Institut
Mines-Télécom

TRINITY
COLLEGE
DUBLIN

Managing Consistency in Wizard of Oz Studies

A Challenge of Prototyping Natural Language Interactions

Stephan Schlögl, Gavin Doherty, & Saturnino Luz

What is Wizard of Oz?

- Wizard of Oz is a prototyping method in which a **human 'wizard' mimics the actions of a system**
- It has its roots in **Natural Language Processing** (cf. Gould et al., 1983)
- It allows researchers and designers to **evaluate potential user experiences without building a fully working system first**

(Buxton, 2007)

Why is Wizard of Oz Relevant for NLP?

- As with graphical user interfaces early and iterative evaluation is **important to cater for high quality software**
- **Technical aspects:**
  Support dialog design, collect language/interaction corpora and test language technology components
- **Design aspects:**
  Explore usability and user experience

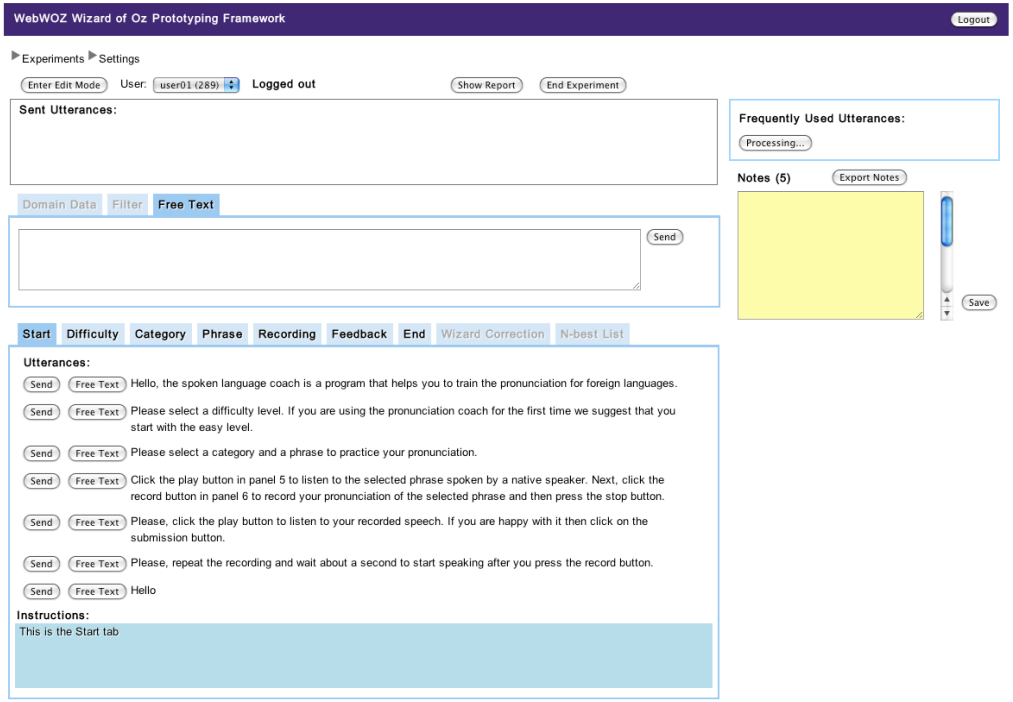
The Challenge of the Wizard Task?

- Follow a defined test protocol (as close as possible…)
- Deal with stress
- Be prepared for the unexpected

- Make **test participants believe that they are interacting with a real system**
  - Act **consistently**
  - Act **predictable**
  - Act **realistic**

Overall Task

- **Simulate text-based natural language system responses using WebWOZ[1]**



[1]https://github.com/stephanschloegl/WebWOZ

Study A

- Interactive system that helps customers to choose an appropriate Internet Connection bundle
- **11 test participants**
- **1 wizard**
- Participant was able to speak to the system in German
- System (=wizard) answers using a set of pre-defined, pre-translated utterances
- **Utterances are displayed on the screen**
- **Further info:** Schneider et al. 2010

