STREAMLInED Challenges: Aligning Research Interests with Shared Tasks

Gina-Anne Levow[1], **Emily M. Bender**[1], **Patrick Littell**[2], **Kristen Howell**[1], **Shobhana Chelliah**[3], **Joshua Crowgey**[1], **Dan Garrette**[1], **Jeff Good**[4] **Sharon Hargus**[1], **David Inman**[1], **Michael Maxwell**[5], **Michael Tjalve**[1], **Fei Xia**[1]
[1] University of Washington, [2] Carnegie Mellon University,
[3] University of North Texas, [4] University at Buffalo, [5] University of Maryland
levow,ebender,kphowell,jcrowgey,garrette,sharon,davinman,mtjalve,fxia@uw.edu
littell@alumni.ubc.ca, Shobhana.Chelliah@unt.edu, maxwell@umiacs.umd.edu

Abstract

While there have been significant improvements in speech and language processing, it remains difficult to bring these new tools to bear on challenges in endangered language documentation. We describe an effort to bridge this gap through Shared Task Evaluation Campaigns (STECs) by designing tasks that are compelling to speech and natural language processing researchers while addressing technical challenges in language documentation and exploiting growing archives of endangered language data. Based on discussions at a recent NSF-funded workshop, we present overarching design principles for these tasks: including realistic settings, diversity of data, accessibility of data and systems, and extensibility, that aim to ensure the utility of the resulting systems. Three planned tasks embodying these principles are highlighted: spanning audio processing, orthographic regularization, and automatic production of interlinear glossed text. The planned data and evaluation methodologies are also presented, motivating each task by its potential to accelerate the work of researchers and archivists working with endangered languages. Finally, we articulate the interest of the tasks to both speech and NLP researchers and speaker communities.

1 Introduction

It is a perennial observation at every workshop on computational methods (or digital tools) for endangered language documentation that we need to find a way to align the interests of the speech and language processing communities with those of endangered language documentation communities if we are to actually reap the potential benefits of current research in the former for the latter.

We propose that a particularly efficient and effective way to achieve this alignment of interest is through a set of "Shared Task Evaluation Challenges" (STECs) for the speech and language processing communities based on data already collected and annotated in language documentation efforts. STECs have been a primary driver of progress in natural language processing (NLP) and speech technology over several decades (Belz and Kilgarriff, 2006). A STEC involves standardized data for training (or otherwise developing) NLP/speech systems and then a held-out, also standardized, set of test data as well as implemented evaluation metrics for evaluating the systems submitted by the participating groups. This system is productive because the groups developing the algorithms benefit from independently curated data sets to test their systems on as well as independent evaluation of the systems, while the organizers of the shared task are able to focus effort on questions of interest to them without directly funding system development.

Organizing STECs based on endangered language data would take advantage of existing confluences of interest: The language documentation community has already produced large quantities of annotated data and would like to have reliable computational assistance in producing more; the NLP and speech communities are increasingly interested in low-resource languages. Currently, work on techniques for low-resource languages often involves simulating the low-resource state by working on resource-rich languages but restricting the available data. Providing tasks based on actual low-resource languages would allow the NLP and speech communities to test whether their techniques generalize beyond the now familiar small sample of languages that are typically studied (see Bender, 2011).

Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 39–47,
Honolulu, Hawai'i, USA, March 6–7, 2017. ©2017 Association for Computational Linguistics

In this paper, we present the design of three possible shared tasks, which together go under the rubric of STREAMLInED Challenges: Shared Tasks for Rapid Efficient Analysis of Many Languages in Emerging Documentation. These proposed tasks are the result of an NSF "Documenting Endangered Languages" (DEL) funded workshop bringing together researchers from the fields of language documentation and description, speech technology, and natural language processing. Our goals in describing these proposed shared tasks are to illustrate the general notion as a way forward in creating useful and usable technology, to connect with other members of the community who may be interested in contributing their data to the shared tasks when they are run, and finally to provide some information to working field linguists as well as language archivists about the size of data sets, type of annotations, and format of data and annotations that would be required to be able to take advantage of any systems that are published as the result of these shared tasks. In the following section (§2), we lay out the design principles we set forth for our shared tasks. In §3, we describe three shared tasks, relating to three different kinds of input data: audio data, data transcribed in varying orthographies, and interlinear glossed text (IGT). §4 briefly motivates the interest of the tasks to the speech and NLP research communities. Finally, in §5 we describe how the software systems created in response to these shared tasks, beyond their potential to benefit working field linguists, can also be of interest to the speaker communities whose languages are being studied.

