EL-STEC: Shared Task Evaluation Campaigns with Endangered Language Data

1 Introduction

Language description and documentation efforts are currently at a critical stage. There are probably not enough linguist-hours left to document all of the languages expected to disappear by the end of this century. At the same time, recent advances in natural language processing and speech technology hold the promise of being able to automate many of the more repetitive tasks taken on by field linguists, greatly speeding up the process of documenting a particular language and allowing more thorough description of more languages before they lose their last fully fluent speakers.

The recent NSF-sponsored workshop at the 2014 meeting of the Association for Computational Linguistics ("The Use of Computational Methods in the Study of Endangered Languages") highlighted once again the need to find a way to align the interests of the speech and language processing and endangered language documentation communities if we are to actually reap the potential benefits of current research in the former to 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 & Kilgarriff, 2006). A STEC involves standardized data for training (or otherwise developing) NLP/speech systems and then a held-out, also standardized, set of 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 amounts of annotated data and would like to have reliable computational assistance in producing more; the NLP and speech communities are currently very 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 actually 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).

The goal of this project is to organize some STECs (at least one involving speech and one that starts from data that is already transcribed) that (i) are based on existing data from language documentation projects, (ii) push the state of the art in technology that could be developed into tools that are truly useful to field linguists and responsive to where

they find bottlenecks in their work-flow and (iii) engage the interest of the NLP and speech research communities. Our process will begin with surveying the existing data in endangered language archives (§3). Next we will design our shared tasks (§4–6), considering both what is possible (aligning amounts and types of data to the requirements of modern systems) and what is useful to field linguists (as worked out in a small workshop hosted at UW). Finally, we will organize and run the shared tasks in conjunction with major conferences in NLP and speech technology (§7).

2 Background: Shared Task Evaluation Campaigns

Shared task evaluation campaigns began their rise to prominence in the natural language processing and speech research communities in the 1980s and 90s. The campaigns of this era include the Resource Managment (RM) (Price et al., 1988) and Air Travel Information Systems (ATIS) (Dahl et al., 1994) tasks in speech recognition and spoken dialog systems, respectively; the Message Understanding Conference (MUC) series (Grishman & Sundheim, 1996) on information extraction; and the Text Retrieval Conference (TREC) (Voorhees & Harman, 2005) in information retrieval and related technologies. These campaigns came about to address serious issues in evaluation and comparability in prior research paradigms. Prior to the development of these shared tasks, comparisons across systems designed for the same tasks were hampered by, among other issues, differences in type and amount data used for system development, differences in type of data used for evaluation, use of proprietary data in system development, and lack of standardized metrics for evaluation. As a result, it was almost impossible to obtain a direct comparison of the effectiveness of different techniques, and it was further profoundly difficult to replicate the results of such work. Furthermore, as resource-intensive machine learning techniques became prevalent, the need for large annotated data sets became a significant bottleneck for research groups seeking to develop and test systems.

The introduction of shared task evaluation campaigns enabled significant strides in the comparability and replicability of research, while lowering the barriers to entry in these areas. In addition to providing an emphasis on systematic quantitative evaluation, the STECs ensured standardization and availability of several key experimental elements:

- The detailed task specification itself: What was a system supposed to do? What inputs did it expect? What outputs should it produce?
- Training and development data: What resources could and should be used in system development?
- Held-out test data: What would the answer key or 'gold standard' results look like? Furthermore, test data would be reserved so that evaluation would be performed on previously unseen items.
- Evaluation metrics and tools: How should the results of the system be scored? What

factors should be considered in assessing a system for a given task? A reference implementation of the scoring system would be provided.

- Timeline: Training, testing, and evaluation material would be released to all participants at the same time.

All of these factors contributed to a level playing field for participants and equitable comparisons of systems.

Other key properties of these STECs also contributed to their popularity and impact. First, resources, both the specific training/testing/evaluation materials and other supporting resources would be made available to the community for the task (and subsequent experimentation if possible) under varying license agreements. Secondly, most shared tasks required, as a condition of access to the data, public reporting of the techniques employed in the systems. As a result, not only could one perform fair comparison across a multitude of systems, but one could readily identify novel effective techniques which would then spread through the scientific community. Successful techniques showcased in shared tasks came to dominate speech recognition, information retrieval, machine translation, and speaker recognition for years. Furthermore, although early STECs were predominantly developed by funders to assess fundees, recently STECs have expanded to allow proposals of tasks by interested individuals or groups. Thus, the tasks can serve as venues to build community around particular tasks or research areas.

Currently, dozens of STECs are held each year involving hundreds of participating research teams. The tasks are highly diverse, ranging from basic speech recognition to information extraction to word segmentation of Chinese microblog data (similar to Twitter). These STECs provide access to data for participants, standardized and comparable evaluations, sharing of research results, and focus and building of communities around research tasks. As a result, they have become major drivers and enablers of research in speech and language processing.

3 Surveying Existing Annotated Data

Our first step in the process of crafting STECs around endangered language documentation is to carry out a thorough census of the data that currently exists in endangered language archives, including ELAR, PARADISEC, ANLA and others. For each language we consider, we will catalog which types of data are available, which types of annotations over that data are available, glossing, phoneme inventories, metadata about speakers, and any other types of annotation we encounter. Some examples are provided in Table 1. For all data and annotations we will also catalog the amount available.

