LingoTurk: managing crowdsourced tasks for psycholinguistics

Florian Pusse and **Asad Sayeed** and **Vera Demberg**
Computational Linguistics and Phonetics/M^{2}CI Cluster of Excellence
Saarland University
66123 Saarbrücken, Germany
{fpusse,asayeed,vera}@coli.uni-saarland.de

Abstract

LingoTurk is an open-source, freely available crowdsourcing client/server system aimed primarily at psycholinguistic experimentation where custom and specialized user interfaces are required but not supported by popular crowdsourcing task management platforms. LingoTurk enables user-friendly local hosting of experiments as well as condition management and participant exclusion. It is compatible with Amazon Mechanical Turk and Prolific Academic. New experiments can easily be set up via the Play Framework and the LingoTurk API, while multiple experiments can be managed from a single system.

1 Introduction

LingoTurk is a crowdsourced experiment management system aimed at the use case where the experiment user interface must be highly customized. Common web browsers now permit the design of experimental user interfaces with highly sophisticated presentations, allowing crowdsourcing environments to be used as laboratories for psycholinguistic experimentation with paradigms that in the recent past could only be run "in-lab".

Crowdsourcing in language science was originally popularized among researchers for the collection of labelled training data. It has recently gained popularity as a platform for collecting experimental data for cognitive modeling (e.g., Gibson et al., 2013; Kush et al., 2015). Experimenters trade direct control over subject demographics and environment for faster, cheaper experiment completion with many more subjects. Crowdsourcing can provide experimenters with a way to access populations which are not locally available (e.g., native speakers of a non-local language). Commercial platforms provide micropayment architectures to provide rewards to users. They also manage abuse, track user reliability, track (usually-pseudonymized) identities, and recruit participants.

We designed LingoTurk to handle the condition where the actual experiment must be hosted outside of the "default" systems provided by crowdsourcing platforms. This is motivated by cases in which there is functionality not directly supported by the crowdsourcing platform but can be managed externally, such as the separation of experimental conditions or the storage of specialized data types. Insofar as common crowdsourcing platforms support external interfaces, LingoTurk provides an easily deployed server-side solution to external experiment management. As its administration functions are also web-based, LingoTurk allows for the steering and management of crowdsourcing experiments to be performed without strong technical skills, such as student assistants in non-technical majors.

The source code is made available at `https://github.com/FlorianPusse/Lingoturk`.

1.1 Crowdsourcing and language science

Crowdsourcing has been a trend in language research for the better part of a decade and has principally been focused on the collection of annotated training data for supervised machine learning in fields like machine translation (Zaidan and Callison-

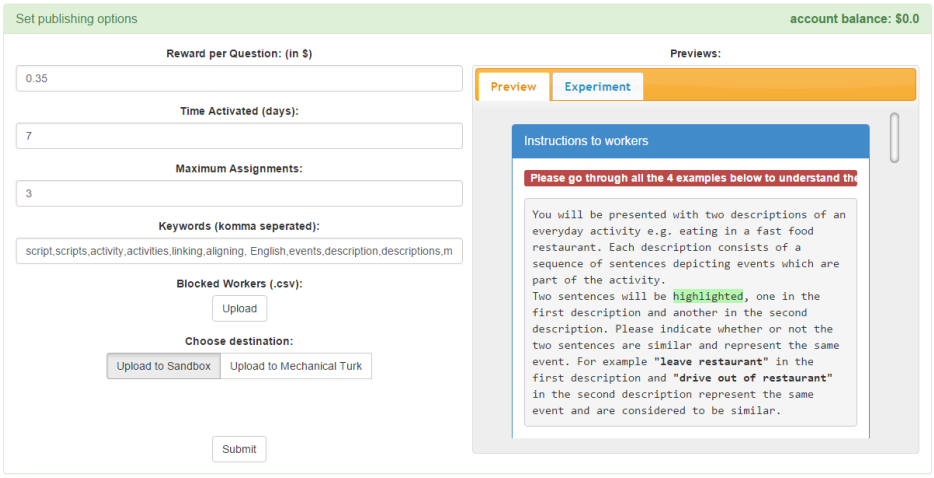

Figure 1: Publishing an experiment to MTurk.

