The NIEUW Project: Developing Language Resources Through Novel Incentives

James Fiumara, Christopher Cieri, Mark Liberman, Chris Callison-Burch*, Jonathan Wright, Robert Parker
University of Pennsylvania Linguistic Data Consortium and *Department of Computer and Information Science
{jfiumara, ccieri, myl, jdwright, parkerrl} @ldc.upenn.edu, {ccb} @cis.upenn.edu

Abstract
This paper provides an overview and update on the Linguistic Data Consortium's (LDC) NIEUW (Novel Incentives and Workflows) project supported by the National Science Foundation and part of LDC's larger goal of improving the cost, variety, scale, and quality of language resources available for education, research, and technology development. NIEUW leverages the power of novel incentives to elicit linguistic data and annotations from a wide variety of contributors including citizen scientists, game players, and language students and professionals. In order to align appropriate incentives with the various contributors, LDC has created three distinct web portals to bring together researchers and other language professionals with participants best suited to their project needs. These portals include LanguageARC designed for citizen scientists, Machina Pro Linguistica designed for students and language professionals, and LingoBoingo designed for game players. The design, interface, and underlying tools for each web portal were developed to appeal to the different incentives and motivations of their respective target audiences.

Keywords: novel incentives, citizen science, language resources

1. Introduction

Human language technologies (HLT), linguistic research, and language teaching all rely heavily on a variety of Language Resources (LRs) and have benefited immensely from decades of linguistic data creation and sharing supported by governments and research institutes. Continued efforts from data centers such as LDC[1], European Language Resources Association (ELRA)[2], LDC for Indian Languages[3], and South African Center for Digital Language Resources (SADiLaR)[4] have made large amounts of data available to the research community. However, the overall amount of LRs available globally still falls short of need. Traditional approaches are unlikely to meet the needs of the research community due to finite resources of funding versus the effort required to create these LRs. LDC's NIEUW projects seeks to close this gap by using novel incentives and workflows to collect linguistic data and judgments and to make these data available world-wide to the research community.

2. Novel Incentives and LR Creation

Given the high cost and extensive effort that goes into collecting and annotating linguistic resources, HLT researchers have increasingly looked to alternative incentive models and crowdsourcing options such as Amazon Mechanical Turk (MTurk). While MTurk's ability to collect large numbers of discrete HITs (Human Intelligent Tasks) at extremely low cost per HIT allows the creation of resources for as little as $1/10^{th}$ the typical cost (Callison-Burch and Dredze 2010), the crowdsourcing platform is not without issues including variable annotation quality (Tratz and Hovy 2010) and, more importantly, ethical concerns of exploitive labor practices (Fort et al. 2011). MTurk is certainly one alternative to traditional linguistic resource creation practices in its utilization of large-scale crowdsourcing rather than employing a small group of experts, but ultimately the platform still relies on the incentive of monetary compensation, even if comparably small amounts of it meted out by microtask. However, even outside of ethical concerns, this reliance on monetary compensation is ineffective when there is a lack of funding, when monetary payment is not permitted, or when potential contributors are motivated by factors other than money and can be problematic when workers use unexpected means to maximize their earnings.

Alternatively, citizen science, social media, and games with a purpose (GWAP) have shown that people are willing to volunteer large amounts of time and effort given appropriate non-monetary incentives which can include entertainment, competition, learning and education, social interaction, demonstrating expertise, and contributing to a social good. Successful examples outside of the HLT community are many, including LibriVox[5] which organizes volunteers to record audiobooks from out-of-copyright works which are then made freely available to the public and Zooniverse[6], an online citizen science platform that has recruited over two million volunteers generating over six hundred million classifications for a variety of research projects in astronomy, zoology, biology, medicine, history, and climate science. Novel incentives have also been used effectively inside of the HLT community including the now defunct The Great Language Game (Skirgård, Roberts & Yencken 2017) which collected tens of millions of language identification judgments and Phrase Detectives, an online game with a purpose which has collected millions of judgments on anaphoric expressions in two languages since going live in 2008 (Poesio et al. 2016).

Building upon these models, the NIEUW project enhances LR development well beyond what project-dependent, direct funding alone can accomplish by creating an infrastructure that enables the ongoing construction of scalable data collection and annotation activities available

[1] https://www.ldc.upenn.edu
[2] http://www.elra.info
[3] https://www.ldcil.org
[4] https://sadilar.org
[5] https://librivox.org
[6] https://www.zooniverse.org

Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People @LREC2022, pages 1–7
Marseille, 25 June 2022.

to the public via the web and mobile devices and designed with appropriate incentive models in mind.

