Proceedings of the Workshop on Computational Methods for Endangered Languages

Volume 1 *Proceedings of the 3rd Workshop on Computational Methods for Endangered Languages Vol. 1 Papers*

Article 3

2-26-2019

Future Directions in Technological Support for Language Documentation

Daan van Esch
Google, dvanesch@google.com

Ben Foley
University of Queensland, ARC Centre of Excellence for the Dynamics of Language, b.foley@uq.edu.au

Nay San
Stanford University, Australian National University, ARC Centre of Excellence for the Dynamics of Language, nay.san@stanford.edu

Future Directions in Technological Support for Language Documentation

Daan van Esch
Google
dvanesch@google.com

Ben Foley
The University of Queensland,
ARC Centre of Excellence for the Dynamics of Language
b.foley@uq.edu.au

Nay San
Stanford University,
Australian National University,
ARC Centre of Excellence for the Dynamics of Language
nay.san@stanford.edu

Abstract

To reduce the annotation burden placed on linguistic fieldworkers, freeing up time for deeper linguistic analysis and descriptive work, the language documentation community has been working with machine learning researchers to investigate what assistive role technology can play, with promising early results. This paper describes a number of potential follow-up technical projects that we believe would be worthwhile and straightforward to do. We provide examples of the annotation tasks for computer scientists; descriptions of the technological challenges involved and the estimated level of complexity; and pointers to relevant literature. We hope providing a clear overview of what the needs are and what annotation challenges exist will help facilitate the dialogue and collaboration between computer scientists and fieldwork linguists.

1 The Transcription and Annotation Challenge

Language documentation projects typically yield more data than can be reasonably annotated and analyzed by the linguistic fieldworkers that collect the data. For example, transcribing one hour of audio recordings can take 40 hours or more (Durantin et al., 2017), and potentially even longer where the language is severely under-studied or under-described. To reduce the annotation burden placed on linguistic fieldworkers, freeing up time for deeper linguistic analysis and descriptive work, the language documentation community has been working with machine learning (ML) researchers to investigate what role technology can play. As a result, today enterprising fieldwork linguists can use a number of toolkits to help accelerate the annotation process by automatically proposing candidate transcriptions. Such toolkits include CoEDL Elpis (Foley et al., 2018), Persephone (Adams et al., 2018; Michaud et al., 2018), and SPPAS (Bigi, 2015).

2 A Technological Development Roadmap

These toolkits are already showing promising results across a number of languages. Arguably, however, their current incarnation only scratches the surface of what is technically possible, even today. In this paper, we aim to describe a number of potential extensions to these systems, which we believe would be reasonably straightforward to implement, yet which should significantly improve the quality and usefulness of the output. Specifically, we look at the annotation steps in the language documentation process beyond just phonemic or orthographic transcription, and describe what assistive role technology can play for many other steps in the language documentation process, such as automatically proposing glossing and translation candidates.

To be clear, this paper does not actually describe complete systems that have been implemented. Rather, our goal is to describe technical projects that we believe would be worthwhile and straightforward to do. We provide examples of the annotation tasks for computer scientists; descriptions of the technological challenges involved and the estimated level of complexity; and pointers to relevant literature on both the computational and the linguistic side. Our hope is that this overview of what the needs are and what annotation challenges exist will help facilitate the dialogue and collaboration between computer scientists and fieldwork linguists.

Our higher-level vision is for a toolkit that lets fieldwork linguists process data at scale, with limited technical knowledge needed on the side of the user. This toolkit should make it as easy as possible to apply the most relevant well-known techniques in natural language processing and machine learning to any language. This will likely involve providing a simple, pre-configured graph-

Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages: Vol. 1 Papers, pages 14–22,
Honolulu, Hawai'i, USA, February 26–27, 2019.

ical user interface; for the purposes of this paper, however, we focus on the underlying technology.

For our purposes, we assume there is an audio corpus where orthographic transcriptions are available for at least some subset of the recordings. If no orthographic transcriptions exist at all, then there are a number of techniques around audio-only corpus analysis that can be applied, such as those explored in the Zero-Resource Speech Challenges (Dunbar et al., 2017), but those approaches are outside of the scope of this paper.

