Proceedings of the LREC 2020 Workshop "Citizen Linguistics in Language Resource Development", pages 1–6
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

LanguageARC: Developing Language Resources Through Citizen Linguistics

James Fiumara, Christopher Cieri, Jonathan Wright, Mark Liberman
University of Pennsylvania, Linguistic Data Consortium
Philadelphia, PA USA
{jfiumara, ccieri, jdwright, my}@ldc.upenn.edu

Abstract

This paper introduces the citizen science platform, LanguageARC, developed within the NIEUW (Novel Incentives and Workflows) project supported by the National Science Foundation under Grant No. 1730377. LanguageARC is a community-oriented online platform bringing together researchers and "citizen linguists" with the shared goal of contributing to linguistic research and language technology development. Like other Citizen Science platforms and projects, LanguageARC harnesses the power and efforts of volunteers who are motivated by the incentives of contributing to science, learning and discovery, and belonging to a community dedicated to social improvement. Citizen linguists contribute language data and judgments by participating in research tasks such as classifying regional accents from audio clips, recording audio of picture descriptions and answering personality questionnaires to create baseline data for NLP research into autism and neurodegenerative conditions. Researchers can create projects on Language ARC without any coding or HTML required using our Project Builder Toolkit.

Keywords: citizen science, crowdsourcing, language resources, novel incentives

1. Introduction

Linguistic research and Human Language Technology (HLT) development have greatly benefited from the large amount of linguistic data that has been created and shared by data centers, governments and research groups around the globe. However, despite these efforts, the amount and variety of available Language Resources (LRs) falls far short of need. Current approaches to LR development are unlikely to solve the dearth of LRs due to both the overall amount of effort required and to the reliance on finite project-focused funding and collection. The Linguistic Data Consortium (LDC)'s NIEUW (Novel Incentives and Workflows) project supported by the National Science Foundation under Grant No. 1730377 was developed to address these issues by utilizing novel incentives and workflows to collect a variety of linguistic data and annotations and make that data widely available to the research community.

2. Language Resources

Human language technologies, linguistic research and language pedagogy all rely heavily on a variety of LRs. Despite the ongoing efforts of data centers such as the LDC[1], European Language Resources Association (ELRA)[2], Chinese LDC[3], LDC for Indian Languages[4] and the Southern African Centre for Digital Language Resources (SADiLaR)[5], multinational projects such as CLARIN[6] and numerous national and regional corpus creation efforts, the public availability of language resources is only a fraction of what is truly needed for linguistic research and HLT development. One predominant factor is simply that there is a large number of languages in the world; over 7000 by some counts (Eberhard, Simons & Fennig 2019). In addition, the number of resources required to develop minimally necessary technologies in any given language is as much as two dozen (Krauwer 1998, Binnenpoorte, et al. 2002, Krauwer 2003). Another contributing issue is that new language resource production frequently does not result in maximum coverage of languages and resources types, but rather tends to increase the size of existing LRs (Cieri 2017).

In summary, the current approaches to developing LRs required for research and HLT development insufficiently address the problem of lack of language resources. If we hope to rectify the scarcity and imbalance of available resources, new methods of data collection and annotation are required.

3. Novel Approaches to LR Creation

A primary reason that current approaches of LR creation are insufficient is that they tend to rely on finite funding resources for a problem that is multiple orders of magnitude greater. While we are not proposing to replace traditional methods of funding LR development, a promising alternative or supplement is to harness renewable resources that rely on incentives other than monetary. Social media, citizen science and games with a purpose (GWAP) have demonstrated that humans are willing to volunteer vast stores of effort given appropriate opportunities and incentives, which include: competition, entertainment, desire to demonstrate expertise, learning and discovery, the desire to contribute to science or a larger social good and participating in a community. Successful examples include the now defunct The Great Language Game (Skirgård, Roberts, & Yencken 2017) which collected tens of millions of language ID judgments and the citizen science platform, Zooniverse[7], which has solicited hundreds of millions of contributions from approximately two million volunteers.

Following similar incentive models, we have identified three overlapping communities that seem the most promising for these efforts: game players, citizen scientists and language students and teachers. Under the NIEUW project, we are creating community platforms for each of

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

these three communities. We have completed online platforms for game players and citizen linguists and a platform designed for Linguistics students and teachers is currently in development.