We argue that the best way to attract and engage non-traditional workforces is to offer a variety of incentives that are organized in separate, but not mutually exclusive, groupings. We have identified three already existing communities that we think present the most promise: 1) citizen scientists who are motivated to participate in linguistic research due to an interest in language and culture or the desire to contribute to research and technology development; 2) language students and professionals such as linguists, transcriptionists and professors who work directly with linguistic data but would benefit from improved tools and infrastructure; 3) game players who seek entertainment, challenge, and competition.

For NIEUW we have created three web portals geared towards each of these communities containing language collection and annotation activities and games designed to appeal to the respective groups. The website design, task size and complexity, tool builders and workflows were all created with the relevant participant populations in mind. Although LDC initially created tasks for these portals, they have since been made available to research collaborators. By allowing language researchers to create their own projects on these portals, the infrastructure not only serves the larger research community but creates a sustainable resource that can continue to grow without being tethered to any particular project goal or funding requirement.

3. LanguageARC : A Portal for Citizen Linguistics

Crowd-sourced contributions to scientific research by the general public (often called "citizen science") has a long history from Edmund Halley who solicited the public to help map solar eclipses (Pasachoff 1999) to bird lovers helping the Audubon Society count or track birds (Root 1988). Recent technologies such as the internet and smart phones have made it even easier for the public to contribute to science. Building on this history, LanguageARC[7] (**A**nalysis **R**esearch **C**ommunity) is a citizen science platform and community dedicated to language research ("citizen linguistics").

Figure 1: Language ARC web portal

3.1 Overview of LanguageARC

LanguageARC brings together researchers and citizen linguists by hosting language research projects that present specific tasks, activities, and goals. Citizen linguists can participate in these research projects by contributing judgements and data, and through project chat room discussions with both researchers and other participants. Projects are comprised of one or numerous tasks with each task consisting of a discrete activity for the participant to perform in response to input data which can be in the form of text, audio, image, or video prompts.

By way of example, the *Fearless Steps* project presents several distinct tasks ranging from beginner to advanced level which ask participants to listen to audio communication clips from NASA Apollo missions and provide judgments or transcriptions of the audio. The task Speaker Count presents an audio clip and asks the participant to identify the number of speakers in that clip and if the speech overlaps across speakers from a restricted set of answer options simply by clicking a button.

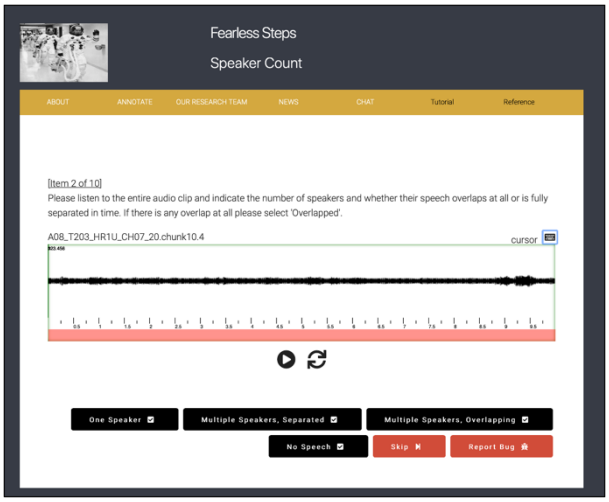

Figure 2: *Fearless Steps* "Speaker Count"

Participants can join the LanguageARC community with as little as a login username and a valid email address for verification, but the registration form also allows the collection of optional demographic information about the participant such as date of birth, gender, and languages spoken. Once registered, a Language ARC member can contribute to any public project accessible from the Project menu page (see Figure 3).

Private projects accessible by invitation only are also possible allowing researchers to restrict access to a task to specified users such as a research lab or students in a class. Private projects are only visible to the invited participants and do not show up to the public.