2 Design Principles

A key goal of the EL-STEC workshop was to identify STECs that would actually align the interests of the communities involved. In discussing specific instances (including those developed in §3), we identified several design principles for our shared tasks:

Realism Whereas shared tasks in speech and NLP are often somewhat artificial, it is critical to our goals that our shared tasks closely model the actual computational needs of working linguists. It directly follows from this design principle that the software contributed by shared task participants should work off-the-shelf for stakeholders in documentary materials who are interested in using it later (e.g., linguists, speaker community members, archivists, etc.). This is ensured in our shared tasks by choosing a virtual environment for the evaluation, such as the Speech Recognition Virtual Kitchen (`http://speechkitchen.org/` Plummer et al., 2014), described in §3.1. Similarly, while some shared tasks artificially limit the data that participants can use to develop (or train) their systems, our shared tasks have a "use whatever you can get your hands on" approach, and we will explicitly point participants to resources that may be useful, such as ODIN (Lewis and Xia, 2010) or WALS (Haspelmath et al., 2008). Furthermore, in contrast to previous NLP and speech research which simulates low-resource languages by using only a small amount of data from actually well-resourced languages,[1] our STECs will be true low-resource environments, bringing in, we expect, complications not anticipated in work on familiar languages.

Typological diversity In order to facilitate the development of truly cross-linguistically useful technology, we will insist on typological diversity in the the languages used for the shared tasks. Specifically, we envision each shared task involving multiple development languages from different language families and instantiating different typological categories. These languages will be represented by data sets split into *training* and *development* data. Teams participating in the shared task will use the training data to train their systems and the development data to test the trained systems. By working with multiple languages from the start in developing their systems, shared task participants are more likely to create systems that work well across languages. To really test whether this is the case, however, each shared task will feature one or more *hidden* test languages, from yet other language families and typological classes. Of course, the shared task organizers will also provide training data for these hidden test languages (of the same general format and quantity as for the development languages), but no further system development will be allowed once that data is released.

Accessibility of the shared task The shared tasks must have relatively low barriers to entry, in

[1] There are some notable exceptions, including Xia et al's (2009) work on language identification in IGT harvested from linguistics papers on the web (Xia et al., 2009), the Zero Resource Speech Challenge (Versteegh et al., 2015), and the recent BABEL (Harper, 2014) program.

order to encourage broad participation. This can be ensured by having the shared task organizers provide baseline systems which provide working, if not necessarily high-performing, solutions to the task. The baseline systems not only establish the basic feasibility of the tasks, but also provide a starting point for teams (e.g. of students) who have ideas about how to improve performance on the task but lack the resources to create end-to-end solutions.

Accessibility of the resulting software We wish to ensure that the research advances achieved during the shared tasks are accessible to all of the constituencies we have identified: NLP and/or speech researchers, linguists working on language documentation, and speaker communities. This in turn means that the software systems submitted for evaluation should be available at a reasonable cost (ideally free) to third parties and reasonably easy to run. Furthermore, the systems should be well documented in papers published in the shared task proceedings, such that future researchers could re-implement the algorithms described and verify their performance against published numerical evaluation results on the development languages.

Extensibility From the point of view of the speech/NLP research communities, one of the merits of shared tasks is the establishment of standard data sets and published state of the art results on those data sets against which future research can be compared. There is a paucity of typologically diverse multilingual data sets. The initial data sets (including both development and test languages) for our shared tasks will immediately become a resource that addresses this need, but we would like these resources to grow over time. Accordingly, each shared task will include documentation of what is required for a new data set to be added and welcome submission of such data sets. On in-take, the existing available systems can be run on the data sets to establish baseline numbers for comparison in future work.

Nuanced Evaluation Rather than having one single metric (such as might be required to anoint a single "winner" of each shared task), the shared tasks will have multiple metrics to allow for nuanced evaluations of the relative strengths and weaknesses of each submitted system.

Having articulated these design principles, we now turn to the explanation of our three proposed shared tasks.