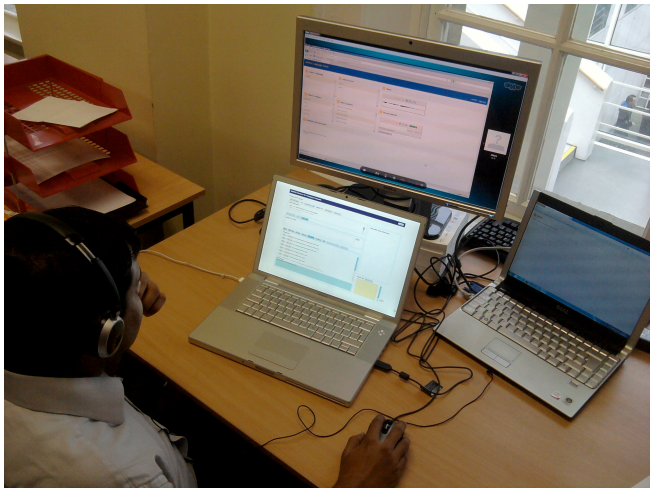
Study B

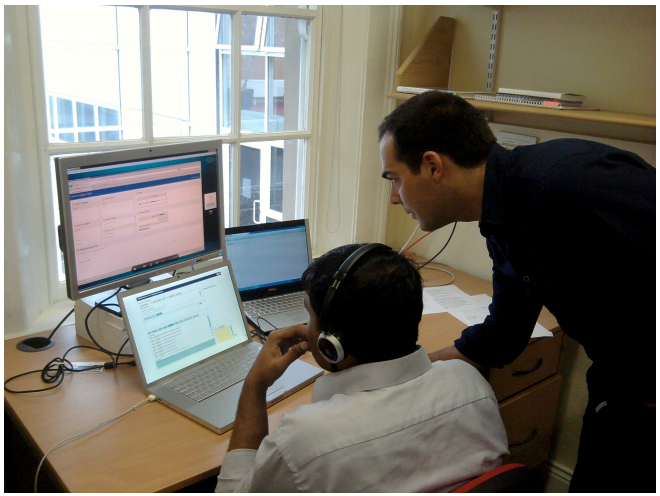
- Extending Study A into the spoken language domain
- **17 test participants**
- **1 wizard**
- Participant was able to speak to the system in German
- System (=wizard) answers using pre-defined, pre-translated as well as pre-recorded utterances
- **2 modes:**
  - Mode 1: Utterances are displayed on the screen
  - Mode 2: Recorded utterances are played
- **Further info:** Schneider 2013