We also assume that the corpus at hand was created in a relatively ubiquitous tool for language documentation — e.g. ELAN (Max Planck Institute for Psycholinguistics; Brugman and Russel, 2004), Praat (Boersma and Weenink, 2001), or Transcriber (DGA, 2014) — for which there would be a ready-to-use conversion tool to convert the corpus into a common format that can be understood by the toolkit. Finally, we assume that the linguist has access to a tool to create keyboards for use on a desktop/laptop system, such as Keyman (SIL), so an appropriate orthography can be used, and that the linguist has access to an entry input method for the International Phonetic Alphabet (IPA).

Broadly, we see two areas where automation toolkits for language documentation can make significant headway still: automatic analysis of existing data in the corpus, and automatic processing of unannotated data. To facilitate the discussion below, Table 1 briefly walks through a reasonably comprehensive set of annotations for an audio recording. Specifically, for each word in our Nafsan example utterance, Table 1 includes:

- a speaker label to indicate which of the speakers in the recording spoke this word
- the orthographic form, or simply spelling, of this word
- an interlinear gloss, which we will describe in further detail below for readers who are unfamiliar with this practice
- a part-of-speech tag, e.g. using the conventions from (Petrov et al., 2012)
- a phonemic transcription, e.g. in the International Phonetic Alphabet

In addition, our example annotation shows a translation of the entire utterance into a single language.[1]

3 Analysis of Existing Data

3.1 Text Analytics

In terms of automatic analysis of existing data, there appears to be a significant amount of low-hanging fruit around text-based analytics. Once a language worker has put the effort into preparing a corpus for use in a toolkit, by normalizing and consolidating files, the data becomes suitable for automatic analysis to obtain corpus-wide metrics. For example, a toolkit could easily emit the following metrics on the orthographic transcriptions in the corpus:

- the total number of words observed[2]
- the number of unique words observed
- a frequency-ranked wordlist, with word counts
- the average number of words in each utterance, and a histogram of words per utterance

These metrics can also be broken down by speaker, if multiple speakers are present in the corpus and speaker annotations are available in the metadata. In that case, a toolkit could also automatically compute statistics such as pointwise mutual information score for the usage of each word by each speaker, or by speaker group, for example if multiple varieties are represented in the corpus. This may point to interesting inter-speaker or inter-variety lexical frequency differences that merit further investigation. Per-speaker lexical analyses could also be used to inform fieldwork elicitation adjustments, e.g. when some words were not elicited from a particular speaker in the data set. For a literature review describing a large number of automatic analysis options in this space, see (Wieling and Nerbonne, 2015).

Frequency-ranked wordlists in particular can also be helpful for quality assurance purposes:

[1] Fieldwork linguists can, naturally, choose to create this translation in any language. In our example, for expository purposes, we used English.

[2] We use "words" here to mean specifically orthographic units separated by whitespace in the target-language orthography; if such separation is not part of the orthography, then more complex approaches are needed, as is the case when users wish to run these analyses based on lexical roots (after morphological analysis).

Speaker	Spelling	Gloss	Part-of-Speech	Phonemes
SPEAKER1	waak	pig	NOUN	wak
SPEAKER1	nen	that	DET	nan
SPEAKER1	i=	3SGREALISSUBJECT	PRON	i
SPEAKER1	p̃as	chase	VERB	p̃as
SPEAKER1	=ir	3PLOBJECT	PRON	ir

Table 1: Annotations for an example recording in Nafsan (ISO 639 erk), DOI 10.4225/72/575C6E9CC9B6A. Here, the English translation would be "That pig chased them."

words at the bottom of the list may occur only as one-off *hapax legomena*, but they may also be misspelled versions of more frequent words. Reviewing a frequency-ranked list may help find orthographic inconsistencies in the data set. In fact, an automatic toolkit could also emit additional information around annotation quality, such as:

- a frequency-ranked character list, with pointers to utterances that contain rarely used characters (on the assumption that they may be data entry mistakes)
- words that seem to be misspellings of another word, based on edit distance and/or context

For datasets in languages where comprehensive, clean wordlists are available, a transcription tool with built-in spell-checking or word-completion features would benefit users. If no clean wordlist exists yet, a frequency-based wordlist is a good start and can be human-curated. Going beyond just orthographic words, finite-state technologies may also be needed, especially for morphologically complex languages where stems may have more variations than could ever be covered by a simple wordlist, as is frequently described in the literature; two relevant recent surveys are (Mager et al., 2018; Littell et al., 2018).

If interlinear glosses are available in the corpus, the toolkit could also easily emit a list of all words and their glosses, highlighting words with multiple glosses, as these may (but do not necessarily) represent data entry inconsistencies; the same is true for part-of-speech tags.