3 Proposed Shared Tasks

In this section, we briefly outline three proposed (families of) shared tasks which we believe to meet the design principles described in §2 above. For each shared task, we describe the input data and output annotation (or other information), the proposed evaluation metrics, and information about required data formats.

3.1 Grandma's Hatbox

The process of documenting an endangered language often begins with recordings of elicitations between the linguist and their consultant. Work continues through transcription, alignment, analysis, and glossing. Each of these steps is time-consuming. As a result, a kind of funneling occurs where less data can be analyzed in as great of a level of detail at each stage in the pipeline. Some recordings may never be transcribed or aligned by the linguist due to insufficient time or resource or even a shift in research priorities. We have defined a cascade of shared tasks focused on processing audio recordings to facilitate the transcription and alignment process, to widen and speed up this pipeline. The tasks further aim to develop technology that would make a backlog of unanalyzed recordings more accessible to other linguists, to archivists, and to members of the community through automatic extraction of information about the languages, participants, genre, and content of the recordings.

We name our cascade of shared tasks "Grandma's hatbox" to describe the sequence of steps to follow when a collection of field recordings is found with little associated information: imagine a box of telegraphically labeled tapes and field notebooks. The specific subtasks are defined below.

1. Language Identification

   High-resource (HRL) versus Low-resource language (LRL): Given the identities of the high-resource language, such as English, and the low-resource language used in the recordings, participants should produce a segmentation of the speech in the recordings, identifying the points at which transitions to and from the HRL occur.

All languages: Given a recording with a known HRL and an unspecified number of unknown LRLs, identify the number of distinct languages being spoken, the time points of transitions between languages, and which of the unknown languages is being spoken in each interval.

Training and test data will include audio files in .wav format with fine-grained language segmentation consistent with the subtasks above for: a) one HRL and one LRL, and b) a single HRL and at least two LRLs. In the case of overlapped speech, both (all) languages should be labeled. To score the results, we will use a language-averaged time-percentage F1-score (a standard measure that balances between precision and sensitivity of classification). This metric overcomes the weakness of a simple accuracy measure, which could perform well simply by always selecting the majority language.

2. Speaker identification

   Speaker clustering and segmentation: Given a collection of audio files with multiple speakers and multiple languages, assign a unique speaker code to each speaker and label all times when each speaker is talking.

   Known speaker segmentation: Given spans of speech labeled with speaker identity for one or more speakers, identify all other spans where each of the known speakers talks.

   Training and test data will be provided with speaker labeling and segmentation. To allow focus on both frequent and infrequent speakers, we will employ two evaluation metrics: F1 scores averaged across all time segments in the corpus and all speakers in the corpus, respectively. As above, in the case of overlapped speech, both (all) speakers should be labeled.

3. Genre classification
   For a span of audio, identify which of a fixed inventory of genres, e.g., elicitation, monolog, LRL dialog, reading, or chanting, is present. This inventory will be provided by the organizers along with the training data. Training and test samples will be provided for the range of genres and scored using F1 score.

4. Automatic metadata extraction by HRL Automatic Speech Recognition (ASR)
   Given a recorded preamble to an interview in an HRL, identify key metadata for the recording, including: date, researcher name, consultant name, and location. Task participants will only have access to the audio itself, with no further metadata. Training and evaluation data will comprise at least 500 examples each per HRL along with corresponding metadata templates, specifying slots and fillers. Evaluation will be based on slot filling accuracy: #(correctly filled slots)/#(total slots).

5. Transcription alignment
   Given noisy, partial text transcripts of 2-3 sentences in length, find the time alignment of the text to the recorded speech. Transcripts may include a mix of HRL and LRL-specific orthography. Training and evaluation data will include transcribed spans across a representative range of orthographic conventions with corresponding time alignments to audio files. This task will be evaluated based on absolute time difference between hypothesized and true boundary times.

For each of these tasks, the organizers will create a baseline system, from input to evaluation, to be distributed to the participants. These baseline systems and participants' submitted systems will all be provided as "virtual machines", encapsulated computing environments containing all the required components to run the systems, which can then be deployed in most common operating system environments. Using this framework for speech systems has been promoted by the Speech Recognition Virtual Kitchen team (Plummer et al., 2014) to deploy speech recognition components and systems in educational settings. This will allow ready system comparison and a natural path to deployment on multiple platforms.

