Online experiments with the Percy software framework – experiences and some early results

Christoph Draxler
BAS Bavarian Archive of Speech Signals
Institute of Phonetics and Speech Processing
Ludwig-Maximilian University Munich, Germany
draxler@phonetik.uni-muenchen.de

Abstract

In early 2012 the online perception experiment software Percy was deployed on a production server at our lab. Since then, 38 experiments have been made publicly available, with a total of 3078 experiment sessions. In the course of time, the software has been continuously updated and extended to adapt to changing user requirements. Web-based editors for the structure and layout of the experiments have been developed. This paper describes the system architecture, presents usage statistics, discusses typical characteristics of online experiments, and gives an outlook on ongoing work. `webapp.phonetik.uni-muenchen.de/WebExperiment` lists all currently active experiments.

Keywords: online perception experiment, tool, WWW, results

1. Introduction

An important part of empirical speech and language research is setting up and running perception experiments. In these experiments, participants are asked for their subjective judgment of specific speech or language features, e.g. the acceptability of a given construct, a categorization of sounds, the perceived regional accent, etc. The audience may be experts in the field, or, and increasingly so, laymen, e.g. in perceptual dialectology (for an introduction see e.g. (Anders et al., 2010), (Anders, 2010)).

Traditionally, perception experiments are run on a computer, either using dedicated hardware and software or via the web in a standard browser. Until the advent of HTML5, web-based experiments were severely restricted due to technical limitations: audio and video playback required dedicated plug-ins and were thus highly platform dependent. With HTML5, these limitations no longer exist, and because mobile devices have become so powerful, it is now possible to run perception experiments not only on traditional computers, but also on mobile devices and even TV sets, allowing researchers to reach new target audiences.

(Reips, 2002a) compares traditional, lab-based experiments with online experiments via the web. He discusses the advantages and disadvantages of both approaches and gives recommendations on how to successfully set up and run experiments via the web. Many of these recommendations also hold for perception experiments, but there are some issues specific to this kind of experiments.

(Reips, 2002b) describes an early web-based tool for the design of web experiments, however (due to the state of technology in 2002) without support for audio or video. LimeSurvey (`www.limesurvey.org`) and UniPark (`www.unipark.info`) are examples of online systems for *surveys*; however, they are general-purpose tools and thus do not provide features required by perception experiments, e.g. limit the number of replay repetitions or measure reaction time. WebExp (Keller et al., 2009) is one of the earliest systems for online perception experiments. (Reips and Lengler, 2005) describe a portal for web experiments; this portal is used to recruit participants and to allow researchers to search for sample experiments in different domains. (Lefever et al., 2007) contains relevant information on response rates, and participant age and sex distribution for online experiments.

2. System Architecture

The online perception experiment system Percy is implemented as a client-server system (Draxler, 2011). The server is responsible for session and data managment, the client provides the user interface for the experiment, usually within a browser.

2.1. Server side

All data is held in a relational database system. The data model consists of two parts: a hierarchy of relations defining the content of the experiments, and a second hierarchy containing the participants and their inputs (1). A *project* is an administrative frame for experiments. An experiment may contain one or more *scripts*, each script consists of one or more sequential *sections*, and a section contains one or more experiment *items*, which are presented in sequential or random order. An experiment *input* is part of an experiment *session*.

The two hierarchies are linked via two relations: each experiment input is linked to an experiment item, and an experiment session links a particular experiment script with an experiment session and a participant.

Percy distinguishes four user roles: project and experiment administrators, script editors and participants.

Currently, the database is held in a PostgreSQL database system (v. 8.2.3) which has shown to be highly performant and extremely stable. The server is a standard Apache Tomcat server (v. 6.0.32).

2.2. Client side

The client loads the experiment HTML file and renders it in the browser. Depending on the state of the experiment, the screen layout changes: first, an introductory screen is shown to give the user some general information about the experiment. Then, a data form is displayed which requests