For a STEC to have lasting and broad impact, the data must be available to be freely distributed. If it is, then even past the official STEC, research groups can come back to the problem and try to improve on the state of the art. Accordingly, we will first seek data from languages that is already available for unrestricted distribution. If we do not find sufficient

Type of Information to Catalog	Details
Type of data	Audio, Text
Recording conditions	Single-/multi-speaker, noise levels
Type of annotation	Transcription of audio; Translation into English, Spanish, French, Russian of other languages of wider communication
Glossing	Lemmas only; or lemmas as well as most grammatical morphemes
Amount of data/annotation	Hours, words, utterances
Phonology	Phoneme inventories
Speaker metadata	Number of speakers, gender, age, other languages spoken
Language metadata	ISO 639-3 code, endangerment status, language family

Table 1: Types and examples of information to be catalogued in archive survey

data of this type, we will then work with field linguists to see whether any data can be "freed".[1]

4 Designing Shared Task Evaluation Challenges

4.1 What Is Possible?

Machine learning techniques have enabled the development of robust, highly effective systems for a wide range of natural language and speech processing tasks. The majority of these techniques exploit supervised machine learning, where ground truth labels paired with data instances are required for classifier training, and classifier performance typically increases as the amount of labeled training data increases. This requirement for labeled training data, however, creates a serious knowledge acquisition bottleneck. Such labeled data is often costly to acquire both in terms of time and money, limiting annotation efforts and the availability of such resources to languages and tasks for which there is substantial financial support.

To overcome this resource acquisition bottleneck, researchers have investigated two main strategies: reducing the need for labeled data through unsupervised and semi-supervised machine learning algorithms and applying domain adaptation techniques to exploit other labeled data. In each of these cases, application to endangered language data provides great opportunities but also poses additional challenges. These additional challenges must be addressed directly for the technology to be of use to endangered language documentation. They also contribute to the interest of the problems for machine learning researchers.

Semi-supervised and unsupervised methods, machine learning approaches that require little or no labeled training data respectively, have demonstrated great promise in addressing

[1]We realize that this will depend on the reasons for the restrictions of access to the data. If the community of speakers from which it was recorded do not wish it to be broadly disseminated, then that must be respected.

the knowledge acquisition bottleneck. A wide range of such techniques have been developed and have demonstrated effectiveness in diverse speech and language processing tasks. As examples, unsupervised approaches have received much attention in part-of-speech tagging and induction as surveyed in (Christodoulopoulos et al., 2010), morphological segmentation (Creutz & Lagus, 2007; Poon et al., 2009), word sense clustering and disambiguation (Lin, 1998; McCarthy et al., 2004; Brody & Lapata, 2008), and topic segmentation (Hearst, 1994; Choi et al., 2001; Matveeva & Levow, 2007), among others. Semi-supervised methods have also been applied to important problems including speech recognition (Hsiao et al., 2013; Tusk et al., 2014; Gales et al., 2014), prosodic event recognition (Levow, 2006; Jeon & Liu, 2009), and parsing (Koo et al., 2008; Druck et al., 2009), and many more.[2] Most experiments in unsupervised and semi-supervised learning, however, have simulated the low-resource setting by artificially impoverishing datasets from resource-rich languages and demonstrating improvements in performance at different points on the learning curve. These approaches further typically assume the availability of a large unlabeled dataset, even when labeled data is severely limited or unavailable. In the case of endangered language data, the availability of labeled data is intrinsically and truly limited. In addition, the absolute amount of data, labeled or unlabeled, may be small.

An alternative framework for approaching low-resource language tasks is through adaptation. Even outside true low-density or endangered language settings, low-resource models have drawn significant interest. Even if a particular language is resource-rich, new domains or new tasks may be viewed as low-resource. For example, English is the canonical example of a resource-rich language, but if one wants to build a system for automatic transcription and searching of university-level lectures, there may be few dedicated resources for a given field of study. In general, adaptation approaches build upon a model created based on a resource-rich language or task that is then retuned or adapted to the new low-resource task. Approaches differ primarily in the mechanism used to adapt to the new setting, and many different models have been proposed.

A common strategy is to adapt the original model by adding a small amount of labeled data from the low-resource task to the original training data. Simple augmentation with such data forms a standard baseline adaptation technique (Daumé III, 2009). More sophisticated adaptation techniques may exploit linguistic similarities between the original training data and the new task setting, such as similarities in phonetic, lexical, syntactic, or semantic structure, depending upon the task. Structural correspondence learning (Blitzer et al., 2006) is one such model for unsupervised adaptation and has been used for adaptation across text domains for part-of-speech tagging of biomedical text (Blitzer et al., 2006), sentiment analysis for different products (Blitzer et al., 2007), and dialog labelling in conversational speech (Margolis et al., 2010). Other approaches used in speaker and speech recognition involve computing transformations on features drawn from different domains (Anastasakos et al., 1996; Dehak et al., 2011; Parrish & Gupta, 2012). Extensions of such approaches will likely be applicable to tasks in documenting endangered languages as well.

[2] `http://aclweb.org/aclwiki/index.php?title=Semi-supervised_Learning_in_NLP`

4.2 Low-resource and "Surprise Language" Tasks

The speech and language processing communities have created a number of shared tasks centered on low-resource efforts. Beyond tracks of specific tasks which stipulate unsupervised or semi-supervised methods as described above, two such general settings are "surprise language" and low-resource language tasks.