The results of each of these cascading tasks can feed into subsequent stages and into the enrichment of the audio archive. General metadata will be stored in a template, including language type and quantity information, speaker number and quantity information, as well as genre, speaker names, recording dates, and so on. Time span information for language and speaker segmentation

as well as alignment can readily be encoded in formats readable by tools such as ELAN[2] (Brugman and Russel, 2004), which has seen increasing adoption for endangered language research and which provides an easy visual interface to time-aligned speech and language annotations.

The techniques developed through the "Grandma's hatbox" ensemble of shared tasks have potential benefit for field linguists, researchers in endangered languages, and language archivists. For those collecting data, these techniques can accelerate the process of transcription and alignment of speech data. They can also facilitate consistent metadata extraction from recordings, including language and speaker information, recording dates and locations, as well as genre. These techniques can likewise be applied to existing archive data or ingestion of new materials, for which detailed metadata or time-aligned transcription were unavailable or as a means of testing the quality of existing metadata and transcriptions. This information can be provided in standard formats consistent with best practices established by the community.

3.2 Orthographic Regularization

One of the central problems facing endangered-language technology is the lack of substantial and consistent text corpora; what texts exist are often written in a variety of orthographies (ranging from the scientific to the naïve), written by transcribers of a wide range of expertise, and frequently written by speakers of significantly different dialects. Only a few endangered languages have a sufficient corpus of expert-transcribed text to enable conventional high-resource text technologies; more often, the number of writers trained in a regular orthography is quite small. Similarly, for any such technology that takes text input, there are often only a small pool of potential users who can form inputs in the particular orthography the system expects.

Any text technology for an endangered language must, therefore, be prepared to work with inexpert and approximate transcriptions, transcriptions using variant orthographies, and even transcriptions in which there is no systematic orthography at all.

We therefore propose an orthographic regularization shared task in which systems are provided a set of text passages in a single endangered language, including both expert transcriptions in a systematic orthography and a variety of inexpert transcriptions. These variant transcriptions can range from attempts at formal orthography from inexpert transcribers (e.g., student work), to historical transcriptions, to dialectal variants, to renderings by writers without any background in a formal orthography (what is sometimes called "naïve transcription"). The task of the system is to normalize the variant transcriptions into their correct orthographic forms, and the results will be judged on a held-out set of text passages for which both expert and inexpert transcriptions are available.

The Orthographic Regularization shared task will have three evaluation conditions, which differ with respect to the presence or absence of parallel transliterations, metadata, and audio.

T1. Only mono-orthographic material is available; no parallel data or metadata can be used.

T2. Parallel data (although likely in small amounts) is available.

T3. Metadata (author ID, date of composition, etc.) and/or audio recordings are available.

The division into three conditions is to stimulate the development of systems that train on a comparatively bare minimum of material, while not discouraging the development of systems that make use of wider (although still commonly available) resources.

Text data will be provided in UTF-8 text format, metadata in JSON format, and audio recordings in WAV format. Data will further be organized based on the task conditions they are permissible in.

The system is tasked with producing a normalization of each test text into each regular, named orthography. That is, given a file in an irregular orthography and one (or more[3]) regular, named orthographies appropriate for the language, the system should produce an output file for each regular, named orthography. System output files will be compared to gold-standard, held-out texts in the desired orthographies, corresponding to the test passages. The evaluation metric will be character

[2] http://tla.mpi.nl/tools/tla-tools/elan/

[3] Some endangered language communities have several competing "official" orthographies. When multiple such orthographies exist and text is available in each, the goal will be to normalize into each orthography, to avoid the appearance of judging one orthography as "correct" and the other as non-standard.

error rate, based on the number of insertion, deletion, and substitution errors at the character level in the system output relative to the gold-standard texts. Only a subset of generated documents may be evaluated, depending on the availability of parallel gold-standard texts in each of the regular, named orthographies.

Systems will be submitted as containers (such as a Docker container or a similar service), and will be immediately available for use by community members via a web interface.

3.3 First-pass IGT Production

Our final proposed shared task concerns the production of interlinear glossed text (IGT) on the basis of transcribed, translated text. That is, we assume a workflow by which linguists collect spoken texts, transcribe them, elicit translations from the speakers they are consulting, and then work on producing IGT, including segmenting the words into morphemes and glossing each morpheme. Given this workflow, it is typical for a given field project to produce more transcribed texts than translated texts and more translated texts than glossed texts. The goal of this task is to even out the last two categories–that is, to create more glossed texts from translated texts.