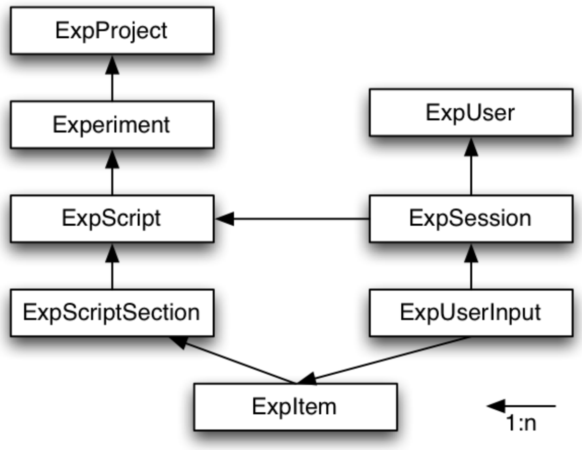


Figure 1: Data model of the Percy relational database

the necessary information about the participant. Note that this form contains participant data fields as well as session fields which are used to collect information about the environment in which the experiment is performed, the audio equipment used, the device type and the browser.

Upon submitting the form, the data is sent to the server and the entire script for the current experiment session is loaded. Note that the server may select different scripts for a given experiment, depending on the chosen selection criteria. Such selection criteria may be to use the least frequently used script, the newest script, a randomly chosen script, etc.

Once the script is loaded, one item after the other is presented to the participant. Each input is immediately transferred to the server so that no input data is lost, even if a participant decides to abort the experiment.

2.2.1. Experiment screen designs

The layout of the experiment pages is specified using JavaScript with CSS. The default design consists of an interactive icon for every audio stimulus and up to two rows of regular buttons labelled with the input options defined in the experiment script. The first row contains the input options proper, the second row contains self-assessment options (see Fig. 2 a) for a default layout with one row of input buttons).

By dynamically generating the screen layout, different designs can be implemented by a programmer. These designs may present several input options to the participants, or contain additional GUI elements, e.g. maps. Fig. 2 shows a selection of screens used in recent experiments.

The icons reflect the state of the stimuli: active when ready for playback, disabled when playing, and marked (usually with a red 'X') if the maximum number of playback repetitions has been reached.

2.2.2. Input processing

HTML5 provides a number of input elements: regular and radio buttons, checkboxes, menus, single and multi-line text fields, and sliders. The dynamic creation of the screen design allows any of these elements to be used in a screen. By programming the GUI, user input can be verified before data is sent to the server. Thus, the consistency of user input can easily be checked, e.g. accept input only if the audio was actually played.

2.3. Experiment Editors

Perception experiments come in different designs and sizes, and the technical implementation of an experiments consists of the definition of the experiment content, and the design of the experiment screens.

An experienced programmer or experiment administrator may enter the experiment contents directly into the database on the server. However, for researchers who simply want to perform an experiment, a graphical experiment editor is easier to use. For Percy, two experiment content editors have been developed by students of the media informatics lab at LMU Munich. The editors take different approaches to the task: one editor is a highly interactive graphical editor, the other takes an explorer-like approach (see Fig. 3 for screen shots of the editors).

Currently, an editor for the screen design of the experiment is being developed.

3. Usage statistics

From Jan. 1st 2012 through Feb. 28th 2014, a total of 38 experiments with 3078 sessions have been run using Percy. 19 of these experiments ran for less than three months, 10 ran less than one year, and 9 very long-term or even permanent experiments have been running for more than one year. 4 experiments were set up in the context of university courses, e.g. a seminar on online experiments, 4 experiments were set up for BA and MA theses, and 24 for research at PhD, post-doc, or project level (the remaining experiments were pilot tests or demos). 31 experiments were set up within the phonetics lab, 7 by external researchers.

Experiments have been carried out in German, English, Spanish, and Estonian.

4. Results

4.1. Global results

4.1.1. Distribution of participant sex

The distribution of participant sex is 1983 female (61.9%) vs. 1145 male (37.3%); 26 (0.8%) participants did not indicate their sex[1].