Burch, 2011) and opinion mining (Sayeed et al., 2012). In these areas, the psyche of the annotators is not the principal object of interest.

This has consequences of the relationship of the "requester"—Amazon's term for the task designer—to the crowdsourcing worker. When psycholinguistic experiments are crowdsourced, the objects of interest are no longer directly the "annotations" themselves, but rather what they reveal about the humans who made them. This means that the relationship of the requester to the workers is quite different from what it is in annotation efforts: qualification for the task is replaced by qualification for the study, and annotator reliability is augmented by the need for experimental control.

Psycholinguistic experiments attempt to confirm a hypothesis about the way in which linguistic stimuli are evaluated by the human mind. Experimental items are therefore often separated by condition Normally each subject should only see one item in each condition.

Furthermore, "learning effects" are often a risk in psycholinguistic experiments. Subjects can get used to the experimental paradigm, and as time goes on, their responses can be said to become less and less the spontaneous reaction of linguistic cognition.

Consequently, fine-grained control over condition presentation and worker exclusion are desiderata of a crowdsourcing platform for psycholinguistic experimentation. LingoTurk is designed to address this need via the self-hosting of experiments while offering integration with existing crowdsourcing platforms.

1.2 Crowdsourcing platforms

Amazon Mechanical Turk (MTurk) is the earliest, most widely-used crowdsourcing platform. MTurk provides a set of standard task designs for crowdsourcing as well as the option to create a custom task. Custom tasks can be hosted on an external server, if they are served to Amazon via the MTurk API.

However, MTurk lacks the architecture for experimental exclusion of workers (subjects) by condition. Nevertheless, its API provides the information to construct one server-side, if experimenters host the task on their own server. The MTurk API also permits the experiment interface to appear as a pane inside the MTurk interface, allowing subjects to experience the task seamlessly.

Prolific Academic (PA) is an alternative platform, useful for needs currently unmet by Amazon, particularly a non-US-centric clientele. PA does not provide an API that allows for full external-question integration, but it allows for worker redirection that permits similar server-side participant tracking.

1.3 Comparison to alternative systems

There are other server-side experiment publishing platforms aimed at psychological and psycholinguistic research: for example, Ibex (Drummond, 2013) and PsiTurk (McDonnell et al., 2012). These platforms have functions that overlap with Lingo-

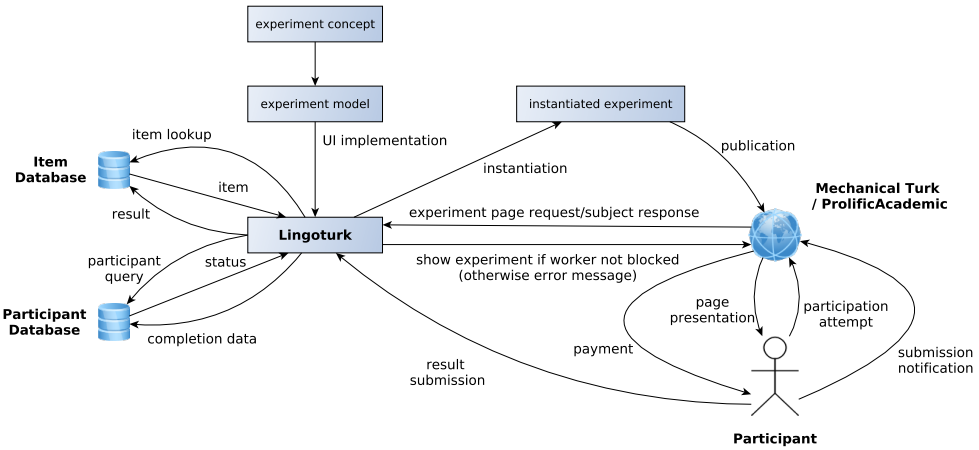

Figure 2: Workflow of crowdsourcing task using LingoTurk.