For this shared task, we will provide for each development language a collection of at least 500 fully glossed IGT instances (typically sentence-like units, as segmented by the transcriber), plus whatever other materials are available for the language. In addition, there will be another 500 IGT instances designated for evaluation. In this shared task, we are assuming that the goal is to produce five-line IGT, where the first line represents the instance in some standard orthography or broad IPA transcription, the second segments the line into morphemes[4], the third glosses each morpheme, and the fifth provides a translation to a language of broader communication. In addition to these relatively standard lines, we also anticipate a fourth line which gives "word glosses". These are phrasal representations of the grammatical and lexical information provided in each source language word that are sometimes produced by linguists as a shorthand and are valued by speaker communities engaged in language revitalization, as they are far more approachable than linguist-oriented glosses. An example for Nuuchahnulth (nuk) is given in (1).

(1) hayimḥ q^{w}icičƛii
hayimḥa q^{w}i-ci-čiƛ-ii
not.know what-go-MO-WEAK.3
not know where she went
'They did not know where she had gone'

The participating teams will develop systems that can be "trained" on the data for a given language, and then produce first-pass segmentation and glossing (both standard glossing and "word glossing") on further data, given as input transcription and translation. The expectation is not that such automatically produced glosses would be perfect, but rather that the first pass glossing, even if somewhat incorrect, will still be useful. For example, it could be good enough that correcting the glosses is faster than doing them by hand or that the automatically produced glosses facilitate searching the relatively unanalyzed portion of the corpus for examples of phenomena of interest.

For each development language, the shared task organizers will provide 500 more instances of IGT designated as "development test" data. System developers can use this data to check system performance, by passing in the transcription and translation lines, and comparing the output segmentation and gloss lines to the "gold standard" to gauge system performance, perform error analysis, and determine how to improve their systems.

The final shared task evaluation will involve one or more hidden test languages. Each participating system will be trained on the data (at least 500 instances of IGT, plus whatever else is available) from the hidden test language(s) and tested against linguist-provided annotations for 500 test instances of IGT per language. Both development and test language data will be formatted in Xigt (Goodman et al., 2015), and the shared task organizers will provide converters between Xigt and formats such as Toolbox[5], FLEx[6], or Elan (Brugman and Russel, 2004) so that linguists can use the resulting systems with their own data.

In selecting development and test languages for this task, we will look for morphologically

[4] Depending on the analytical style of the linguists producing the data and the traditions for that language area, this line might have one canonical 'underlying' form for each morpheme, or it might allow different allomorphs. Participating systems will be expected to reproduce the style of the input data.

[5] http://www.sil.org/resources/software_fonts/toolbox

[6] http://www.sil.org/resources/software_fonts/flex

complex languages, but attempt to find typological diversity along dimensions such as prefixing/suffixing and agglutinating/fusional, as well as language family and areal diversity. To serve as a development of test language for this shared task, a project would need at least 1000 fully glossed instances of IGT for that language. For the resulting software to produce useful output to the linguist, the glossed IGT should be representative of what else is in the text (e.g., if the text is mostly transcribed narratives, it is important for the training IGT to include a good sample from narratives).

This task differs from what is already accomplished by the glossing assist function in FLEx (Baines, 2009) in several ways. First, where FLEx produces all possible analyses, the systems participating in this shared task will be asked to choose from among possible outputs the one deemed most likely (on the basis of the training data). Second, where FLEx typically assumes "surface-true" segmentation for the morpheme-segmented line, systems participating in this shared task will be expected to produce underlying forms if that is what is provided in the training data. Finally, where FLEx requires direct input from the linguist if it is to have information about constraints such as affixes only attaching to particular parts of speech, it is anticipated that participating systems will pick this information up from the training data.

4 Intellectual Merit: Research Interest in Speech/NLP

All three of our proposed shared tasks not only solve problems of relevance to field linguists, they also carry inherent research interests for speech and NLP researchers. All three tasks share the properties that they produce data sets and benchmarks to allow researchers to test whether their proposed language-independent solutions work across a broad range of language types. Furthermore, they allow researchers to explore truly low-resource scenarios. These contrast with the typical simulated low-resource scenarios in that the latter involve decisions about which data to keep, and this might not be representative of what an actual low-resource situation might be like. Each task has additional inherent research interest of its own, as detailed below.