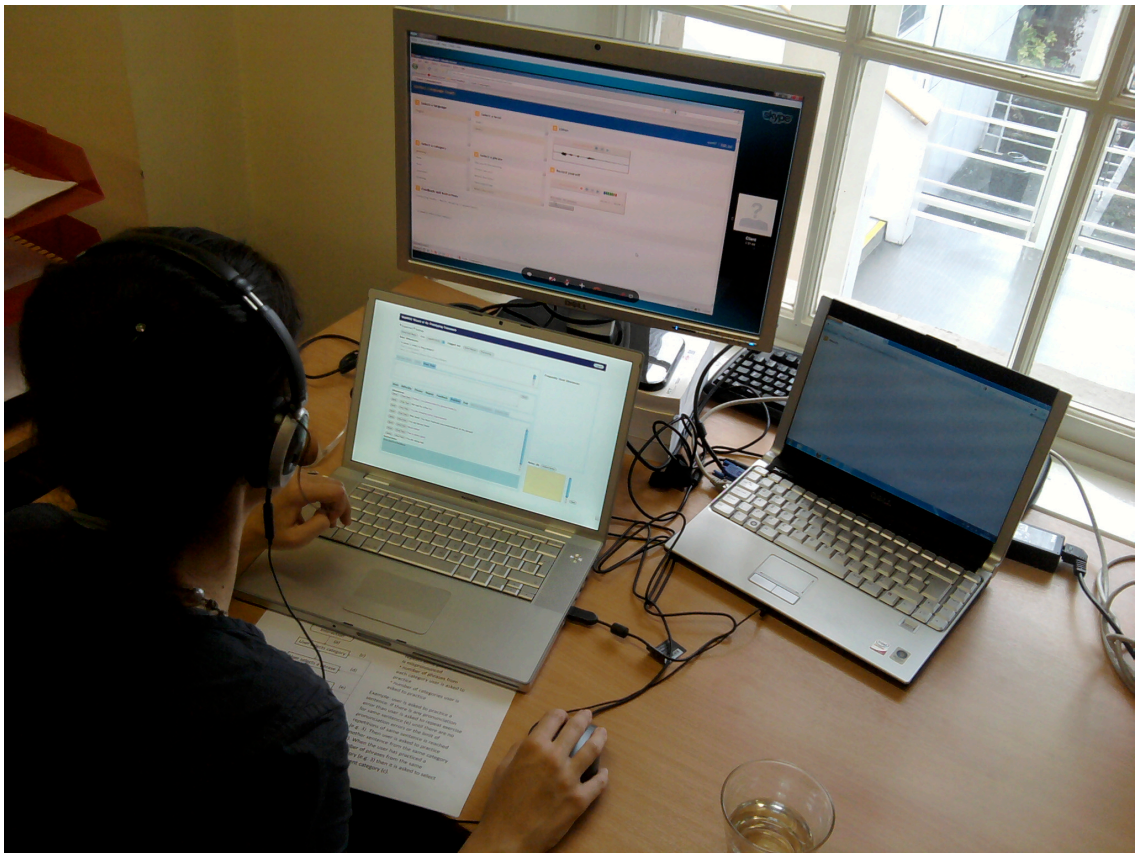
Study C

- WOZ used with online language pronunciation trainer
- **13 test participants**
- **1 wizard**
- Participant was training her/his pronunciation of predefined English sentences
- System (=wizard) was giving textual feedback based on the evaluation results
- **Further info:**
  Cabral et al., 2012a/b