[7] https://languagearc.org

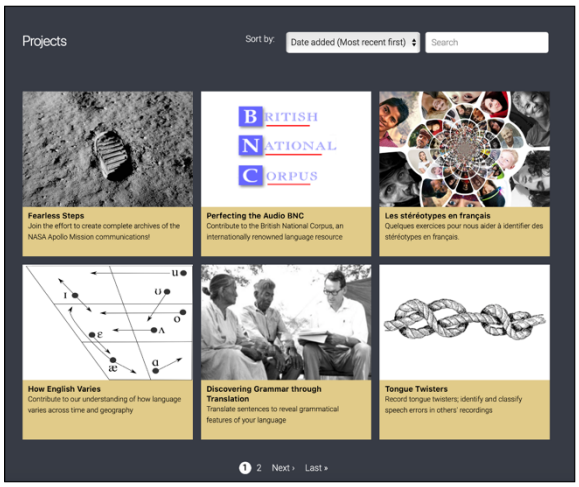

Figure 3: Project Menu

3.2 LanguageARC Project Structure

Language ARC is primarily organized by projects with each project presenting an image, a title, a call-to-action subtitle, and a brief project description. Each project also features the option to include research institute logos, bios for team members, and project message boards for building community and providing a place for participants to interact with researchers and each other. Projects are then sub-organized by tasks which also include their own titles and images, as well as options for tutorials and reference guides to provide any needed background information and task instructions to the participants. Each task includes a tool to collect data and judgments from participants who iterate over items in a dataset.

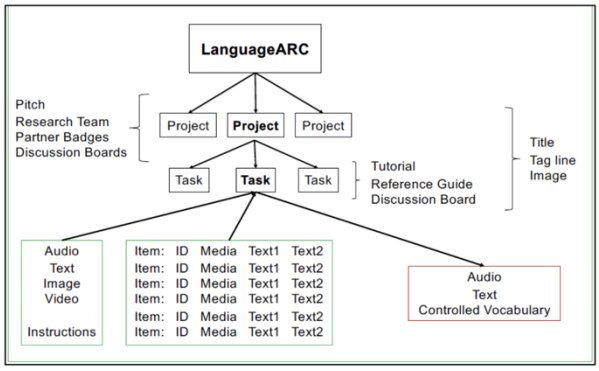

Figure 4: Project structure flow chart

3.3 Building Projects and Tasks

Annotation and data collection tasks on Language ARC are created with a modified toolkit that LDC has built and used to collect millions of annotations and create hundreds of LRs. The toolkit has been modified and adapted to be portable to multiple environments including the web. The toolkit is also open source and can even be deployed to a laptop and taken into the field where there might be no internet access. In order to make Language ARC as widely accessible to the research community as possible, we have created an easy-to-use Project Builder which allows users with little to no coding or programming knowledge to create annotation tasks by uploading formatted data and answering questions in series of templates. The Project Builder guides the user through a step-by-step series of templates from general information (project name, description) to specific task details (data, manifest, and tool options).

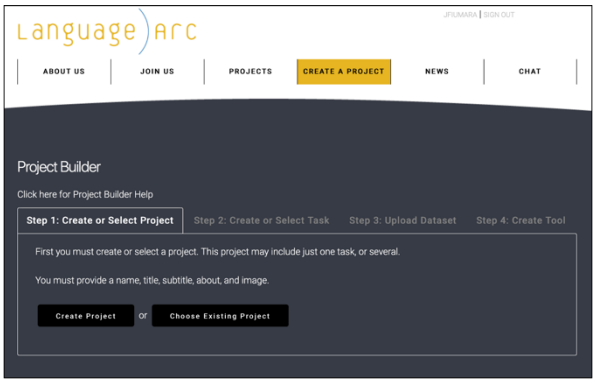

Figure 5: Project Builder main menu

In the first step users create and set-up the basic details of the project including name, description, project image, discussion forums, and research team. Step 2 allows one to create a new task or select an existing task to update. Every project must have at least one task to start and additional tasks may subsequently be added to the project. Tasks can also include Reference Guides and Tutorials which are created with a markdown editor and may include text, images, audio, or video materials.