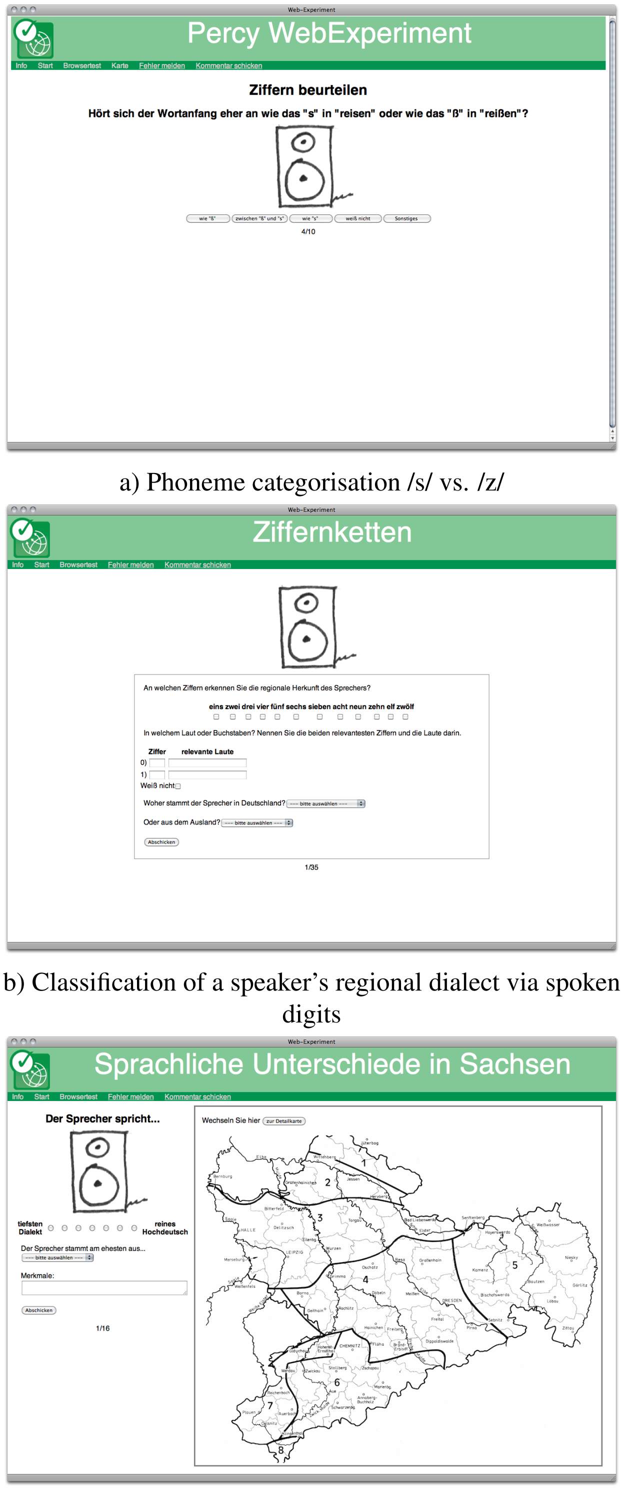
This approx. 2:1 ratio of female vs. male participants is quite common for online experiments, especially when they address an academic audience (see e.g. (Lefever et al., 2007) for similar results).

4.1.2. Distribution of participant age

In general, participants were asked to indicate their age at the time of participation. This input field was checked for consistency (i.e. empty input or values outside the range of 0..150 were not accepted). Table 1 shows the age and sex distribution of the participants.

Clearly, in academic environments, most participants are students.

[1] At some point in time, the participant data input form was changed to make entering the participant sex mandatory.


Web-Experiment
Percy WebExperiment
Info Start Browsertest Karte Fehler melden Kommentar schicken

Ziffern beurteilen

Hört sich der Wortanfang eher an wie das "s" in "reisen" oder wie das "ß" in "reißen"?

wie "ß" | zwischen "ß" und "s" | wie "s" | weiß nicht | Sonstiges
4/10

a) Phoneme categorisation /s/ vs. /z/

Web-Experiment
Ziffernketten
Info Start Browsertest Fehler melden Kommentar schicken

An welchen Ziffern erkennen Sie die regionale Herkunft des Sprechers?

eins zwei drei vier fünf sechs sieben acht neun zehn elf zwölf

□ □ □ □ □ □ □ □ □ □ □ □

In welchem Laut oder Buchstaben? Nennen Sie die beiden relevantesten Ziffern und die Laute darin.