Some Pictures







Further info:
Cabral et al. 2012a/b

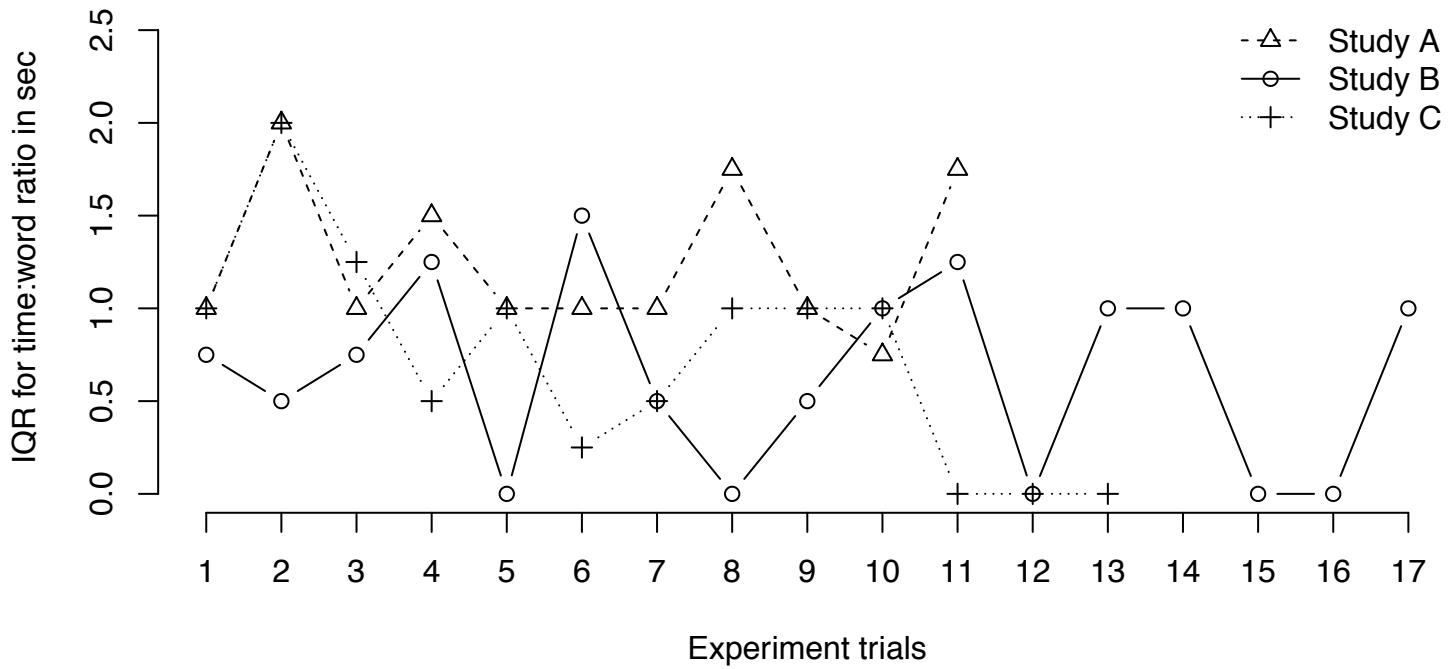
Summary

- **3 wizard studies**
  - 1 wizard interacting with **11 participants (Study A)**
  - 1 wizard interacting with **17 participants (Study B)**
  - 1 wizard interacting with **13 participants (Study C)**
- Wizards **select/generate text utterances** to be sent to participants
- **1 challenge:**

  Estimate a participants reading speed in cases where a follow-up utterance needs to be sent
- **Meta Analysis:** Wizard consistency

Results Meta Analysis

- **Wizards have problems estimating reading speed consistently**



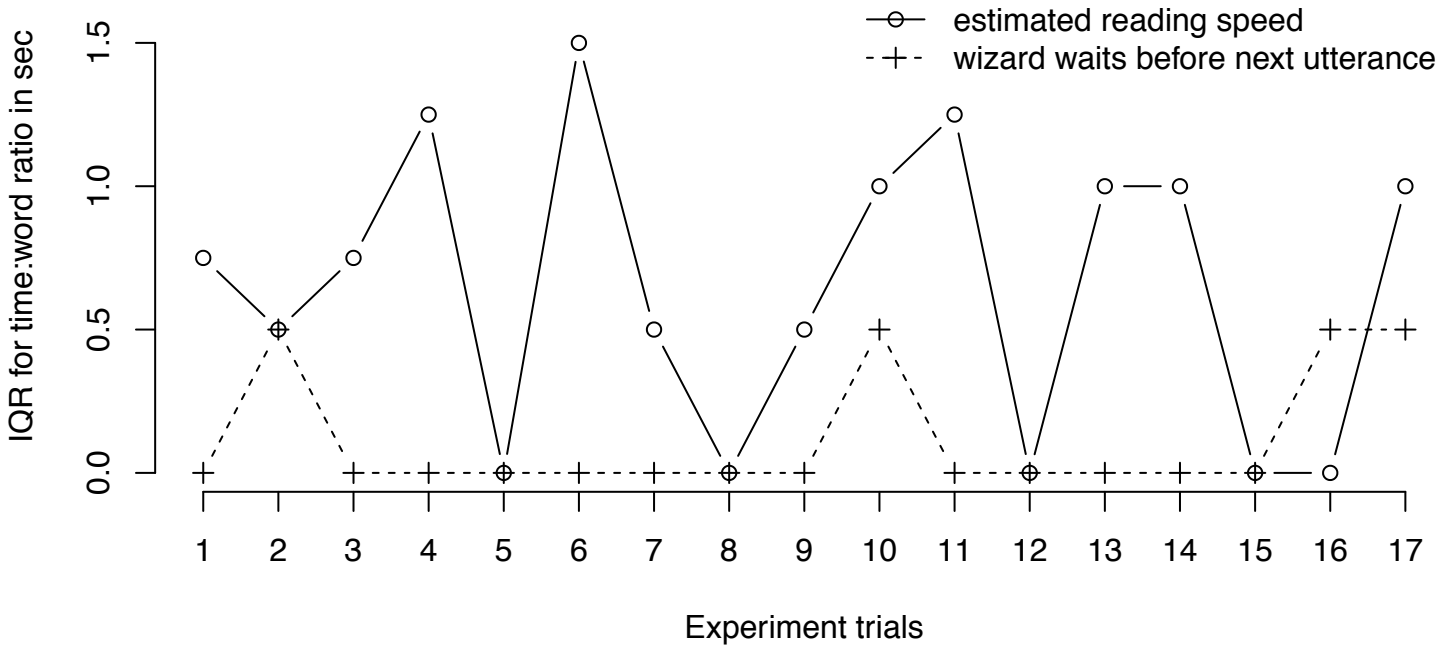
Interquartile Range (IQR) Values for Text-based Interaction; Comparison of studies A, B and C.