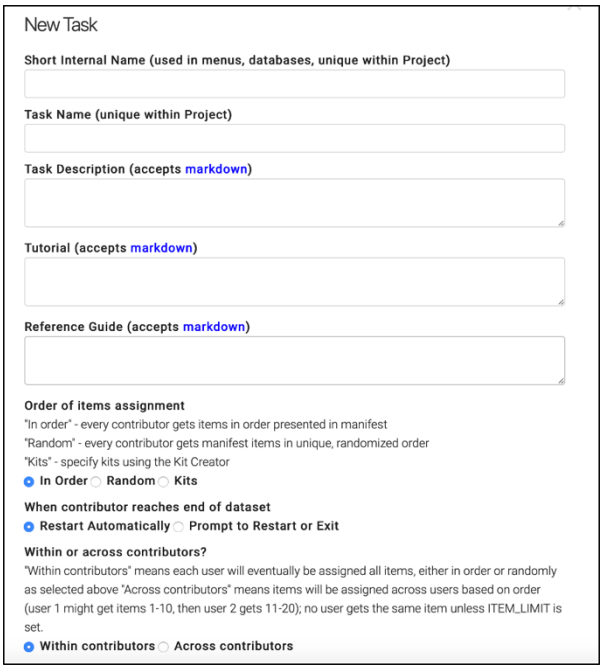

Figure 6: Task set up template

After the basic details of the project and task are created, the third step is to upload the input data which can be text, audio, image, or video data. Input data must also be accompanied by a tab-delimited manifest file, with

column headers, that orders and labels the data for the tool.

ID	File	Label
1	doc1.txt	Document One
2	doc2.txt	Document Two
3	doc3.txt	Document Three
4	doc4.txt	Document Four

Figure 7: Tab delimited manifest

The final step is to create the annotation tool which can also be accomplished simply by answering questions in the template. Certain fields are required such as those indicating the input media type and indicating appropriate columns in the manifest for the media files and prompts. The tool creation form reads the uploaded manifest file so that column headers from your manifest are choices in the dropdown boxes where necessary. Users can provide two columns of text that will be included with – and can vary with – each input file, for example, a label or piece of text description. Participant responses can be in the form of an audio recording, a text box, or controlled responses (buttons or multiple-choice check boxes). Additionally, options for "skip" or "report bad item" buttons are available. Access to the Project Builder is by approval only. Researchers interested in creating a project should reach out via the Contact Us page on the website.

3.4 Current Projects & Future Work

LanguageARC officially launched in October 2019 with a handful of in-house created projects. Since then, the portal has grown to include eleven currently active projects, one completed project, and a half-dozen projects in prototype status. Projects are available in multiple languages including varieties of English (American, British, South African), French, Mandarin and Sesotho, and prototype projects in Swedish, Italian, and Arabic. Current LanguageARC projects support sociolinguistic research, data annotation and collection for corpus building and NLP development, and even collecting linguistic data to support clinical research. There are a wide range of activities and tasks available for citizen linguists to engage with including translation tasks, transcription tasks, listening to audio clips and making judgements about dialect, recording oneself describing pictures, and answering psychological surveys to build general population control data for clinical research.

Some recently added projects include *Fearless Steps* initiative led by UTDallas-CRSS which seeks to audit, categorize, and transcribe audio recordings from NASA space missions; *Les stéréotypes en français* which solicits judgments about stereotypes in French language and culture; and *South African CDI* which collects data about childhood language development in Sesotho and South African English.

Although it is more difficult to evaluate the complexity, usefulness and benefits of HITs across such a diversity of projects and tasks, to date LanguageARC has presented 132,010 HITs to 800 unique userIDs.

LanguageARC infrastructure and toolkit will continue to be developed as required to support current projects and future projects as new needs arise. However, the primary goal currently is to build and sustain the Language ARC community which includes both citizen linguists and researchers. Social media has been a primary tool to reach potential participants. LDC is increasing its outreach both by diversifying our social media presence to new domains (such as YouTube and Instagram) and increasing the number of publicity and advertising campaigns on both social media and relevant citizen science organizations such as SciStarter[8]. LDC will continue to promote LanguageARC to the research community through our newsletter, professional listservs, and presentations at conferences and workshops.

4. MachProLx : Tools for Language Professionals and Students

Language professionals, such as linguists and language teachers, and students may have different incentives than an amateur who wants to contribute to scientific research. In order to meet the needs and incentives of professionals and students, LDC created a separate portal Machina Pro Linguistica,[9] or MachProLx for short.

Figure 7: Machina Pro Linguistica home page

While this portal is built upon the same general framework and toolkit as LanguageARC, it includes additional features such as the ability to create additional project pages using a markdown editor and a version of LDC's powerful web-based transcription tool called *LDC webtrans*. While MachProLx is primarily intended for restricted user groups such as students in a particular class or researchers in a lab, projects can also be made publicly available.