In surprise language tasks, research groups that have been developing general purpose techniques for a particular problem are presented with a previously unannounced language (not considered in system development), typically with minimal resources, and asked to develop a system for that language on a short timeline. One well-known example of a surprise language task was the DARPA-sponsored surprise language dry run (and later full exercise) held in 2003 as part of the Translingual Information Detection, Extraction, and Summarization (TIDES) program (Oard et al., 2003; Oard, 2003). Participants in the dry run were asked to rapidly develop language processing resources for Cebuano (ISO 639-3: ceb), a language widely spoken in the southern Philippines but with few established tools for automatic language processing. The resources were intended to support development of systems for cross-language information retrieval (CLIR), enabling discovery of relevant documents in a target language (Cebuano) based on search queries issued in a source language (here, English). Using resources ranging from textbooks and scholarly resources to online dictionaries and translations, participants developed tools for automatic shallow morphological analysis (stemming), statistical machine translation, and a full CLIR system in roughly 2.5 days. Similarly, "surprise" versions of subtasks or languages have been introduced into recent shared tasks in Information Extraction, such as the "Surprise Slot Filling" subtask of the Text Analysis Conference (TAC) 2010 Knowledge Base Population (KBP) task (Chen et al., 2010), Information Retrieval, where "surprise" languages are planned for a transliterated search task at FIRE2014[3], and code-switching detection, with a "surprise" genre at the First Workshop on Computational Approaches to Code Switching[4].

There have been a number of speech and language projects addressing low-resource tasks, with a recent speech-related conference hosting a special day devoted to low-resource approaches for speech processing and the "Spoken Language Technologies for Under-resourced Languages" workshop series. A significant recent stream of work under the IARPA-funded BABEL program (Harper, 2014) has brought a renewed focus on automatic processing of low-resource and relatively less-studied languages. In particular, the program highlights speech processing tasks including spoken term detection (essentially keyword spotting) and automatic transcription of conversational telephone speech in languages ranging from Lao (lao) to Zulu (zul) to Assamese (asm) to Pashto (pus). These systems are trained using only relatively small amounts of transcribed audio; about 10 hours of such transcribed audio are provided, although additional unlabeled data is available along with materials in other languages. In contrast, current high-quality speech recognition systems for English may be trained on 600-1000 hours of audio.

[3] http://research.microsoft.com/en-us/events/fire13_st_on_transliteratedsearch/fire14st.aspx

[4] http://emnlp2014.org/workshops/CodeSwitch/call.html

These efforts have spurred research on languages with fewer available text and speech resources in a resource-constrained setting. They have also significantly broadened the range and diversity of languages receiving high levels of attention from the speech and language processing community. Results on tasks to date indicate the need for further research, but also demonstrate that current techniques achieve differing levels of effectiveness on different languages and also that drawing on resources from many languages can prove useful in improving effectiveness on languages with few available resources (Mangu et al., 2013; Tusk et al., 2014; Gales et al., 2014). Application to such a diverse range of languages has also generated some new insights. For example, it has been found that pitch information can improve automatic speech recognition across many languages, not just those with lexical tone, as had been previously assumed (Metze et al., 2013). The BABEL program also has incorporated the surprise language paradigm. Systems in the initial phase developed tools for four languages (Cantonese (yue), Pashto (pus), Tagalog (tgl), and Turkish (tur)) and then were tested by building a new system for Vietnamese (vie) in four weeks.

4.3 What Is Useful? Targeting High-Value Tasks

Our goal with this proposal is to stimulate research in NLP and speech technology that is of interest both to the NLP/speech tech communities as well as to the language documentation community. Given data with a variety of annotations, there will be a range of tasks we can define for the STECs that involve extending those annotations to more data. However, not all such tasks are equally useful for field linguists. Accordingly, we plan to invite a small group of field linguists, working on different projects in different field conditions, to UW to demonstrate for us their work-flow in collecting and annotating data and to discuss with us where the most promising points for computational assistance are. In particular, we are looking for tasks which correspond to real bottlenecks in field linguistic research and for which having answers which are frequently but not always right are still useful. These can be tasks where it's faster to correct computer output than to annotate completely by hand or tasks where having an approximate answer can support further work, even without correction.

We will plan to hold this workshop about 6 months into the grant period, when we are mostly through our survey of existing field data and so have a sense of which tasks are possible given the available data. We will seek to bring five field linguists together from around the US and Canada and will look for linguists working with typologically diverse languages and under different field conditions. For example, the fieldwork process is very different working with communities with established orthographies and literate speakers who can be hired to do transcription and even translation than it is in communities where there is no established orthography nor speakers comfortable with the writing system.

The primary purpose of our workshop will be to inform the choice of STECs that we organize under this grant, but we will collect the information about work-flow and likely useful points of computational methods and make it available through the project web page.

5 Sample Tasks

As noted above, the specific tasks we focus our STECs on won't be decided until we've completed both our survey of available data and more importantly the workshop with field linguists to identify the most high-value tasks we can approach. However, in order to make the discussion in this proposal more concrete, in this section we outline a couple of potential tasks and describe why we think they would be of interest to both the speech/NLP and documentary linguistics communities.

Ideally, our tasks would target development of a full speech-to-text transcription system with high phonetic fidelity, needing little or no additional tuning for a new language. However, such a system is still well beyond the state-of-the-art for speech recognition; standard systems trained on only ten hours of transcribed audio often have word error rates of 70% or more on conversational telephone speech (Hsiao et al., 2013). Though the use of more sophisticated techniques and additional untranscribed audio can reduce the error rates by 15% or more absolute (Hsiao et al., 2013), this level of accuracy would be unacceptable for immediate use in documentation of endangered languages, where furthermore even ten hours of transcribed audio may not be available for training. Automatic, language-independent phonetic transcription is an even more distant goal: Pure phone recognition remains a fundamental challenge for automatic methods, which rely heavily on word and word sequence (language model) constraints for effectiveness[5] with best published results of about 17% phone error rate (Graves et al., 2013), on very clean read speech where word level recognition achieves less than 3% word error rate. Thus, we focus on speech and language processing tasks that are more constrained but which we believe will stimulate important basic computational research while yielding effective tools for language documentation in the relatively short term.