Critical Reflection

- Wizards were **not actively asked** to estimate a participant's reading speed
- Studies were **not specifically designed** for this analysis
- Wizards **might not have had sufficient training**

- We report on a **wizards' natural actions**
- One wizard highlighted that she was **reading utterances in her mind**
- A study showed that the time spent on wizard training is often **less than 30 minutes and in our case no improvements over the course of several sessions were noticed**

Comparison Text vs. Speech in Study B



Interquartile Range (IQR) Values for Text-based Interaction vs. Speech-based Interaction Study B.

One-tailed paired Student's t-test: t(16)= 3.9105, **p=0.0006**

Discussion

- WOZ is a **valuable prototyping method** but its **dependency on a human wizard** makes it susceptible for errors
- **Inconsistent** wizard **behavior may bias study results** (Note: While in the discussed experiments inconsistencies did not lead to significantly reduced user satisfaction ratings, such might be a problem when it comes to stricter experimental settings)
- **Additional support for wizards** (e.g. through additional contextual information or timing functionalities) can improve the validity of the method

Acknowledgments

References

- Buxton, B. (2007). Sketching User Experiences. Morgan Kaufman.
- Cabral, J. P., Kane, M., Ahmed, Z., Abou-Zleikha, M., Székely, É., Zahra, A., Ogbureke, K. U., et al. (2012a). Rapidly Testing the Interaction Model of a Pronunciation Training System via Wizard-of-Oz. Proceedings of LREC. Istanbul, Turkey.
- Cabral, J. P., Kane, M., Ahmed, Z., Székely, É., Zahra, A., Ogbureke, K. U., Cahill, P., et al. (2012b). Using the Wizard-of-Oz Framework in a Pronunciation Training System for Providing User Feedback and Instructions. Proceedings of IS ADEPT. Stockholm, Sweden.
- Gould, J. D., Conti, J., & Hovanyecz, T. (1983). Composing letters with a simulated listening typewriter. Communications of the ACM, 26(4), 295-308.
- Schlögl, S., Doherty, G., Luz, S., & Karamanis, N. (2010a). WebWOZ: A Wizard of Oz Prototyping Framework. Proceedings of ACM EICS (pp. 109-114). Berlin, Germany.
- Schlögl, S., Doherty, G., Karamanis, N., Schneider, A. H., & Luz, S. (2010b). Observing the Wizard: In Search of a generic Interface for Wizard of Oz Studies. Proceedings of Irish HCI (pp. 43-50). Dublin, Ireland.
- Schlögl, S., Schneider, A. H., Luz, S., & Doherty, G. (2011). Supporting the Wizard: Interface Improvements in Wizard of Oz Studies. Proceedings of BSC HCI. Newcastle Upon Tyne, UK.
- Schneider, A. H., Sluis, I. V. D., & Luz, S. (2010). Comparing Intrinsic and Extrinsic Evaluation of MT Output in a Dialogue System. Proceedings of IWSLT (pp. 329-336). Paris, France.
- Schneider, A. H. (2013). Intrinsic and Extrinsic Component Evaluation in Interactive Multilingual Speech Applications. Thesis: Trinity College, University of Dublin.