4.1 MachProLx Features

While the MachProLx portal is similar to LanguageARC, there are a few things that set it apart, in addition to the differing user community and corresponding incentive model. One added feature that was motivated by the potential needs of this user community is the ability to create multiple project pages using a markdown editor. This allows, for example, a professor to integrate a syllabus and multimedia instruction materials into a portal project.

[8] https://scistarter.org

[9] https://machprolx.org/

Within a project page one can link to the project tasks and to other created markdown pages allowing maximum flexibility in presenting both annotation tasks and instructional or background material.

MachProLx features the same general underlying toolkit as LanguageARC and any tool or task that one can create in the latter can also be created in the former. However, MachProLx also includes a version of LDC's web-based transcription tool, webtrans (Wright et al. 2021). While the general toolkit in LanguageARC does allow the creation of basic transcription tasks, it is designed for quick, atomized activities in line with the incentives and workflow model for a citizen scientist. For example, simple transcription of brief utterances from short audio clips of a few seconds duration. However, this structure is not sufficient for all research requirements such as the need for detailed, time-aligned transcripts for long duration audio recordings (e.g., a sociolinguistic interview).

To meet these different needs, the transcription tool in MachProLx is designed to create detailed time-aligned transcripts from either single or dual channel audio while still presenting a relatively simple and easy-to-use interface. The top of the tool presents a waveform and the bottom portion the segmented transcript. The two parts are interactive: highlighting a section of the waveform allows the creation of a new transcript segment or highlights the corresponding portion of the transcript for already created segments and clicking a transcript segment will highlight the corresponding portion of the waveform allowing for adjustments to the segment boundaries. The tabular transcript is comprised of time stamps, transcript, and optional speaker and section labels. Downloaded transcripts include these fields plus audio file name in a tab delimited format. The blue lines under the waveform indicate segments. Various functions such as playback, scroll, merge segments, and segment boundary adjustments can be done with keyboard controls. Transcript segments can be marked with speaker labels to indicate speaker turns and section labels to organize parts of the transcript are customizable. First and second pass transcription tasks can be connected so that transcripts created in a first pass can be edited and corrected in a second pass.

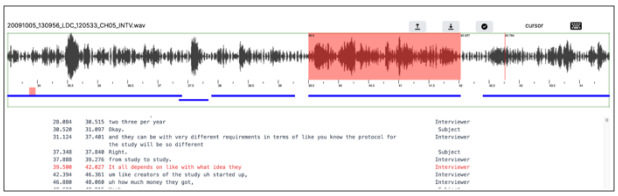

Figure 8: Transcription tool interface

4.2 MachProLx Projects and Plans

Since MachProLx is primarily designed for restricted groups most projects on the portal will not be visible to individuals who have not been invited to a project. Currently, there are a few prototype projects being piloted by research colleagues for lab and classroom use. There are also training and education projects currently available to the public and planned open access projects for the near future. We have created the project *Learning to Transcribe* which anyone who wants to learn how to transcribe linguistic data can work on. This learning project also benefits other MachProLx projects by providing a general, shared space for transcription instruction and practice. *Learning to Transcribe* provides reference materials on how to use the webtrans transcription tool, some general transcription specification guidelines, and a practice task where participants can try their hand at transcribing sociolinguistic interviews.

One project in development that will be open to anyone who wishes to participate (after signing up for an account) will be the *Penn Sociolinguistic Archive* project. The Penn Sociolinguistic Archive is a large collection of sociolinguistic interviews recorded by University of Pennsylvania Professor William Labov and his students and collaborators over the past five decades. It consists of recordings from 5813 separate sessions, covering dialects of English spoken across the United States and the world. Selections from this archive will be available to transcribe (after any sensitive or identifying audio has been masked) for research and educational use.

5. LingoBoingo: Language Games Portal

The third portal created under the NIEUW project is dedicated to language games and gamified activities. LingoBoingo[10] differs from the portals for citizen linguists and language professionals and students as it is primarily a webpage that hosts links to external language games in order to pool recruiting resources, improve discoverability and develop collaborations among researchers.