Ziffer	relevante Laute
0)	
1)	

Weiß nicht □

Woher stammt der Sprecher in Deutschland? --- bitte auswählen ---

Oder aus dem Ausland? --- bitte auswählen ---

b) Classification of a speaker's regional dialect via spoken digits

Web-Experiment
Sprachliche Unterschiede in Sachsen
Info Start Browsertest Fehler melden Kommentar schicken

Der Sprecher spricht...

tiefsten Dialekt ○ ○ ○ ○ ○ ○ ○ reines Hochdeutsch

Der Sprecher stammt am ehesten aus... --- bitte auswählen ---

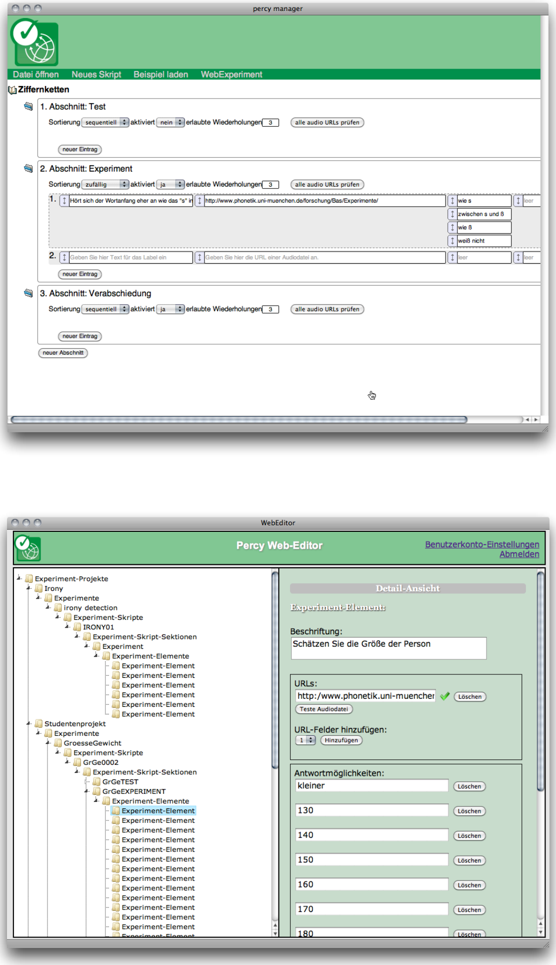
Merkmale:

c) Indicate the regional background of Saxonian speakers on a map

Figure 2: Screen designs for different acoustical perception experiments

sex	10-19	20-29	30-39	40-49	50-59	60+
f	156	1230	311	68	67	39
m	56	547	267	97	133	30

Table 1: Sex and age counts for the participants; ages below 10 and above 80 are not counted



Ziffernketten

1. Abschnitt: Test

Sortierung: sequentiell; aktiviert: nein; erlaubte Wiederholungen: 3

2. Abschnitt: Experiment

Sortierung: zufällig; aktiviert: ja; erlaubte Wiederholungen: 3

1. Hört sich der Wortanfang eher an wie das "s" in | http://www.phonetik.uni-muenchen.de/forschung/Bas/Experimente/ | wie s | zwischen s und ß | wie ß | weiß nicht | leer

2. Geben Sie hier Text für das Label ein | Geben Sie hier die URL einer Audiodatei an. | leer | leer

3. Abschnitt: Verabschiedung

Sortierung: sequentiell; aktiviert: ja; erlaubte Wiederholungen: 3

- Experiment-Projekte
  - Irony
    - Experimente
      - irony detection
        - Experiment-Skripte
          - IRONY01
            - Experiment-Skript-Sektionen
              - Experiment
                - Experiment-Elemente
                  - Experiment-Element
                  - Experiment-Element
                  - Experiment-Element
                  - Experiment-Element
                  - Experiment-Element
                  - Experiment-Element
  - Studentenprojekt
    - Experimente
      - GroesseGewicht
        - Experiment-Skripte
          - GrGe0002
            - Experiment-Skript-Sektionen
              - GrGeTEST
              - GrGeEXPERIMENT
                - Experiment-Elemente
                  - Experiment-Element
                  - Experiment-Element
                  - Experiment-Element
                  - Experiment-Element
                  - Experiment-Element
                  - Experiment-Element
                  - Experiment-Element
                  - Experiment-Element
                  - Experiment-Element
                  - Experiment-Element
                  - Experiment-Element
                  - Experiment-Element
                  - Experiment-Element

Detail-Ansicht

Experiment-Element:

Beschriftung: Schätzen Sie die Größe der Person

URLs: http:/www.phonetik.uni-muencher

URL-Felder hinzufügen: 1

Antwortmöglichkeiten: kleiner, 130, 140, 150, 160, 170, 180

Figure 3: Screen shots of the experiment content edtiors

4.1.3. Rate of email addresses provided

In most experiments, the participant input form contains an optional email field. This field is needed to be able to identify individual participants, e.g. to allow them to take part in a raffle. 52.5% filled in this field, and there is virtually no difference between the sexes.