Turk, but do not cover LingoTurk's full design goals. LingoTurk provides an administration GUI front end for experimenters (figure 1), so that the day-to-day management does not have to be performed via the command line by technically-skilled researchers. LingoTurk is also integrated with the Play Framework, which is intended to accelerate the development of complex, highly scalable web applications using a well-engineered Model-View-Controller (MVC) paradigm in Java and Scala. The MVC paradigm facilitates the development not only of the subject-facing experiment UI, but also user-friendly experiment item entry and testing views for stimulus preparers.

2 Design and workflow

The MVC paradigm combined with database integration makes LingoTurk a platform for reliably engineered experimental interfaces that can handle complex data structures as well as easy object-oriented extensibility. LingoTurk is intended for a self-hosting use case; once a web server has been set up, the Play Framework enables LingoTurk to be a turn-key solution for experiment administration.

Figure 2 shows the design of LingoTurk in terms of its overall workflow. LingoTurk manages communication with the crowdsourcing platform as well as governs interactions with a participant database which keeps track of experimental conditions and exclusions and an item database that keeps track of stimuli and responses. LingoTurk selects the pages to be served to the crowdsourcing platform based on information provided by the platform. On MTurk, that means that a worker who is ineligible to see an experimental condition will be presented with a page that informs of them of this and asks them to return the task.

Creating a LingoTurk experiment based on an existing interface type (section 3) is done from the graphical administrative console, which is itself a web site. The experimenter instantiates an experiment type and fills the stimuli into web forms that are designed to handle experimental conditions. The experimenter also uses the administrative interface to provide credentials for the crowdsourcing platform as well as to preview and publish the experiment and retrieve the results. Excluded worker IDs (such as those who participated in previous runs of the experiment) can also be uploaded to LingoTurk this way. New experiment designs can be developed and added to the interface using Play Framework-based HTML and Scala templates; common Javascript libraries are provided by default, and an API is provided for server communication.

LingoTurk also allows for the creation of quality control questions that can be used to exclude poorly-performing workers after a threshold of wrong answers is reached. To use this feature, the experimenter must include stimuli with correct or expected responses.

3 Experiment interfaces

LingoTurk has been used for experiments performed by researchers at Saarland University. Here, we

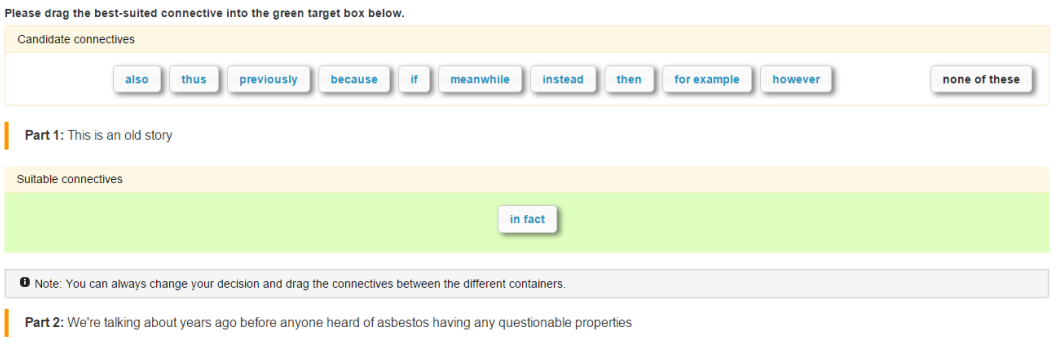

Figure 3: The discourse connective drag-and-drop task.

discuss a couple of examples of interfaces that are included with LingoTurk. We provide other paradigms in the package.