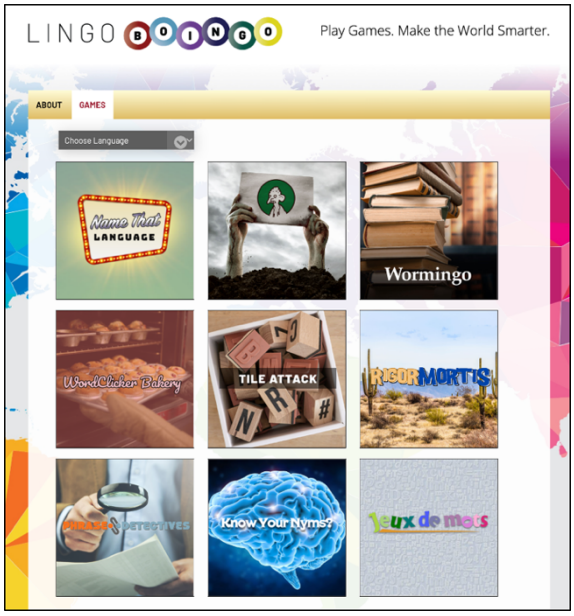

Figure 9: LingoBoingo games menu

LingoBoingo currently hosts nine language games developed by researchers at the University of Pennsylvania's Linguistic Data Consortium and Department of Computer and Information Science, the

[10] https://lingoboingo.org/

University of Essex, Queen Mary University of London, Sorbonne Université, Loria (the Lorraine Laboratory of Research in Computing and its Applications), Inria (the French National Institute for Computer Science and Applied Mathematics), and the Université de Montpellier.

The nine games on the portal are available in English, Italian and French languages and present a variety of game types including Zombilingo, a zombie themed GWAP that allows for the dependency syntax annotation of French corpora (Fort et al. 2014), the previously mentioned Phrase Detectives, the text annotation game WordClicker Bakery (Madge et al. 2019), and the French vocabulary game, Jeux de mots.

Inspired by The Great Language Game, LDC created its own language identification game, called Name That Language (NTL). With a name and visual design inspired by vintage television game shows, Name That Language elicits judgments of languages spoken in brief audio clips taken from broadcast and conversational telephone speech to be used in language recognition and confusability research.

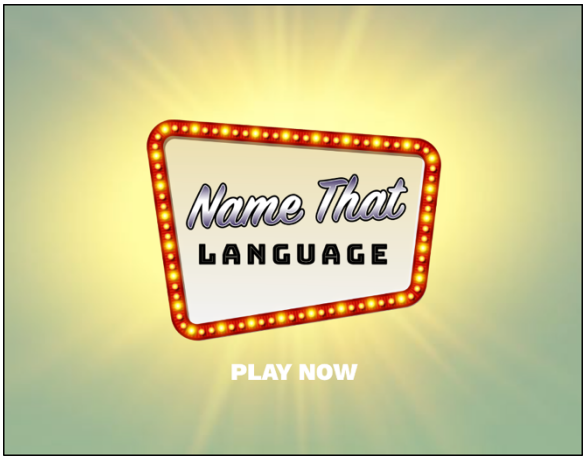

Figure 10: Name That Language splash page

One of the primary ways that Name That Language differs from The Great Language Game is the inclusion of both known clips drawn from published corpora subjected to expert language annotation and suspected clips drawn from broadcasts or conversations purported to be in the target language but not verified. This allows the game to not only collect information about language confusability based on player responses, but also to also to reliably determine the language spoken in suspected audio clips in order to build robust and accurate corpora for language recognition research and technology development.

The game interface was dynamically designed for visually appealing display on computer monitors and mobile devices. The interface presents game logo, scoreboard, audio play and pause controls, buttons with possible languages, and Next and New Game buttons to move on to next clip or restart the game. Players listen to short ~10 second clips of audio and select the believed language by clicking the corresponding button. The known or suspected language is always included as a choice and the number of distractors increases as the game progresses. Players receive 10 points for each correct answer and lose one of three 'lives' for each incorrect answer. The goal is to maximize points earned before losing all three lives.