4.1.4. Response rate

Recruiting experiment participants is a tedious and time-consuming effort with an uncertain outcome. Anecdotically, we present recruitment data for two experiments; section 4.2.2. presents data from a third and international experiment.

In a seminar on perception experiments, students set up an online experiment asking participants to estimate the size and weight of a speaker based on listening to a read sentence. The students then employed four recruitment methods: addressing people at lunchtime in the cafeteria, sending personal emails to friends, and distributing a call for paticipation via Facebook or via the university mailing list. The university mailing list vastly outperformed the other recruitment methods: within five days, 280 participants were recruited, compared to 54 via the other methods. Response is fast, but also very short-lived: after five days, the number of participants dropped dramatically to one or two per day.

The second example is an experiment on voice identification. In the first call for participation, we asked for German native speakers with no hearing impairments via the university mailing list. In a second call for participation, we sent emails to people who had provided email addresses in previous experiments and who were male German native speakers.

As in the other example, response was swift, but short-lived: in the five days following the first call for participation, 80 participants could be recruited, and 48 of them completed their sessions. In the five days following the targeted email, 52 participants were recruited, 33 of which completed their experiment sessions. Clearly, having a participant database is quite useful in recruiting specific participants.

As a side-effect the targeted email was useful to assess the quality of the participant database. The targeted email was sent to 319 people; 11 email addresses were invalid, and two participants requested to be removed from the database.

4.1.5. Experiment duration

The average session duration, including all aborted sessions, was 16:08 minutes. Counting only the completed experiment sessions, the average duration was 12:12 minutes. The aborted sessions have a longer average duration, namely 15:24 minutes. Two reasons for this may be that a) participants found the experiment too difficult and thus too time-consuming, or that b) participants encountered technical difficulties such as low data transfer rate, which made the experiment slow and boring.

4.2. Experiment-specific results

4.2.1. Regional variation in spoken digits

The Ph@ttSessionz speech database (Draxler and Steffen, 2005) contains recordings of 1102 adolescent speakers recorded via the Internet in 41 grammar schools in Germany. The speech material consists of spoken digits, digit chains, spellings, date and time expressions, and phonetically rich sentences as well as spontaneous speech.

The research question to be answered by a perception experiment was whether the spoken digits are sufficient to determine the regional origin of a speaker. Digits are interesting because they can be presented in a non-orthographic numerical format, occur in every dialect region, and they are, with the exception of the digit 7 ('sieben'), monosyllabic.

In this series of experiments, we focused on the following regional variants on the phoneme level:

- voiced word-initial fricative and plosive in 'sieben' (/z i: b @ n/) and 'drei' (/d R aI/) respectively,
- lip rounding in the vowel 'u' in 'null' (/n U l/),
- monophthongization of the word-initial diphthong in 'eins' (/? aI n s/),
- realization of the word-final vocalic R /6/ in 'vier' (/f i:6/).

Following the literature, it could be expected that

- the fricative in 'sieben' would be produced as a voiceless fricative in southern Germany,
- the plosive in 'drei' would be produced as devoiced or even voiceless plosive in western and eastern Germany,
- the back vowel in 'null' would be produced with lip rounding in eastern Germany
- the vocalic R in 'vier' would be produced as a vowel in northern and eastern Germany

For this experiment, a total of 4647 digit recordings by 1007 spekers were selected. In order to keep the experiment duration short, every participant was presented only approx. 45 stimuli.

The target audience were laymen, and hence the questions were formulated in a non-technical way: *Klingt der Wortanfang eher wie das 's' in 'reisen' oder das 'ß' in 'reißen'?* (Does the word begin sound more like the 's' in 'reisen' or the 'ß' in 'reißen'?).

The experiment started in Nov. 2010. Until the end of February 2014, a total of 632 sessions were run, with 165 male and 466 female participants (in one experiment, no sex was given). The completion rate is 85.9%.

To check whether participants were actually able to detect phonemic differences in the stimuli, three test items were presented at the beginning of each experiment session. In the first item, participants were asked whether they could perceive a difference in the phoneme of interest. The results for the different phonemes differ substantially – 16.1% of the participants did not perceive a difference between word-initial /s/ and /z/ in 'sieben' (seven), 5.0% did not perceive a difference in the medial /u/ in 'null', 8.8% did not perceive a difference between the voiced and voiceless initial plosive in 'drei' (3), and 2.5% did not perceive a difference in the word-final /6/ in 'vier' (4).