Our games portal, LingoBoingo[8], currently includes nine language games developed by LDC and colleagues at University of Pennsylvania's Department of Computer and Information Science, the 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. Lovers of language, grammar and literature can test their knowledge, compete against other players and earn high scores in a variety of linguistic games. Among these nine games is LDC's own Name That Language![9] game which is inspired by The Great Language Game and has already collected nearly 450,000 judgments since October 2018.

However, the bulk of the NIEUW effort has been dedicated to building our citizen science platform, LanguageARC[10].

4. Citizen Linguistics

Contributions to scientific research by the public have a long history, e.g. Edmund Halley soliciting assistance from the public to map solar eclipses (Pasachoff, 1999) and the annual Christmas Bird Count organized by the Audubon Society which started in 1900 (Root, 1988). The advent of the internet, smartphones and social media have only increased the public's ability and incentives to contribute to scientific research endeavors. Following this history, LanguageARC (**A**nalysis **R**esearch **C**ommunity) is a citizen science platform and community dedicated to language; henceforth, "citizen linguistics" and "citizen linguists."

Figure 1: Citizen Linguist portal, LanguageARC.

4.1 LanguageARC Overview

LanguageARC hosts multiple *projects* to which citizen linguists can contribute. A project may contain one or multiple *tasks* and each task is composed of a discrete activity that can be applied to multiple *items* or input data.

For example, the project *From Cockney to the Queen* seeks to identify and understand how people speak across London and Southwest England in relation to various demographics. One task asks contributors to listen to an audio clip and identify the region which the speaker likely comes from, while another task asks contributors to record themselves discussing their own experiences and understandings of language differences across geographic areas. In these tasks, the *items* include audio clips and maps and the contributions include speech recordings and judgments made via button selections.

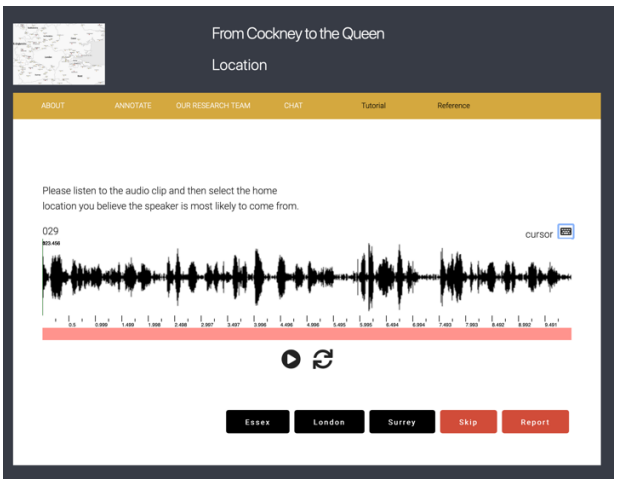

Figure 2: LanguageARC task

Individuals can become a member of the LanguageARC community by providing as little as login ID and email address used for verification purposes, although the registration form also provides a space to collect optional demographic information such as gender, date of birth, languages spoken and geographic regions where one has lived. Once someone joins the LanguageARC community they can participate in any public project on the platform which can be found on the Project menu page (Fig. 3).

LanguageARC also allows the option for private projects which can be accessed by invitation only (though one is still required to join LanguageARC in order to access private projects). Private projects will only be visible to those who have been invited and added to the project. This gives researchers the ability to create a task for a restricted group of contributors such as members of their lab, postdocs or students in one of their courses.

[8] https://lingoboingo.org
[9] https://namethatlanguage.org
[10] https://languagearc.org

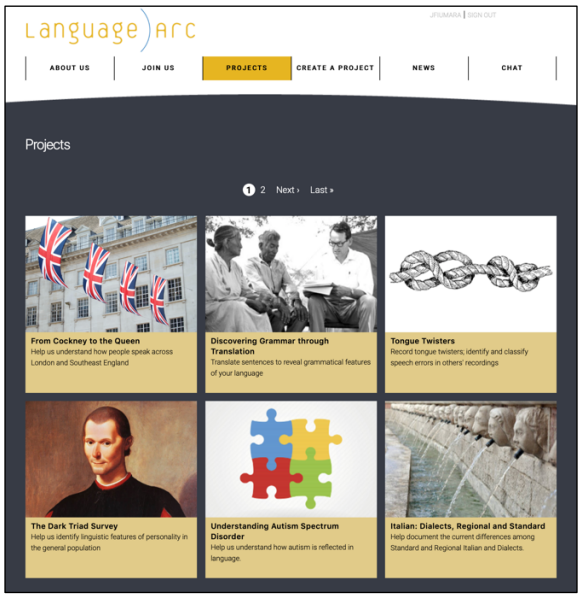

Figure 3: LanguageARC Project menu

Future updates to the project menu page will include search and filter options allowing the ability to search by keyword and filter by categories such as date added, alphabetical by name, the target language of the project and which projects need the most assistance from the community.