5.1 Candidate Speech-Based Task: Forced Alignment of Transcribed Speech to Audio for Non-ideal Conditions

A key step in analyzing recorded speech corpora is aligning transcriptions of the speech to particular time spans in the audio waveform. Such alignments at the word or phone level enable a wide range of more detailed phonetic and acoustic-prosodic analysis of the speech, including analyses of duration, prosody, and intonational patterns. Manual alignment is very time-consuming and requires significant phonetic expertise, especially for phone-level alignment. Automatic forced alignment, in contrast, uses a manual transcription of the speech and speech recognition tools to automatically perform the time-alignment. Although such alignments may not be entirely consistent with human alignments and may contain some outright errors, the combination of automatic alignment and manual correction can still represent a significant savings in time and effort over a fully manual process.

A number of forced alignment tools exist, typically built upon established toolkits for speech recognition, such as the Penn Phonetics Lab Forced Alignment (P2FA) (Yuan &

[5]The CMUSphinx speech recognizer's webpage (`http://cmusphinx.sourceforge.net/wiki/phonemerecognition`) comes with an explicit "caveat emptor" warning on phone recognition.

Liberman, 2008) or the ProsodyLab-Aligner (Gorman et al., 2011) which provide wrappers for the Hidden Markov Model Toolkit (HTK)[6]. Some tools, including the ProsodyLab-Aligner used in the LingSync language documentation toolkit (Dunham et al., 2014), also support acoustic model training, to adapt to a new dialect or language, when given a new pronunciation dictionary. These systems work quite well for word alignment under ideal conditions, given single-speaker per channel recordings, on short spans of audio (or with pre-segmented spans (Strunk et al., 2014)), perfect and complete transcriptions, and clear, canonical speech input. However, their performance degrades quickly if any of these assumptions are violated: if speech is overlapped or mumbled, if transcription is incomplete or imperfect (any missing words), or if the recording is long (more than a few seconds) or noisy (any background noise at all). All pronunciation variation also needs to be included manually. In addition, alignment quality at the phonetic level has not generally been explicitly optimized in existing alignment tools.

Several of these issues have already drawn some attention in the Speech technology community, including alignment of long recordings (Katsamanis et al., 2011; Prahallad & Black, 2011), speech in noise (Virtanen et al., 2013), and enhanced phonetic alignment (Yuan et al., 2013; Stolcke et al., 2014). However, they are still far from resolved. Our proposed alignment STEC would encompass alternate tracks addressing one or more of these key alignment challenges.

5.2 Candidate Text-Based Task: Morphological Segmentation of Unglossed, Untranslated Text

Morphological analysis is a key aspect of language description. Identifying the basic building blocks of the language provides scaffolding for other levels of linguistic analysis (see e.g. Chelliah & de Reuse, 2010, Ch 12). Creating an inventory of the language's morphemes also facilitates dictionary development. Performing such segmentation and analysis can be difficult and time-consuming. The task of automatic morphological segmentation involves developing algorithms to identify morpheme boundaries in running text. This could conceivably be of use to field linguists working with collections of narratives that are transcribed but not glossed, even if the automatic segmentation is not 100% accurate. If it were accurate enough, the proposed bondaries could lead the linguist doing the glossing to investigate morphological processes not yet documented.

From the computational perspective, automatic morphological segmentation has been the subject of significant research (Brent et al., 1995; Goldsmith, 2001; Creutz & Lagus, 2007; Dasgupta & Ng, 2007). Furthermore, a shared task series has focussed on the development of tools for unsupervised morphological segmentation (Kurimo et al., 2010), with impressive results on a range of languages of different morphological typologies, including English , Turkish, and Finnish. However, a number of assumptions about resource availability underlie the methods proposed to date. For example, Poon et al. (2009, p.209) state "Unsupervised morphological segmentation is attractive, because in every language there

[6] `http://htk.eng.cam.ac.uk/`

are virtually unlimited supplies of text, but very few labeled resources." In the case of endangered languages, these assumptions about the ready availability of large quantities of unlabeled text data are not supported. Thus, the application of the proposed computational techniques on true low-resource endangered languages will provide an opportunity to assess their true efficacy and level of resource dependence, while developing a useful resource for linguists working with endangered languages.

We would like to reemphasize that these are just potential tasks; the actual choice will depend on both what is possible given the data available and what is most helpful, according to the linguists we work with at the workshop.

6 Languages

It is not possible to say ahead of time which languages we will be working with, but we lay out here the factors we will take into consideration:

Data Availability Different NLP/speech processing tasks have different requirements in terms of amounts of training data. We will be selecting tasks that maximize all of (a) interest to documentary and descriptive linguistics, (b) interest to the NLP and speech technology research communities, and (c) suitability given the size of the data sets available, both in terms of what is presently in the archives for our use in running the STECs but also in terms of what can feasibly be collected in a typical field project.