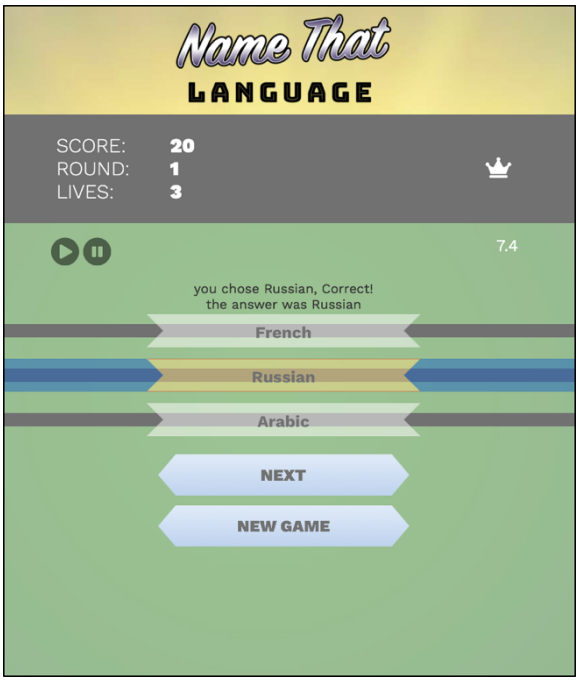

Figure 11: Game play

The initial version of the game included at least 80 known audio clips in 13 languages and approximately 600 suspected clips in each of 9 of those languages. To date, the game has presented 862,608 HITs to 83,991 unique userIDs (a player can have more than one userID) of which 85% have yielded judgements that we can use for Language ID. Together they allow users of the data to determine the language or identify bad clips with a high degree of confidence through the use of a simple voting algorithm (Cieri et al. 2021). The *NTL Language Recognition* corpus resulting from this effort will be released via LDC at no cost. It contains 6680 audio files and a snapshot of the judgements, more than 720,000 database records indicating the file name, known or suspected language, other language choices offered during game play, city and country of the player, date and time of the judgment, and other fields necessary for game administration.

Future plans for Name That Language include adding new audio clips and increasing the number of languages available.

6. Conclusion

LDC's NIEUW project has created infrastructure and tools to dramatically increase the store of LRs by employing novel incentives and workflows proven to work in multiple scientific disciplines and industries. The three NIEUW web portals (LanguageARC, MachProLx, and LingoBoingo) are designed to appeal to different participant communities (citizen scientists, language professionals, and game players) through distinct visual design, workflows, activities, and incentives employed in outreach. These portals are open to the research community who can create

their own linguistic data collection and annotation projects benefiting from the tools, infrastructure, and participant community developed by the NIEUW project.

7. Acknowledgements

The authors would like to acknowledge the support of the National Science Foundation under grant CISE Research Infrastructure (CRI) 1730377.

8. Bibliography

Callison-Burch, C. and Dredze, M. (2010). Creating Speech and Language Data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, pp. 1-12.

Cieri, C., Fiumara, J., Wright, J. (2021). Using Games to Augment Corpora for Language Recognition and Confusability. Proc. Interspeech 2021, 1887-1891.

Fort, K., Adda, G. and Cohen, K. (2011). Last Words: Amazon Mechanical Turk: Gold Mine or Coal Mine?. Computational Linguistics, 37(2):413–420.

Fort, K., Guillaume, B., Chastant, H. (2014). Creating Zombilingo, a Game With A Purpose for dependency syntax annotation. Gamification for Information Retrieval (GamifIR'14) Workshop, Apr 2014, Amsterdam, Netherlands.

Madge, C., Bartle, R., Chamberlain, J., Kruschwitz, U., and Poesio, M., "The Design Of A Clicker Game for Text Labelling," *2019 IEEE Conference on Games (CoG)*, 2019, pp. 1-4.

Pasachoff, J. M. (1999). "Halley as an eclipse pioneer: his maps and observations of the total solar eclipses of 1715 and 1724," Journal of Astronomical History and Heritage, vol. 2, no. 1, pp. 39-54.

Poesio, M., Chamberlain, J., Kruschwitz, U., and Madge, C. (2016). Novel Incentives for Phrase Detectives. 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož Slovenia.

Root, T. (1988). Atlas of Wintering North American Birds: An Analysis of Christmas Bird Count Data, Chicago, IL: University of Chicago Press.

Skirgård, H., Roberts, S. G., & Yencken, L. (2017). Why are some languages confused for others? Investigating data from the Great Language Game. PloS one, 12(4).

Tratz, S. and Eduard Hovy, E. (2010). A Taxonomy, Dataset, and Classifier for Automatic Noun Compound Interpretation. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden. Association for Computational Linguistics, pp. 678–687.

Wright, J., Parker, R., Zehr, J., Ryant, N., Liberman, M., Cieri, C., and Fiumara, J. (2021). High quality recordings and transcriptions of speech via remote platforms. The Journal of the Acoustical Society of America. 150. A356-A356.