every participant who finished the experiment entered in a prise draw. Another one was to ask personal information at the beginning of the experiment.
  The information load was also reduced, stage after stage [Frick et al., 1999]. Thus, the participants go through heavy pages, with a lot of text in the beginning of the experiment and, when the classification stage starts, the pages are much lighter, presenting only the dialogue, the classification form, and a very small set of instructions.
  In addition, the hard-hurdle and warm-up techniques [Reips, 2002b] were used to reduce the negative impact of drop-out. In a nutshell, hard-hurdle will try to tell the participant how difficult the experiment might be, while the warm-up will try to discourage the participant which is more susceptible to drop out.
  Our final measure against drop-out was to make the web pages as simple as possible [Reips, 2002b], using plain html and server-side programming only. That way, virtually any browser could show the experiment without any problem.

3 Results and Analysis

Figure 2 shows the drop-out numbers for this experiment. The bars show the number of participants that visited a specific page and chose to go further. In our experiment, the warm-up and hard-hurdle techniques are represented by the bars from the "Pres" page to the "Log" page in the Figure.

The actual experiment started in "His". At this moment a dialogue sequence was assigned to the participant, and the first dialogue in that sequence was shown.

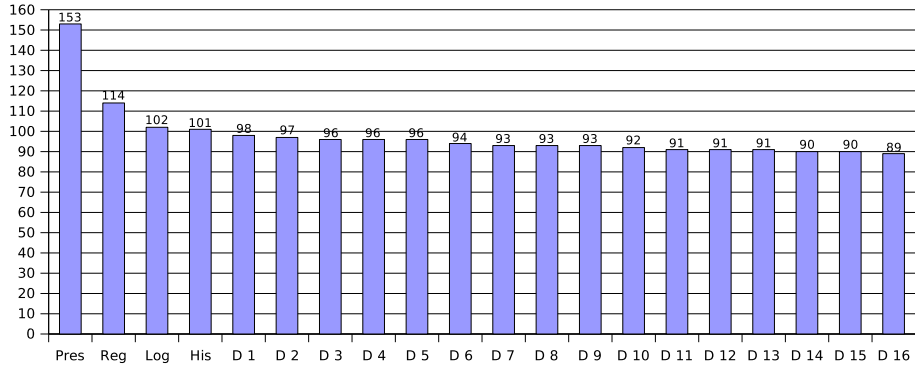

Fig. 2. Drop-out levels

The Figure shows what Reips called the "natural dropout curve" [Reips, 2002b]. In our case, it was a well defined exponential curve in the number of participants who decided to move from one page to the next one. From the "Pres" page to the "Log" page we lost 52 participants, for 153 registered and 101 saw the first dialogue. In the experimental phase, on the other hand, we lost only 12 participants, considerably less. This was in conformity with the initial expectancy that the drop-out of people checking in just out of curiosity would happen before the second, decisive, phase began [Reips, 1996], suggesting that our measures for avoiding drop-out actually worked.

Figures 3 to 5 compare the number of participants that finished the experiment and those who dropped out, according to gender, knowledge area, educational level, age and Brazilian Region where they came from. As we have these data only for those participants who actually provided them, the total amount of participants to be considered is 114, *i.e.*, the ones in the "Reg" column in Figure 2.

These figures show no proportionally great difference between those participants who finished the experiment and those who dropped out. According to the gender (Figure 3), 70% of the participants who finished the experiment were men, while 30% were women. From those who dropped out 68% were men and 32% women.

The educational level (Figure 3) also presents very little difference, with 48% of those who finished the experiment being undergraduate and 52% post-graduate students. From the dropped out graph, we see 52% of undergraduates and 48% of post-graduates.