4.2 LanguageARC Structure

LanguageARC presents each project by its title, a call to action subtitle, a project image and a brief project description in the form of a pitch. Other project features include a section to highlight the members of the research team and a place for logos and links to the research team's supporting organizations and sponsors. Each project also has the option to have their own project message boards to support community building and provide a place for the citizen linguists to interact with the researchers and each other. Each individual task within a project may have its own title, call to action, task image and message as well as tutorials and reference guides to provide background and instructional materials to the citizen linguists.

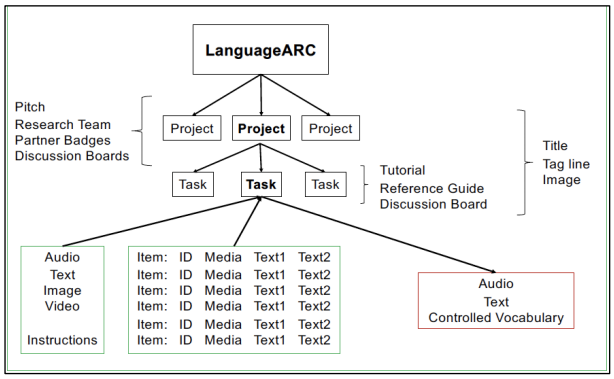

Figure 4: Project structure flow chart

Figure 4 shows the overall structure of LanguageARC described above. The figure also outlines the basic structure of tasks which consist of an *input* (audio, text, video, image), a *tool* which allows contributor interaction with the input, and an *output* (audio, text, controlled vocabulary).

4.3 Toolkit and Project Builder

LanguageARC was created using a modified version of a toolkit that the LDC has built and used to create millions of annotations across more than 100 language resource projects over the past decade. The toolkit has been adapted, modified and extended to make it portable to new environments including on the web. The toolkit has also been made open source and is capable of being deployed to a laptop and taken into the field where there may be no internet access. The modified toolkit source code will be made available on GitHub or similar repository and may be used by researchers outside of the LanguageARC platform. In order to make LanguageARC accessible to as wide a group of researchers as possible, we have created a Project Builder that allows users with no coding or software development experience to easily create and deploy annotation and collection tasks by uploading appropriately formatted data and answering a number of questions presented within a series of templates.

The Project Builder provides a series of ordered templates that takes the creator step-by-step through the build process from general information (e.g., project name, description) to specific task details (e.g., input data, tool features).

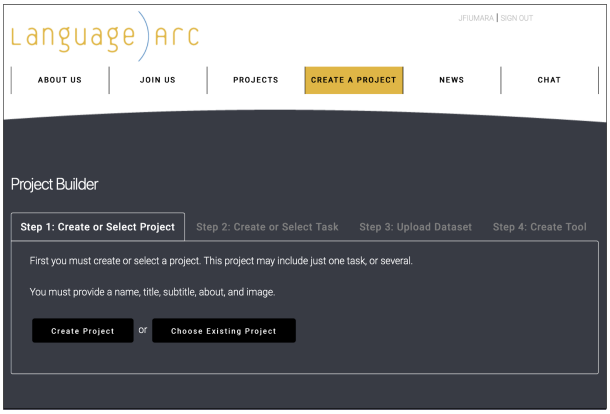

Figure 5: Project Builder menu

In Step 1 you can create a new project or select an already existing project to update. After the basic project information is created, Step 2 allows you to create a new annotation task or select a current task for updating. Each project must have at least one task, but may have multiple tasks within a single project. Task tutorials and reference guides can also be created with markup language and can include images, videos and audio clips.