Beyond size of data set, a primary concern is access rights. In order for an STEC to have maximum impact, it must be possible for later researchers to return to the same data set with minimal hassle to reproduce previous results and extend the state of the art. We will restrict our attention to data that can be freely used for research purposes, or, if sufficient data in that category cannot be found, work towards 'freeing' data in cases where the restrictions come from the researchers involved and not the communities.

Endangerment Status While the main purpose of this grant is to stimulate fundamental research on algorithms that will be useful for all language documentation projects, the particular languages involved in the STECs will benefit from increased research activity. Accordingly, among languages for which there is sufficient data collected and the data can be freely used for access purposes, we will prioritize languages that are most endangered, using the ElCat's Language Endangerment Index (Heaton & Okura, 2013). We will also prioritize languages which are still the subject of active field work, such that insights gained through additional work with the data can be most helpful for on-going documentation.

Typological and Areal Diversity One goal shared by the NLP/speech technology community and field linguistics is an interest in algorithms that work for any natural language.[7]

[7]In working with audio data, we're restricting our attention to spoken rather than signed languages. Any STECs focusing on glossing or other aspects of working with textual data should ideally include signed language data if such are available that meet the other criteria above.

However, algorithms developed and tested on one language won't necessarily work comparably well on others (Bender, 2011). Accordingly, we will endeavor to include in our STECs multiple languages which are furthermore typologically and areally well-separated, to the extent possible given the considerations of data availability and endangerment status.

7 Organizing Shared Task Evaluation Challenges

We plan to organize shared tasks to facilitate and accelerate the work-flows of field linguists working with endangered languages. We will target both spoken language and text-based tasks, with at least one task in each class. The tasks may involve multiple tracks within a broad task, for example with each track addressing different endangered languages.

We plan to co-locate these STECs with larger conference venues, possibly under the auspices of a more general workshop. A number of speech and language technology conferences and workshops have established a framework for conducting shared tasks and providing forums for reporting and disseminating the results. For example, in speech processing, the biennial IEEE Automatic Speech Recognition and Understaning Workshop (ASRU) publishes a call soliciting proposals for shared tasks, while Interspeech, the flagship conference of the International Speech Communication Association, often structures shared task activities through Special Sessions of their main conference. In the Natural Language Processing community, the SemEval and Computational Approaches in Natural Language Learning (CoNLL) workshops solicit and vet shared task proposals for multiple evaluation tasks associated with their workshop events. These workshops are typically held jointly with the main or chapter conferences of the Association for Computational Linguistics (ACL, NAACL, EACL, IJCNLP, EMNLP).

For each of these cases, a proposal for the STEC is submitted to the event organizers. Accepted STECs are expected to follow event guidelines for the timeline for conducting the shared task itself and are provided with a presentation and publication venue at the associated workshop or special session. STEC proposals are typically solicited 9-12 months in advance of the workshops, with the tasks to be conducted over a period of 4-6 months including training and evaluation phases. Short papers are then submitted describing systems and subjected to an accelerated review process.

Based on our resource survey and our needs assessment workshop, we will identify the tasks to be performed and the data to be employed in the STECs. We will submit STEC proposals to the most appropriate venues. After acceptance, we will conduct the STECs according to the timelines and guidelines of the collocating event, with the following main activities.

- **Participant recruitment:** Participants will be recruited through public calls for participation, both through the events' mailing lists and through broader speech and language processing mailing lists. These lists include LinguistList, the Corpora list, and professional society distribution lists such as those for ACL and ISCA. We will also directly solicit participation from researchers known to work in the task area.

- **Task data preparation:** We will create sample, training, and test portions of the selected data sets. We will also reformat data as needed to support the computational needs of the shared task participants and to be consistent with data format standards in the related computational fields. This pre-processing will also allow for consistent presentation of data resources across different languages or drawn from different archives. Data will include ground truth annotations as necessary for training and evaluation.

- **Data dissemination:** The prepared data and any software required for pre-processing will be provided to the archive sites, to be distributed to the STEC participants. This model will ensure both that the original archives maintain control of distribution of the data and that any derived formats are made available and stored by the archives for future access. We have already secured the cooperation of three large archives (ELAR, PARADISEC, and ANLA) for these activities. (Please see included letters of support.)

- **Evaluation:** We will select or develop an evaluation metric appropriate to each of the tasks. Where the STEC task has an established evaluation metric already determined by the speech and language processing community (e.g. word error rate for speech recognition tasks), we will adhere to that metric. Where no such standardized metric exists, we will develop or adapt our own. Software for evaluation and result format validation will be made available to participants through the shared task website.

- **Task execution:** Sample, training, and test data will be made available to participants according to the task timeline. Final results submissions from participants will be received through the STEC website. We will compute the official evaluation scores for the submitted runs and return the results to the participants.

- **Workshop:** The results and approaches of task participants will be presented at a workshop. The organizers will present an overview paper describing the findings, successes, and remaining challenges from the task activities. Particularly novel or effective methods will be invited to give oral presentations of their systems. All participants will be encouraged to present their work in short papers and poster presentations. The workshop will provide an opportunity to assess the task and shared lessons learned. Following the workshop, a website will be maintained devoted to the shared tasks, providing links to the data archives and published results, as well as access to necessary software. Participants will be encouraged to make their systems available, linked to the site either directly or through their own website.

8 Project Deliverables and Timeline

The first nine months of the project will focus on design and development of the STECs. The process will begin with a survey of applicable resources from existing archives. It will also encompass the workshop with field linguists to assess the needs of those working in this area. On that basis, we will design the STECs to be performed, for which suitable and sufficient data is available and which are responsive to field linguistics needs.