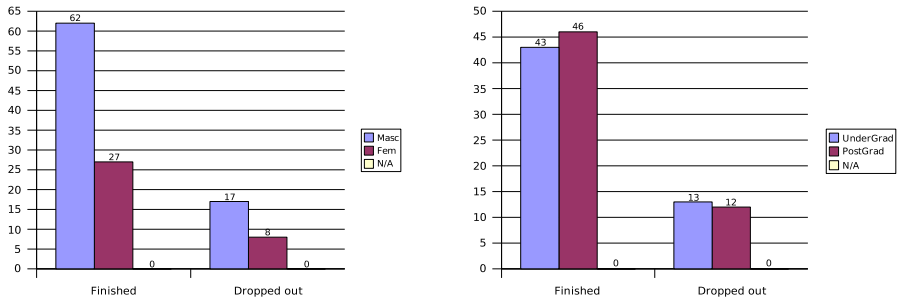

Fig. 3. Number of participants, according to gender and educational level.

According to the knowledge area (Figure 4), 79% were from "exact sciences", 12% from "human sciences" and 9% from "biological sciences". As for drop-out, the numbers are 76%, 20% and 4%, respectively. There is a difference between the two last categories (human and biological sciences) in each group. However, a χ^2 test for the non-zero categories revealed that this difference was not significant at all ($\chi^2(1, 114) = 0,0074$, at $p = 93.14\%$).

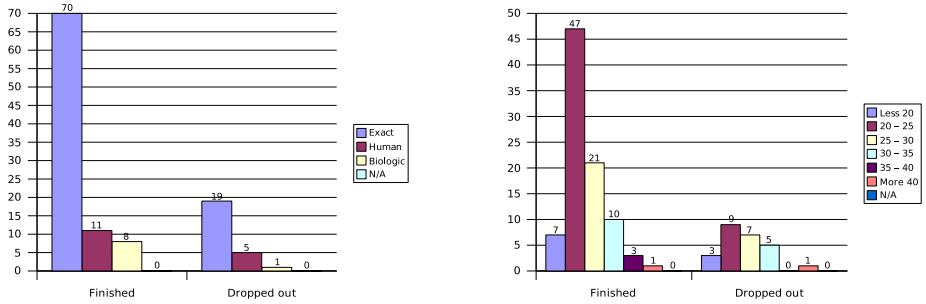

Fig. 4. Number of participants, according to the knowledge area and age.

These figures suggest that our measures for avoiding systematic drop-out, such as the random distribution of the participants in the experimental conditions (the dialogue order), might have worked. The possible exceptions could be age (Figure 4) and Brazilian Region of origin (Figure 5).

Age presents some discrepancies mainly in the 20 to 25 and 30 to 35 years old ranges. From those participants who finished the experiment, 53% were in the 20 – 25 range, while 11% were in the 30 – 35. From those who dropped out, these percentages are, respectively, 36% and 20%. Despite this difference, the numbers are too small to allow any conclusion.

This phenomenon seems to happen in the Region graph as well (Figure 5). We see differences in the sets with participants coming from the South Region

and Foreigners. Again, the number of participants in these groups and those who dropped out is too small to draw any conclusion.

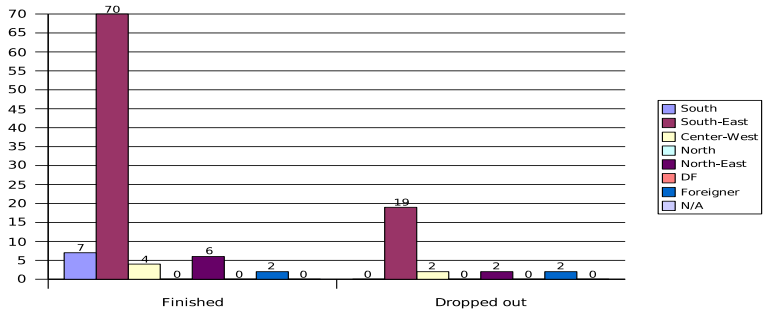

Fig. 5. Number of participants, according to the Region of origin.

Another important point in this experiment was the distribution of students amongst the possible dialogue sequences. As we had 16 dialogues, there were $16! = 20.92 \times 10^{12}$ possible conditions. A quick look at the distributed conditions showed 101 different sequences – the exact number of participants who took part in the "real" experiment. That means we actually had no duplicate sequence, all were different. This certainly helped to minimise any consequence that a particular sequence might have had.