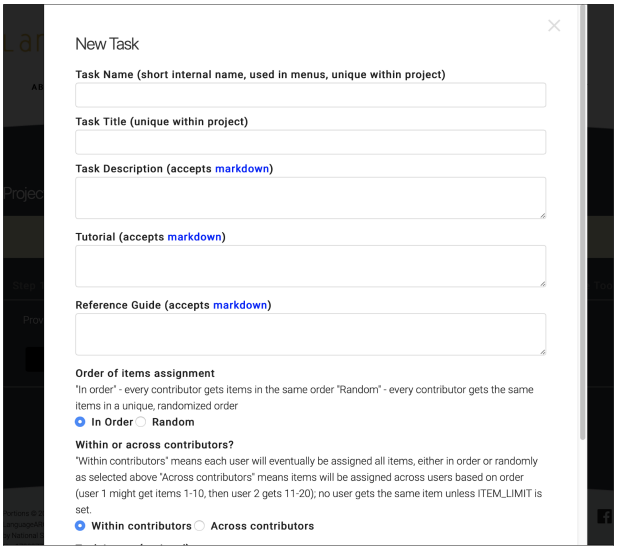

Figure 6: Task creation template

Step 3 in the Project Builder is to upload your input data (image, audio, video or text) and a tab delimited manifest file that orders and labels the input data. Finally, the last step in the Project Builder is to create the tool itself.

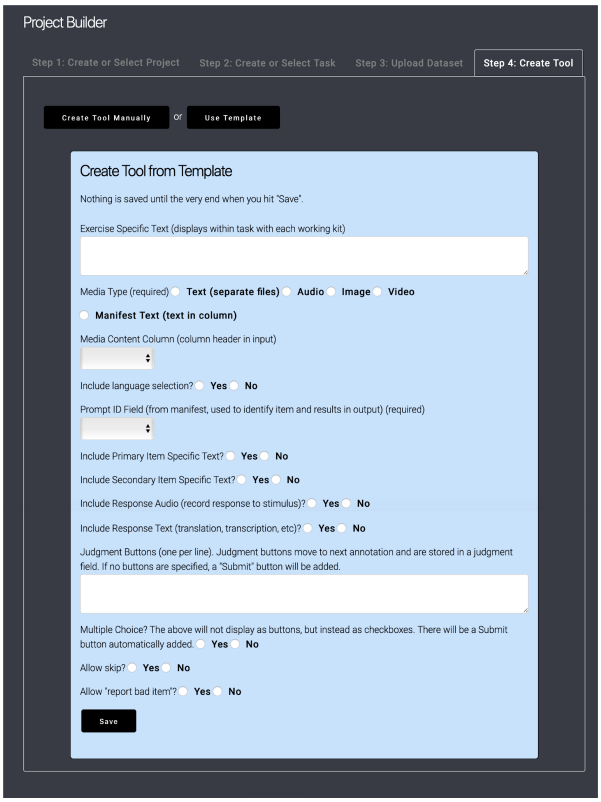

Figure 7: Tool builder template

Building the tool is also accomplished by answering a series of questions that tells the software what the input data is, which relevant data columns to select in the manifest, and what type of annotation interactions and outputs are desired.

Currently, the Project Builder is only available internally to LDC researchers. In the near future, the ability to create Projects will be available to the wider research community. Additional interactive instructions and guidelines for building projects will be included on the website. There will be a process where built projects will need to be approved prior to being made publically available.

Overall, the Project Builder has been designed so that with no coding knowledge required and just a small amount of prep work to prepare input data, projects and tasks can be created in as little as one hour or less.

5. Projects on LanguageARC

LanguageARC currently hosts a small number of projects created by the LDC and colleagues. Projects will be added on an ongoing basis and the number should grow exponentially once the Project Builder is made available to the larger research community. We will describe a few of the projects below to provide more in depth examples of the kinds of collection and annotation projects LanguageARC is capable of supporting.

5.1 From Cockney to the Queen

The project *From Cockney to the Queen* was developed in collaboration with researchers from the Linguistics department at the University of Essex. The goal of the project is to collect data and judgments to support sociolinguistic research into perceptions of regional accents in London and Southeast England. The project contains seven different tasks that ask citizen linguists to classify accents based on a variety of demographic information such as ethnicity, social class and geographic location.

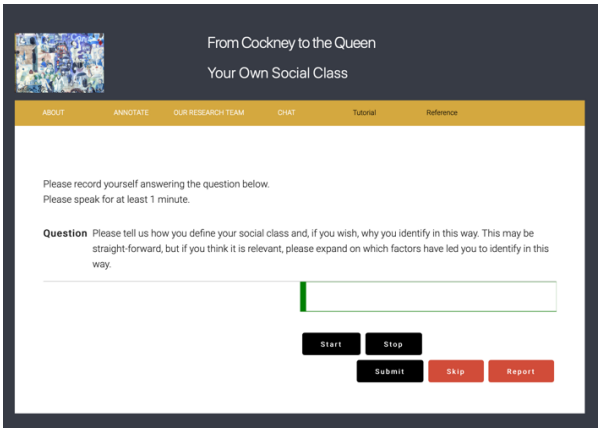

Figure 8: Speech recording task

Additional tasks allow contributors to provide their own experiences and definitions of these demographic features by uploading audio recordings. By using the audio player and audio recording widgets in the Project Builder Toolkit organized around multiple demographic features (ethnicity, social class and location), *From Cockney to the Queen* can collect large amounts of both judgments about accents and raw linguistic data.