The "Grandma's hatbox" shared task suite spans a range of speech processing technologies, including language identification, speaker identification, slot filling, and alignment. Shared task regimes exist for some of these broad areas, such as the NIST speaker (NIST, 2016) and language recognition (NIST, 2015) tasks. The slot filling task also bears some similarities to spoken dialog system tasks, such as the Air Travel Information System (Mesnil et al., 2013) task and components of the Dialog State Tracking Challenge tasks (Williams et al., 2016). However, the setting of endangered language field recordings poses new and exciting challenges, while leveraging techniques developed for other languages in high resource settings. In addition to using languages and language families not typically used in the classic tasks, the recording conditions and audio quality differ from those in typical controlled settings. Both language and speaker segmentation must operate over short, possibly single-word spans, a finer granularity than even 2 second train/test conditions in some tasks (NIST, 2016). Furthermore, these recordings can contain substantial fine-grained code-mixing, with individual speakers talking different languages, and may attract interest from a growing community interested in code-switching in text and speech. The slot filling task will operate over less-structured human-directed speech, rather than the computer-directed speech prevalent in dialog systems tasks listed above. Finally, the alignment task requires not only noisy, partial, multilingual alignment, but alignment over non-standard orthographies. These new challenges will push the state of the art in these speech processing tasks.

The orthographic regularization shared task builds on other work on orthographic regularization in widely spoken languages (see, for example (Mohit et al., 2014; Rozovskaya et al., 2015; Baldwin et al., 2015) on social media text and Dale and Kilgariff (2011) on text produced by language learners), but pushes the frontiers of work in this area in several ways: While this proposed shared task has much in common with these previous shared tasks, endangered language text normalization poses additional interesting problems. In languages like English or Arabic, there is usually a single, established orthography in which almost all users have formal schooling and extensive digital corpora in this orthography that establish "correct" practices. Endangered languages often only have small amounts of material available, often non-normalized and/or in conflicting orthographies; there may be more material avail-

able in need of normalization than there is material that establishes correct practices. On the other hand, there are fewer individual authors, meaning that author identification can potentially lead to greater gains, and supplementary material like audio is likely to be available for at least some of the texts (because much endangered language text is transcribed from audio recordings).

The first-pass IGT production shared task resembles earlier shared tasks on morphological analysis, most notably the Morpho Challenge series (Kurimo et al., 2010). It differs, however, in working with words in context (rather than word lists), and in going beyond segmentation of words into morphemes to associating morphemes with particular glosses. The presence of the translation line also provides a new source of information in producing the glosses, not available in previous shared tasks. Finally, the task of producing word-glosses is a novel one, with connections to low-resource machine translation.

5 Broader Impacts: Benefits to Speaker Communities

Beyond helping with the project of endangered language documentation, the shared tasks described here all also hold potential interest for speaker communities, especially those interested in language revitalization.

The techniques developed through the "Grandma's hatbox" ensemble of shared tasks will allow more rapid and automatic extraction of information describing the content of recordings. By providing easy access to information about the languages, speakers, and types of recorded materials, they will make such recordings more accessible to speaker communities. This automatically extracted information will allow simple search and navigation within and across recordings based on language, speaker, genre, and even content, through aligned transcriptions, allowing speaker communities to more easily engage with recorded materials.

The orthographic regularization shared task will produce technology which, in our experience, is among the most requested and most used among endangered-language communities. Many communities have collections of texts in heterogeneous orthographies, and writers have often been trained in different orthographies (and trained to varying degrees), so the possibility of normalizing texts (both old and new) to a consistent format can solve many practical problems communities face.

At best, such technologies can even help to diffuse "orthography conflicts" between dialects, regions, schools, or generations. For example, as several students of the SENĆOŦEN language told one of the authors, their parents' generation (the last generation of fully fluent speakers) had been taught a particular orthographic tradition, and since that time their schools have adopted a different orthography, developed within (and preferred by) the community. The two orthographies are visually quite different, and students and parents therefore have difficulty writing to each other in their language. Technology that could render the students' writing into their parents' orthography (and meanwhile correct some student errors), or render their parents' writing into the students' orthography, would better enable the kind of intergenerational collaboration that the students need to learn and preserve their language.