A mixed model analysis of the experiment data did not show any significant relationship between either the speaker's regional origin or the participant's regional origin – expressed in the federal state – and the perceived sound. The reasons for this may be that a) the recording sites were not distributed evenly over Germany, b) that the number of judgments for each stimulus varies greatly (from 1 to more than 80), and c) that some regional characteristics do not occur very often, e.g. monophthongization of 'aI' in 'eins' (one).

However, a visualization of the participant's input based on a geographic map of Germany reveals that there is a marked geographical distribution of phoneme realizations and their geographic location. This implies that the German federal states alone are not sufficient to classify dialects. It may be interesting to relate the perception experiment results to objective acoustical measurements.

4.2.2. Dark and clear /l/

In her experiment 'Laterals' Daniela Müller asks participants to categorize manipulated /l/-Stimuli into 'clear' or 'dark' /l/. For this, she has set up an experiment with 128

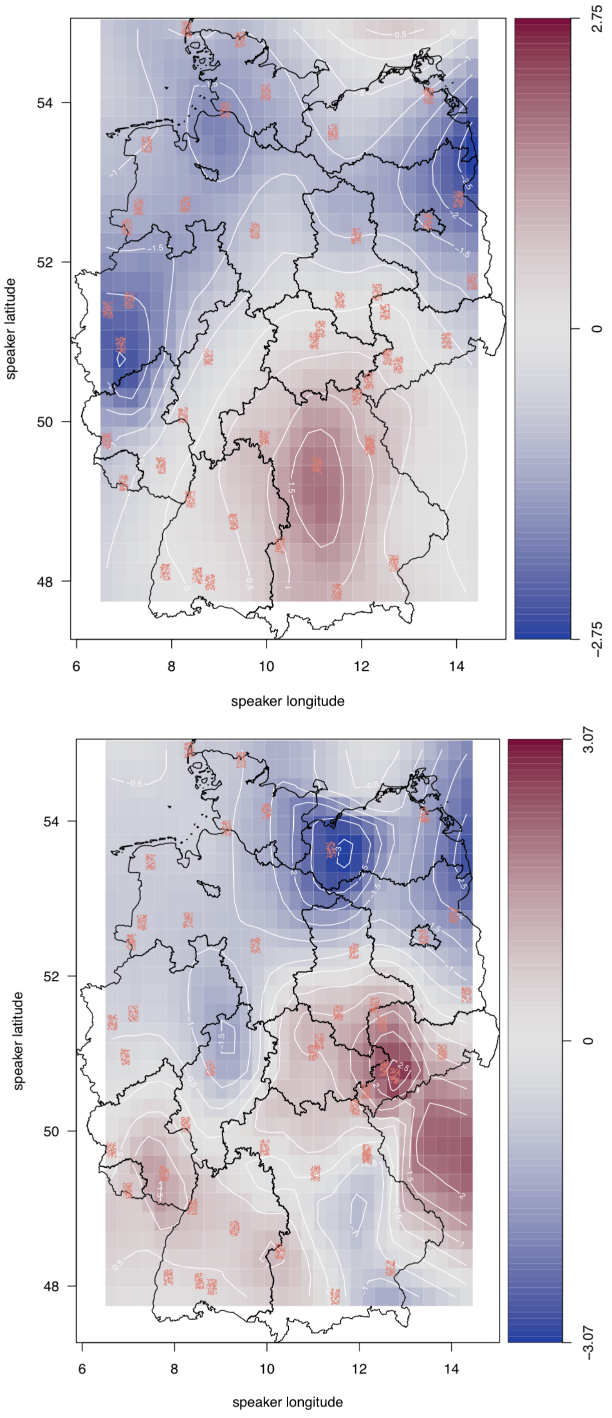


Figure 4: Geographical distribution of /z/ vs. /s/ in *sieben* (seven) and /d/ and /t/ in *drei* (three). Red areas show high values for the selected feature, i.e. voiceless or devoiced fricative and voiceless or devoiced plosive.

stimulus items with a slider input element ranging from ’clear’ to ’dark’ (expressed numerically as an interval between 0 and 100).