5.2 Discovering Grammar Through Translation

The project, *Discovering Grammar Through Translation,* elicits translations from contributors to create bilingual data in English and the native language of the citizen linguist. Using the Elicitation Corpus created at Carnegie Mellon University's Language Technologies Institute,[11] this translation task includes contextual information to elicit translations that reveal grammatical features of languages such as gender, number and tense.

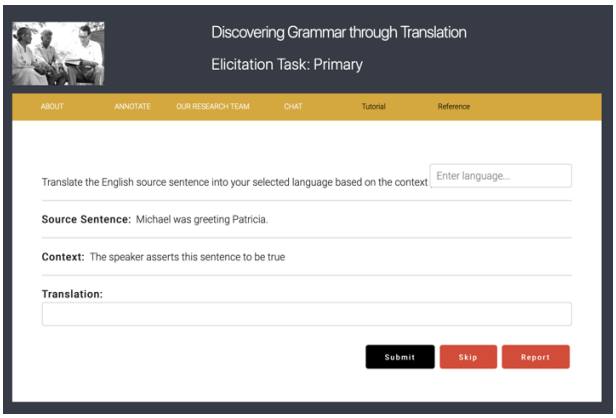

Figure 9: Translation task

The translation task requires that the contributor select a language for the task. The language selection box presents a scrollable list of languages containing all of the > 61,000 names used to refer to the world's 7400 languages with their ISO Language Code in parentheses. A source sentence for translation is provided along with contextual information to guide the translation. For example:

Source: Michael was greeting Patricia.
Context: The speaker asserts this sentence to be true.

Translations are entered into a text box and can be edited until the submit button is selected.

5.3 Clinical NLP Projects

The application of natural language processing to brain disorders such as autism spectrum disorder and frontotemporal degenerative disorders has shown great promise in increasing scientific understanding and clinical diagnosis (Cho et al. 2019, Parish-Morris et al. 2017). In order to identify and study the linguistic patterns and correlates of clinical conditions, researchers need extensive data from the general population to serve as a baseline for psychometric norming. LanguageARC can help collect these baseline datasets by creating tasks that mimic activities used in clinical settings allowing analysis of similar data across those with known clinical disorders and the general population.

5.3.1 Understanding Autism Spectrum Disorder

The Linguistic Data Consortium and the Center for Autism Research at Children's Hospital of Philadelphia[12] have been collaborating to develop LRs and apply human language technologies to the study of autism spectrum disorder (Parish-Morris et al. 2016).

The LanguageARC project *Understanding Autism Spectrum Disorder* asks contributors to complete two related tasks.

Figure 10: Understanding Autism Spectrum Disorder tasks

The first task asks the citizen linguist to answer the 50-questions Autism Quotient (AC) survey developed by the Autism Research Centre at Cambridge University.[13] While the AQ elicits self-report of traits associated with Autism Spectrum Disorders, LanguageARC's use of the instrument is not for purposes of individual diagnosis and no results are returned to contributors.

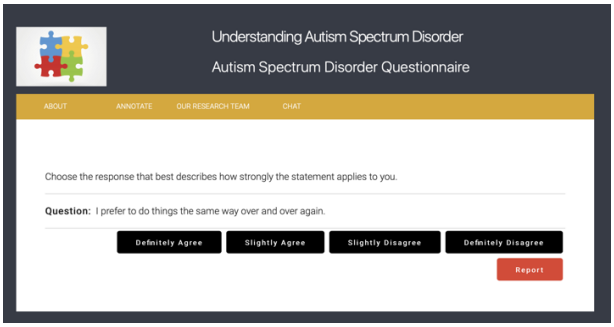

Figure 11: Questionnaire task

A second task asks contributors to complete a series of picture descriptions via an audio recording tool. Picture description tasks are commonly used in clinical settings. The combination of the two tasks allows the project to collect AQ results and corresponding linguistic data via the picture description from the overall general population allowing the creation of a large baseline dataset to assist in clinical research.

[11] https://www.lti.cs.cmu.edu
[12] https://www.centerforautismresearch.org
[13] https://www.autismresearchcentre.com

It should be repeated that these tasks designed for citizen linguists are not intended to provide diagnosis and do not provide test scores or feedback to the contributor.