Drag-and-drop discourse connectives Demberg et al. (2015) present the problem of discourse relation prediction: specifically, how do speakers interpret an implicit gap between propositions? They investigated this question through an experiment that allowed subjects to fill in the gap between two sentences, implicitly connected in the Penn Discourse Treebank, with an explicit discourse connective. For this purpose, they used a drag-and-drop paradigm, wherein subjects selected connectives from a set of labelled tiles and dragged them into a target zone (figure 3). They divided this task into phases in order to narrow down the discourse type, with LingoTurk presenting a subsidiary tableau of connective phrases depending on the result of the first tableau. The selection of connective tableaux is controlled via the item entry interface on the administrative side of LingoTurk. Data analysis for this task is still on-going.

The advantage of a drag-and-drop paradigm is that it requires the subject to make an explicit choice but also to use a little bit of effort in doing so. This reduces the bias that might be introduced by the least-effort of choosing the first or the last item (Sayeed et al., 2011).

Script alignment by connector drawing Wanzare et al. (2016) use the LingoTurk system to present a task involving the alignment of collected narratives to prepared scripts (e.g., for baking a cake) for an on-going project that is investigating the psycholinguistic aspects of script knowledge as well as developing script corpora. In this paradigm, steps of the narrative (an account of an action collected from subjects during previous research) are presented as tiles in a column on the left side of the window, while steps of the standardized script are presented on the right side. Subjects draw connections between narrative steps and standardized script steps.

4 Demonstration

For our demonstration at the conference, we will bring a computer and present a running instance of the LingoTurk system. We will proceed through the construction of example experiments which will be pushed through to the MTurk Sandbox (MTurk's testing server). We will make use of our built-in experimental paradigms to demonstrate the versatility and convenience of LingoTurk. We will also demonstrate the underlying practical details of developing new experimental paradigms and integrating them into LingoTurk.

5 Future work

There are considerable opportunities to expand the system. One possible direction is integration with other platforms that have been gradually emerging, such as ClickWorker. We are exploring the possibility of integration with the CrowdFlower platform; an important challenge in this case is the integration with CrowdFlower's built-in quality control system. Another direction is increasing the customisability of experiment designs by including a graphical web design tool, reducing the need to interact directly with the Play Framework when developing new experimental paradigms.

References

Demberg, V., Sayeed, A., and Pusse, F. (2015). Discourse annotation via mechanical turk. In *First Action Conference of the TextLink COST Initiative*, Louvain-la-Neuve, Belgium.

Drummond, A. (2013). Ibex farm. *Online server: http://spellout.net/ibexfarm*.

Gibson, E., Bergen, L., and Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056.

Kush, D., Lidz, J., and Phillips, C. (2015). Relation-sensitive retrieval: Evidence from bound variable pronouns. *Journal of Memory and Language*, 82:18 – 40.

McDonnell, J., Martin, J., Markant, D., Coenen, A., Rich, A., and Gureckis, T. (2012). psiturk (version 1.02)[software]. new york, ny: New york university.

Sayeed, A. B., Boyd-Graber, J., Rusk, B., and Weinberg, A. (2012). Grammatical structures for word-level sentiment detection. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 667–676, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sayeed, A. B., Rusk, B., Petrov, M., Nguyen, H. C., Meyer, T. J., and Weinberg, A. (2011). Crowdsourcing syntactic relatedness judgements for opinion mining in the study of information technology adoption. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '11, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wanzare, L., Zarcone, A., Thater, S., and Pinkal, M. (to appear 2016). A crowdsourced database of event knowledge sequence descriptions for the acquisition of high-quality script knowledge. In *Language resources and evaluation conference*, Portorož, Slovenia.

Zaidan, O. F. and Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 1220–1229, Stroudsburg, PA, USA. Association for Computational Linguistics.