The final twelve months of the project will focus on the performance on the STECs we have designed to develop tools to support documentation and research on endangered languages.[8] Following the standard timeline for shared task campaigns, we will submit formal proposals for the shared tasks to the most relevant speech and language processing fora, conduct the shared tasks, and support presentation of results and approaches at a workshop linked to a major speech and/or language processing venue.

The project will yield the following deliverables:

- Survey of archived endangered language data, providing a detailed collation of quantity of resources; types, formats, and quantities of annotation; and level of resource availability
- STEC design documents
- Sample, training, and evaluation test data: Standardized training and evaluation test splits of the data will be created; formats and annotations will be systematized.
- Task software: Any scripts developed for corpus data handling and task evaluation will be made available.
- STEC workshop report and proceedings

The aligment of deliverables to the project timeline is summarized in Table 2.

Activity	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Archive Survey	•	•					
Field Linguist Workshop		•					
Task Design			•				
Task Proposal				•			
Task Activities					•	•	
Task Workshop							•

Table 2: Project timeline and milestones

9 Results from Prior NSF Support

9.1 Learning Tone

NSF IIS-0414919, Amount: $313,500, Role: PI, Period: 2004-2009

Intellectual Merit: This project built on recent phonetic research to develop an articulatorily motivated model of tone, focusing on minimally supervised models. A common

[8]Total project duration is 24 months to accommodate different possible dates for shared task workshops; a hiatus may be needed to ensure staffing during the STEC and workshop.

model spanned tone and pitch accent language families, ranging from high-resource languages (Mandarin (cmn) and English (eng)) to low-resource Bantu languages, such as isiZulu (zulu) and isiXhosa (xho). Supervised classification experiments demonstrated improvements due to contextual modeling for all languages. Results on unsupervised clustering and semi-supervised learning approached those for supervised learning while using dramatically less training data. (Levow, 2005, 2006; Surendran & Levow, 2008; Wang & Levow, 2008; Levow, 2009a, 2009b, 2008).

Broader Impact: Three REU students were trained in prosodic analysis and automatic recognition under this project. Two Ph.D. students were trained under this grant. Project software is available at: `http://faculty.washington.edu/levow/projects/tai`.

9.1.1 Other NSF-supported Activities

BCS #0729515; Title:"Dyadic Rapport within and across Cultures : Multimodal Assessment of Human-Human and Human-Computer Interaction"; Role: PI; Period: 01/15/2008-12/31/2012; Amount: $415,643 Intellectual Merit: Levow's project created an audio-video corpus of dyadic interactions to investigate multi-modal cues to conversational rapport in three language-cultural groups: American English (eng), Iraqi Arabic (acm), and Mexican Spanish (spa). Analysis of this multi-modal data identified differences in interactional feedback across these groups and contrasts across feedback types; it further enabled automatic prediction of this feedback based on acoustic-prosodic cues (Levow & Wang, 2012; Wang & Levow, 2011; Levow & Duncan, 2012; Levow et al., 2010).

Broader impacts: Two REU students were trained in computational speech processing and a team of graduate and undergraduate students were trained and participated in analysis of multi-modal data.

IIS #1351034; Title: "EAGER: ATAROS: Automatic Tagging and Recognition of Stance"; Role: PI; Period: 09/15/2013–08/31/2015; Amount: $257,836 Intellectual Merit: In ATAROS, Levow and colleagues created the ATAROS corpus, designed to elicit high rates of stance-taking at varying strengths in dyadic conversation. The project developed novel dynamic acoustic measures to assess stance-related speaking style, and demonstrated highly effective recognition of stance strength and polarity (Freeman et al., 2014; Luan et al., 2014).

Broader impacts: The project has trained one REU student, three undergraduate students, and four graduate students from Electrical Engineering, Linguistics, and Computational Linguistics in analysis and automatic recognition of stance in speech. Corpus and documents are available from `http://depts.washington.edu/phonlab/projects.htm`.

9.2 AGGREGATION

BCS #1160274; Title "AGGREGATION: Automatic Generation of Grammars for Endangered Languages from Glosses and Typological Information [ctn, ing,

inh]"; PI: Bender; Period 9/15/12–3/15/15; Amount: $224,039, with REU supplement BCS #1358097, $4,032 Intellectual Merit: AGGREGATION grant supported by DEL) is investigating computational methods for the development of precision grammars on the basis of interlinear glossed text (IGT) collected in field projects. Project outcomes to date include the development of Xigt, an XML representation for IGT and the development of algorithms that can take in IGT and create grammars. The resulting grammars currently have low coverage, but for the sentences they can handle they produce richer representations than what is in the IGT input, including spanning syntactic and semantic analyses. For any STECs we develop that use IGT annotations, we anticipate using the Xigt format for data encoding. Furthemore, Bender's experience in working with field data in the context of the AGGREGATION project will inform our work on defining STECs which are both approachable and useful.

Broader Impacts: The REU student is now a PhD student in Computational Linguistics. The work is on-going, but anticipated broader impacts include support for richer language documentation.

10 Conclusion

In this project, we propose to promote fundamental research on speech and language processing which could lead to effective tools for endangered language documentation by building shared task evaluation challenges (STECs) around data from endangered language archives.

Intellectual Merit STECs have long been recognized as a driving force in research progress in natural language processing and speech technology. There is growing interest in low-resource languages among these communities and a strong demand for independently prepared data sets and standardized evaluations. By developing STECs on the basis of endangered language data and in consultation with field linguists about the points in their work-flow most likely to benefit from automated analysis, we can align the interests of these two communities to their mutual benefit.