4 Conclusion

In this paper we described the technical details of a web experiment. The experiment was designed so that participants go through 16 different dialogues, classifying each of them according to a list of categories.

To deal with classic experimental issues, as well as new ones, brought on by the use of the Internet, we followed a number of strategies from the literature (see the included bibliography). The strategies described in this paper seemed to work very well. Our experiment developed exactly as predicted. More specifically, drop-out rates were nicely moved to a zone external to the real experiment. Participant distribution was approximately the same for those who finished the experiment and those who dropped out, meaning that no bias was introduced due to drop-out.

Another point worthy of mention is the fact that the experiment ran for less than one hour, from the call for participation. That means, in one hour, we already had more than 30 participants who finished the experiment. In the subsequent hours, this number increased to 89, which was a huge surprise. For we were hoping to have 30 participants finish the experiment in one week, and suddenly, in one single hour, we had much more than that.

Also, the experiment was almost cost free. With the exception of the prises we paid, we had no additional cost, for everything was run using the university's infrastructure. Not to mention the fact that all data was collected and analysed automatically by scripts, saving a lot of the experimenters' time. All in all, our experience with the web as a medium for carrying out research with human beings was very satisfactory.

Finally, it is mostly advisable that someone willing to design a web experiment follows the tips in the bibliography we present here. We also hope that our testimony can encourage researchers to explore this new medium, gathering more information about its possibilities and limitations.

Acknowledgements

We thank all the participants for their volunteer work.

References

[Birnbaum, 2004a] Birnbaum, M. H. (2004a). *Handbook of Methods in Social Psychology*, chapter Methodological and Ethical Issues in Conducting Social Psychology Research Via the Internet, pages 359–382. Sage.

[Birnbaum, 2004b] Birnbaum, M. H. (2004b). Human research and data collection via the internet. *Annual Review of Psychology*, 55:803–832.

[Frick et al., 1999] Frick, A., Bächtiger, M.-T., and Reips, U.-D. (1999). Financial incentives, personal information and drop-out rate in online studies. In Reips, U.-D., Batinic, B., Bandilla, W., Bosnjak, M., Grf, L., Moser, K., and Werner, A., editors, *Current Internet Science - Trends, Techniques, Results*. Online Press, Zurich.

[Gosling et al., 2004] Gosling, S. D., Vazire, S., Srivastava, S., and John, O. P. (2004). Should we trust web-based studies? a comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, 59(2):93–104.

[Hewson, 2003] Hewson, C. (2003). Conducting research on the internet. *The Psychologist*, 16(6):290–293.

[Lang, 2002] Lang, M. (2002). The use of web-based international surveys in information systems research. In *Proceedings of European Conference on Research Methodology for Business and Management Studies (ECRM – 2002)*, pages 187–196, Reading, England.

[Mueller et al., 2000] Mueller, J. H., Jacobsen, D. M., and Schwarzer, R. (2000). *Psychological Experiments on the Internet*, chapter What Are Computing Experiences Good For: A Case Study in On-Line Research. Academic Press, San Diego, USA.

[Reips, 1996] Reips, U.-D. (1996). Experimenting in the world wide web. In *Proceedings of the 26th Society for Computers in Psychology Conference (SCiP – 96)*, Chicago, USA.

[Reips, 2002a] Reips, U.-D. (2002a). Internet-based psychological experimenting: Five dos and five don'ts. *Social Science Computer Review*, 20(3):241–249.

[Reips, 2002b] Reips, U.-D. (2002b). Standards for internet-based experimenting. *Experimental Psychology*, 49(4):243–256.

[Roman et al., 2004] Roman, N. T., Piwek, P., and Carvalho, A. M. B. R. (2004). A web-based experiment on dialogue classification. Technical Report ITRI-04-15, Information Technology Research Institute – University of Brighton, Brighton, UK.