She published a first call for participation on linguist list on Aug. 8th 2013, a second call was sent directly to colleagues, now with a lottery of Amazon vouchers as an incentive, on Sept. 17th 2013. The first call yielded 71 experiment sessions, the second 301 sessions. For the first call, 41 sessions (56.94%) were performed within 2 days after the call had been published. For the second call, 274 sessions (91.03%) where performed within 7 days after the call had been sent.

Of these 372 sessions, 204 (54.84%) sessions were complete, 168 (45.16%) incomplete.

The average duration of a complete experiment session is 16:25 minutes, which is rather long. If participants decided to abort the experiment, they did so quite early: almost 50% of the incomplete sessions were aborted within the first 10 items and within two minutes after the first input.

199 participants answered the final question on how difficult they found the experiment. 16 (8.04%) found it to be very difficult, 122 (61.31%) difficult, 58 (29.16%) easy, and 3 (1.51%) very easy.

A preliminary analysis of the experiment shows that trained linguistic participants distinguish clear and dark /l/ categorically.

In the examined languages English, German and Greek, the /l/-quality is allophonic. No participant made a categoric distinction between clear and dark /l/. The degree of phonetical expertise does not have an influence on the results.

The stimuli contained two cues for the classification as either clear or dark /l/, namely 1) spectral variation and 2) the duration of the transition from vowel to lateral. In general, the participants used both cues, but the first cue seems to be much more relevant. This is particularly true for English listeners who almost never used the second cue.

It is thus safe to state that in the three languages analysed, the spectral variation of the /l/ is the primary cue, and transition duration a secondary cue. A publication with an in-depth analysis of the results is currently being prepared.

5. Summary and Conclusion

The online perception experiment system Percy has shown to be robust and suitable for performing audio experiments via the web. The most successful means of recruiting participants seem to be mailing lists – participants respond quickly (within a week), and in large numbers. It is not strictly necessary to offer an incentive to the participants, but doing so is likely to increase the number of participants.

Until now, setting up an experiment using Percy requires the help of technical staff. However, editors for the contents of an experiment, and for its layout, are currently being implemented and they should be available soon. With these editors, researchers will be able to design their own experiments. However, for security reasons, experiments will only be allowed to go online after they have been checked by the system administrator.

One major issue is that the reasons for aborting an experiment are unknown. One means to collect additional information on why participants are aborting the session may be to add an explicit ’exit’ button, and then to ask the participant why he or she wants to terminate the session.

Percy is provided as a free service to academia, and the source code is available upon request to the author of the paper.

5.1. Acknowledgements

The author thanks the Statistical Consulting Unit, Department of Statistics, Ludwig Maximilian University Munich for its support of the analysis of the German dialect perception experiment data.

6. References

Anders, C., Hundt, M., and Lasch, A., editors. (2010). *Perceptual Dialectology – Neue Wege der Dialektologie*. de Gruyter, Berlin.

Anders, C. (2010). Die wahrnehmungsdialektologische rekodierung von laienlinguistischem alltagswissen. In Anders, C., Hundt, M., and Lasch, A., editors, *Perceptual Dialectology – Neue Wege der Dialektologie*, pages pp. 67–87. de Gruyter, Berlin.

Draxler, C. and Steffen, A. (2005). Ph@ttSessionz: Recording 1000 adolescent speakers in schools in Germany. In *Proc. Interspeech*, pages 1597–1600, Lisbon.

Draxler, C. (2011). Percy – an HTML5 framework for media rich web experiments on mobile devices. In *Proc. Interspeech*, pages 3339–3340, Florence, Italy.

Keller, F., Subahshini, G., Mayo, N., and Corley, M. (2009). Timing accuracy of web experiments: A case study using the webexp software package. *Behavior Research Methods, Instruments and Computers*, 41(1):1–12.

Lefever, S., Dal, M., and Matthiasdottir, A. (2007). Online data collection in academic resarch: advantages and limitations. *British Journal of Educational Technology*, Vol 38(No 4):pp. 574–582.

Reips, U. and Lengler, R. (2005). The web experiment list: A web service for the recruitment of participants and archiving of internet-based experiments. *Behavior Research Methods, Instruments, and Computers*, 37(2):287–292.

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

Reips, U. (2002b). Wextor: A web-based tool for generating and visualizing experimental designs and procedures. *Behavior Research Methods, Instruments, and Computers*, 34(2):234–240.