Broader Impacts The STECs that we will design will stimulate fundamental research on algorithms that could lead to high-value tools for language documentation. The algorithmic advances achieved will also be applicable to non-endangered low-resource languages, helping to spread the benefits of language technology to a broader part of the world's population. We will make the data sets and evaluation scripts freely available. This means that even past the organized STEC, researchers will be able to revisit the task and search for (and demonstrate) improvements over the state of the art.

References

Anastasakos, T., McDonough, J., Schwartz, R., & Makhoul, J. (1996). A compact model for speaker-adaptive training. In *Proceedings of the International Conference on Spoken Language Processing* (p. 1137-1140). doi: 10.1109/ICSLP.1996.607807

Belz, A., & Kilgarriff, A. (2006). Shared-task evaluations in HLT: Lessons for NLG. In *Proceedings of the Fourth International Natural Language Generation Conference* (pp. 133–135). Sydney, Australia: Association for Computational Linguistics.

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

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 440–447).

Blitzer, J., McDonald, R., & Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 120–128).

Brent, M. R., Murthy, S. K., & Lundberg, A. (1995). Discovering morphemic suffixes: A case study in minimum description length induction. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society.*

Brody, S., & Lapata, M. (2008). Good neighbors make good senses: Exploiting distributional similarity for unsupervised WSD. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)* (p. 65-72).

Chelliah, S. L., & de Reuse, W. J. (2010). *Handbook of descriptive linguistic fieldwork.* Springer.

Chen, Z., Tamang, S., Lee, A., Li, X., Lin, W.-P., Snover, M., Artiles, J., Passantino, M., & Ji, H. (2010). CUNY-BLENDER TAC-KBP2010 entity linking and slot filling system description. In *Proceedings of the Third Text Analysis Conference (TAC 2010).*

Choi, F., Wiemer-Hastings, P., & Moore, J. (2001). Latent semantic analysis for text segmentation. In *Proceedings of EMNLP* (p. 109-117).

Christodoulopoulos, C., Goldwater, S., & Steedman, M. (2010). Two decades of unsupervised POS induction: How far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 575–584).

Creutz, M., & Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, *4*(1).

Dahl, D. A., Bates, M., Brown, M., Fisher, W., Hunicke-Smith, K., Pallett, D., Rudnicky, A., & Shriberg, E. (1994). Expanding the scope of the ATIS task: the ATIS-3 corpus. In *Proceedings of the ARPA Human Language Technology Workshop* (pp. 43–48). Morgan Kaufmann.

Dasgupta, S., & Ng, V. (2007). High-performance, language-independent morphological segmentation. In *Proceedings of Human Language Technology (NAACL).*

Daumé III, H. (2009). Frustratingly easy domain adaptation. *CoRR*, *abs/0907.1815*. Retrieved from `http://arxiv.org/abs/0907.1815`

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(4), 788-798.

Druck, G., Mann, G., & McCallum, A. (2009). Semi-supervised learning of dependency parsers using generalized expectation criteria. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 360–368).

Dunham, J., Cook, G., & Horner, J. (2014). LingSync & the online linguistic database: New models for the collection and management of data for language communities, linguists and language learners. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages* (pp. 24–33).

Freeman, V., Chan, J., Levow, G.-A., Wright, R., Ostendorf, M., & Zayats, V. (2014). Manipulating stance and involvement using collaborative tasks: An exploratory comparison. In *Proceedeings of Interspeech 2014.*

Gales, M., Knill, K., Ragni, A., & Rath, S. (2014). Speeech recognition and keyword spotting for low resource languages: Babel project research at CUED. In *Proceedings of the 4th Workshop on Spoken Language Technologies for Under-resourced Languages* (pp. 16–23).

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, *27*(2), 153-198.

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, *39*(3), 192-193.

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Proceedings of ICASSP 2013.*

Grishman, R., & Sundheim, B. (1996). Message understanding conference - 6: A brief history. In *Proceedings of Coling 1996* (pp. 466–471).

Harper, M. (2014). *IARPA BABEL Program.* Retrieved from `http://www.iarpa.gov/Programs/ia/Babel/babel.html` (Accessed September 2014)

Hearst, M. (1994). Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics.*

Heaton, R., & Okura, E. (2013). The catalogue of endangered languages in context. In *3RD International Conference on Language Documentation and Conservation (ICLDC).* Retrieved from `http://scholarspace.manoa.hawaii.edu/handle/10125/26144`

Hsiao, R., Ng, T., Grezl, F., Karakos, D., Tsakalidis, S., Nguyen, L., & Schwartz, R. (2013). Discriminative semi-supervised training for keyword search in low resource languages. In *Proceedings of ASRU 2013* (pp. 440–445).

Jeon, J. H., & Liu, Y. (2009). Semi-supervised learning for automatic prosodic event detection using co-training algorithm. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 540–548).

Katsamanis, A., Black, M., Georgiou, P., Goldstein, L., & Narayanan, S. (2011). Sailalign: Robust long speech-text alignment. In *Proceedings of Workshop on New Tools and*

Methods for Very-Large Scale Phonetics Research.

Koo, T., Carreras, X., & Collins, M. (2008). Simple semi-supervised dependency parsing. In *Proceedings of ACL-08: HLT* (pp. 595–603).

Kurimo, M., Virpioja, S., Turunen, V., & Lagus, K. (2010). Morpho challenge 2005-2010: Evaluations and results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology* (pp. 87–95).