Of the outputs provided by the first-pass IGT production shared task, the word glosses are anticipated to be the most interesting to speaker communities. This style of information presentation is much more accessible to language learners than glosses produced for linguists, and the ability to produce it automatically for additional texts will facilitate the development of language learning materials as well as making otherwise inaccessible texts into objects of interest for language learners.

6 Conclusion: Next Steps

We have described how shared task evaluation challenges can be used to align the research interests of the speech and natural language processing communities with those of the language documentation and description community and articulated design principles for creating shared tasks that achieve this goal. In addition, we have described three particular shared tasks which we believe to meet those design principles. The next steps are to secure funding to actually run one or more of these shared tasks as well as getting them accepted to appropriate venues and to solicit data collections, either from active language documentation projects or from language archives to use as development and test data sets in these tasks.

Acknowledgments

We are also grateful for the contributions of Mark Hasegawa-Johnson, Russ Hugo, Jeremy Kahn, Lori Levin, Alexis Palmer, and Laura Welcher, during the EL-STEC workshop. This work has been supported by NSF #: 1500157. Any opinions expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

David Baines. 2009. Fieldworks language explorer (FLEx). *eLEX2009*, page 27.

Timothy Baldwin, Young-Bum Kim, Marie Catherine de Marneffe, Alan Ritter, Bo Han, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. *ACL-IJCNLP*, 126:2015.

Anja Belz and Adam Kilgarriff. 2006. Shared-task evaluations in HLT: Lessons for NLG. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 133–135, Sydney, Australia. Association for Computational Linguistics.

Emily M. Bender. 2011. On achieving and evaluating language independence in NLP. *Linguistic Issues in Language Technology*, 6:1–26.

H. Brugman and A. Russel. 2004. Annotating multimedia/ multi-modal resources with ELAN. In *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*.

Robert Dale and Adam Kilgarriff. 2011. Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, ENLG '11, pages 242–249, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael Wayne Goodman, Joshua Crowgey, Fei Xia, and Emily M. Bender. 2015. Xigt: Extensible interlinear glossed text. *Language Resources and Evaluation*, 2:455–485.

M. Harper. 2014. IARPA BABEL Program. `http://www.iarpa.gov/Programs/ia/Babel/babel.html`. Accessed September 2014.

Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, editors. 2008. *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich. http://wals.info.

Mikko Kurimo, Sami Virpioja, Ville Turunen, and Krista Lagus. 2010. Morpho Challenge competition 2005–2010: Evaluations and results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 87–95. Association for Computational Linguistics.

William D. Lewis and Fei Xia. 2010. Developing ODIN: A multilingual repository of annotated language data for hundreds of the world's languages. *Journal of Literary and Linguistic Computing*, 25:303–319.

Gregoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Interspeech 2013*.

Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. 2014. The first QALB shared task on automatic text correction for Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 39–47.

NIST. 2015. 2015 Language Recognition Evaluation Plan. `https://www.nist.gov/file/325251`. Downloaded October 8, 2016.

NIST. 2016. 2016 NIST Speaker Recognition Evaluation Plan. `https://www.nist.gov/file/325336`. Downloaded October 8, 2016.

Andrew Plummer, Eric Riebling, Anuj Kumar, Florian Metze, Eric Fosler-Lussier, and Rebecca Bates. 2014. The Speech Recognition Virtual Kitchen: Launch party. In *Proceedings of Interspeech 2014*.

Alla Rozovskaya, Houda Bouamor, Nizar Habash, Wajdi Zaghouani, Ossama Obeid, and Behrang Mohit. 2015. The second QALB shared task on automatic text correction for Arabic. In *ANLP Workshop 2015*, page 26.

Maarten Versteegh, Roland Thiolliere, Thomas Schat, Xuan Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. 2015. The zero resource speech challenge 2015. In *Proceedings of Interspeech 2015*, pages 3169–3173.

Jason D. Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Fei Xia, William Lewis, and Hoifung Poon. 2009. Language ID in the context of harvesting language data off the web. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 870–878, Athens, Greece, March. Association for Computational Linguistics.