5.3.2 The Dark Triad Survey

The Dark Triad Survey is a questionnaire used by psychologists to measure the personality traits of narcissism, psychopathy and Machiavellianism. As with the autism spectrum survey, this task is not intended as diagnostic and no scores are reported to the citizen linguist participants.

Similar to *Understanding Autism Spectrum Disorder*, the *Dark Triad Survey* project presents two tasks to the citizen linguist. The first is a 27-question survey used to measure dark triad personality traits and the second is a series of picture description tasks. The results will be aggregated with those of the other contributors to show how the whole population performs on these language tasks and provide data for investigating linguistic markers of personality type.

6. Project Reports and Recruitment

Project managers can access the output data collected through their tasks by selecting the report option within their user dashboard. Reports are tab delimited and contain details of every annotation made by users within the task including ID# to identify the project, task, and tool (which change if you update the task); a user ID and geographic location; the date and time of the annotation; and the content of the annotation if it is text entry or controlled vocabulary selections (i.e., button selections). For user annotations in audio format (such as picture description audio recordings) a separate download function is currently being developed.

LanguageARC requires the recruitment of two kinds of contributors: researchers and volunteer contributors. In this early phase of the project, LDC is both creating its own research projects and working with external colleagues to populate the portal with research projects. LDC has also been promoting LanguageARC in other venues likely to reach language researchers such as LREC and LinguistList. Building and sustaining a community of volunteer "citizen linguists" is perhaps an even bigger challenge. LDC is working to build its volunteer community by publicizing LanguageARC through a variety of venues and social media sites including advertising on Facebook and Twitter and promoting through related citizen science communities such as SciStarter.

7. Conclusion

LanguageARC uses novel incentives to address the limitations of current approaches to developing LRs that rely on project-constrained funding. By appealing to the motivations of citizen science, LanguageARC seeks to develop a community of citizen linguists and researchers working toward common goals. The powerful but easy-to-use Project Builder Toolkit and user friendly participant interface allows the creation of a wide variety of data collection and annotation tasks suitable for non-expert contributors. The data that results from projects and tasks developed with NSF funds will be made freely available to the research community.

8. Acknowledgments

The authors would like to acknowledge the support of the National Science Foundation under Grant No. 1730377.

9. Bibliographical References

Binnenpoorte, Diana, Catia Cucchiarini, Elisabeth D'Halleweyn, Janienke Sturm and Folkert de Vriend (2002) Towards a roadmap for Human Language Technologies: Dutch-Flemish experience in Proceedings of the workshop "Towards a Roadmap for Multimodal Language Resources and Evaluation" at LREC 2002, Las Palmas, Canary Islands, June.

Cho, Sunghye, Mark Liberman, Neville Ryant, Meredith Cola, Robert T. Schultz, Julia Parish-Morris (2019) Automatic detection of Autism Spectrum Disorder in children using acoustic and text features from brief natural conversations. Interspeech: 20th Annual Conference of the International Speech Communication Association Graz, September 15-19, 2019.

Cieri, C. (2017) Addressing the Language Resource Gap through Alternative Incentives, Workforces and Workflows, Keynote Speech at the 8th Language & Technology Conference, November 17-19, Poznań, Poland.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2019. Ethnologue: Languages of the World. Twenty-second edition. Dallas, Texas: SIL International. http://www.ethnologue.com.

Krauwer, Steven (1998) ELSNET and ELRA: Common past, common future, ELRA Newsletter, Vol. 3:2, May.

Krauwer, Steven (2003) The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap, in International Workshop Speech and Computer (SPECOM-2003).

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.

Parish-Morris, Julia, Mark Liberman, Christopher Cieri, John Herrington, Benjamin Yerys, Leila Batman, Joseph Donaher, Emily Ferguson, Juhi Pandey, Robert Schultz Linguistic Camouflage in Girls with Autism Spectrum Disorder. *Molecular Autism,* September 30, 2017

Parish-Morris, Julia, Christopher Cieri, Mark Liberman, Leila Bateman, Emily Ferguson, Robert T. Schultz (2016). Building Language Resources for Exploring Autism Spectrum Disorders. LREC: 10th Edition of the Language Resources and Evaluation Conference Portoroz, May 23-28, 2016.

Strötgen, J. and Gertz, M. (2012). Temporal tagging on different domains: Challenges, strategies, and gold standards. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pages 3746–3753, Istanbul, Turkey, may. European Language Resource Association (ELRA).