Levow, G.-A. (2005). Context in multi-lingual tone and pitch accent prediction. In *Proceedings of Interspeech 2005.*

Levow, G.-A. (2006). Unsupervised and semi-supervised tone and pitch accent recognition. In *Proceedings of HLT-NAACL 2006* (pp. 224–231).

Levow, G.-A. (2008). Automatic prosodic labeling with conditional random fields and rich acoustic features. In *Proceedings of IJCNLP 2008* (pp. 217–224).

Levow, G.-A. (2009a). Assessing context and learning for isiZulu tone recognition. In *Proceedings of Interspeech 2009* (pp. 716–719).

Levow, G.-A. (2009b). Investigating pitch accent recognition in non-native speech. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* (pp. 269–272).

Levow, G.-A., Duncan, S., & King, E. (2010). Cross-cultural investigation of prosody in verbal feedback in interactional rapport. In *Proceedings of Interspeech 2010* (pp. 286–289).

Levow, G.-A., & Duncan, S. D. (2012). Contrasting cues to verbal and non-verbal backchannels in multi-lingual dyadic rapport. In *Proceedings of Interspeech 2012.*

Levow, G.-A., & Wang, S. (2012). Employing boosting to compare cues to verbal feedback in multi-lingual dialog. In *Proceedings of SLT-2012.*

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the ACL/COLING* (p. 768-774).

Luan, Y., Wright, R., Ostendorf, M., & Levow, G.-A. (2014). Relating automatic vowel space estimates to talker intelligibility. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (Interspeech 2014).*

Mangu, L., Soltau, H., Kuo, H.-K., Kingsbury, B., & Saon, G. (2013). Exploiting diversity for spoken term detection. In *Proceedings of ICASSP.*

Margolis, A., Livescu, K., & Ostendorf, M. (2010). Domain adaptation with unlabeled data for dialog act tagging. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing* (pp. 45–52). Uppsala, Sweden: Association for Computational Linguistics. Retrieved from `http://www.aclweb.org/anthology/W10-2607`

Matveeva, I., & Levow, G.-A. (2007). Topic segmentation with hybrid document indexing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 351–359).

McCarthy, D., Koeling, R., Weeds, J., & Carroll, J. (2004). Finding predominant senses in untagged text. In *Proceedings of the 42th meeting of the Association for Computational Linguistics* (p. 280-287).

Metze, F., Sheikh, Z., Waibel, A., Gehring, J., Kilgour, K., Nguyen, Q. B., & Nguyen, V. H.

(2013). Models of tone for tonal and non-tonal languages. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (p. 261-266).

Oard, D. W. (2003). The surprise language exercises. *ACM Transactions on Asian Language Processing*, *2*(2), 79–84.

Oard, D. W., Doermann, D., Dorr, B., He, D., Resnik, P., Weinberg, A., Byrne, W., Khudanpur, S., Yarowsky, D., Leuski, A., Koehn, P., & Knight, K. (2003). Desparately seeking Cebuano. In *Proceedings of NAACL-HLT 2003 (Companion Volume)* (pp. 76–78).

Parrish, N., & Gupta, M. (2012). Dimensionality reduction by local discriminative gaussians. In *Proceedings of the International Conference on Machine Learning.*

Poon, H., Cherry, C., & Toutanova, K. (2009). Unsupervised morphological segmentation with log-linear models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL* (p. 209-217).

Prahallad, K., & Black, A. W. (2011). Segmentation of monologues in audio books for building synthetic voices. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(5), 1444-1449.

Price, P., Fisher, W., Bernstein, J., & Pallett, D. (1988). The DARPA 1000-word resource management database for continuous speech recognition. In *1988 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-88)* (p. 651-654 vol.1).

Stolcke, A., Ryant, N., Mitra, V., Yuan, J., Wang, W., & Liberman, M. (2014). Highly accurate phonetic segmentation using boundary correction models and system fusion. In *Proceedings of ICASSP 2014.*

Strunk, J., Schiel, F., & Seifart, F. (2014). Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. In *Proceedings of the Language Resources and Evaluation Conference (LREC).*

Surendran, D., & Levow, G.-A. (2008). Can voice quality improve Mandarin tone recognition? In *Proceedings of ICASSP 2008* (pp. 4177–4180).

Tusk, Z., Golik, P., Nolden, D., Schlute, R., & Ney, H. (2014). Data augmentation, feature combination, and multilingual neural networks to improve ASR and KWS performance for low-resource languages. In *Proceedings of Interspeech.*

Virtanen, T., Singh, R., & Raj, B. (Eds.). (2013). *Techniques for Noise Robustness in Automatic Speech Recognition.* Wiley.

Voorhees, E. M., & Harman, D. K. (Eds.). (2005). *TREC: Experiment and evaluation in information retrieval.* Cambridge, MA: The MIT Press.

Wang, S., & Levow, G.-A. (2008). Mandarin Chinese tone nucleus detection with landmarks. In *Proceedings of Interspeech 2008* (pp. 1101–1104).

Wang, S., & Levow, G.-A. (2011). Contrasting multi-lingual prosodic cues to predict verbal feedback for rapport. In *Proceedings of ACL-2011* (pp. 614–619).

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. In *Proceedings of Acoustics '08.*

Yuan, J., Ryant, N., Liberman, M., Stolcke, A., Mitra, V., & Wang, W. (2013). Automatic phonetic segmentation using boundary models. In *Proceedings of Interspeech.*