Where phoneme annotations are present, a forced-alignment with the orthographic transcription may be carried out to surface any outliers, as these may represent either orthographic or phonemic transcription errors (Jansche, 2014). Of course, a phoneme frequency analysis could also be produced and may yield interesting insights.

It should also be straightforward, though the linguistic value of this would remain to be determined, to apply an unsupervised automatic morphemic analyzer like Morfessor (Virpioja et al., 2013).

3.2 Visualization

Language documentation toolkits will benefit from integration with information visualization tools. Visual data analysis can give a linguist a general overview of the nature of a corpus, and narrow in on data to support new, detailed insights. For example, a broad view of a language collection can be given by visualizing metadata from the corpus, representing the activity of the documentation process as a heatmap of the recording dates/times/participants. Date and time visualizations could show patterns within recording locations and events, indicating how long it took to create the corpus, and how old the data is.

Visualizations showing which time spans within a corpus have been annotated on particular annotation tiers would allow a linguist to quickly obtain a sense of the state of transcription. A representation of transcription state showing what has been transcribed, the types of existing transcriptions, with summaries of metadata, will assist the language worker to understand the shape of their parallel layered data, and develop well-constructed corpora. This information could help prevent unexpected data loss when moving from complex multi-tier transcription to other modes of representation (Thieberger, 2016).

Detailed visualizations can also represent linguistic features of the data, giving the linguist on-demand access to explore a variety of phenomena — e.g. using ANNIS (Krause and Zeldes, 2016) — such as compositional structure of linguistic units (hierarchical data), associations between words (relational data) and alternative word choices (Culy and Lyding, 2010).

Finally, a toolkit could output a high-level overview of statistics and visualizations for a given corpus as a landing page: a description of the collection for online publishing, or to fa-

cilitate discovery of the data within an archive (Thieberger, 2016).

4 Extending Automatic Processing of Unannotated Data

4.1 Audio Analysis

Generally, part of the data set at hand will be missing at least some of the annotation layers in our example above. At the most basic level, it may not be known what parts of an audio collection contain human speech, what parts contain other types of sound (e.g. only nature or car noise), or are even just quiet. A speech vs. non-speech classifier, known as a voice activity detector or VAD (Ramirez et al., 2007), should be built into the toolkit. One open-source VAD is the WebRTC Voice Activity Detector, with an easy-to-use Python wrapper in (Wiseman, 2016). Beyond voice activity detection, speaker identification or diarization may also be necessary. In this regard, the work recently done as part of the First DIHARD Speech Diarization Challenge (Church et al., 2018) is particularly relevant for fieldwork recordings, which are challenging corpora for these technologies.

Practically, entirely unannotated data sets are rare: usually at least the language of the data set is known. Where it isn't, or where multiple languages exist within a given recording, it will also be helpful to be able to identify the language spoken in each part of the audio, based on the recording alone, though this is a hard technical challenge (Gonzalez-Dominguez et al., 2014). As a first step, multiple existing recognizers could be executed, and their output could be analyzed to determine which one produces the most likely transcript, using an approach like that by (Demšar, 2006).

4.2 Automatic Phonemic and Orthographic Transcription

Once the location(s) and the language(s) of speech within the target audio recordings are known, the relevant audio fragments can be processed by speech recognition systems such as Elpis (Foley et al., 2018) or Persephone (Adams et al., 2018) in order to produce automatic machine hypotheses for phonemic and orthographic transcriptions. Today's Elpis and Persephone pipelines are maturing, and are relatively complete packages, but they could be made easier to use for people without backgrounds in speech science.

4.2.1 Pronunciations

For example, Elpis currently requires the linguist to provide a machine-readable grapheme-to-phoneme mapping. However, these may already be available from current libraries (Deri and Knight, 2016), derivable from existing databases like Ausphon-Lexicon (Round, 2017), or present in regional language collections such as the Australian Pronunciation Lexicon (Estival et al., 2014) and Sydney Speaks (CoEDL, 2018). These resources could be integrated directly into the toolkit. Elpis currently also offers no support for handling not-a-word tokens like numbers (van Esch and Sproat, 2017), but this could be supported using a relatively accessible grammar framework where the number grammar can be induced using a small number of examples (Gorman and Sproat, 2016).

4.2.2 Text Corpora

Toolkits should also facilitate easy ingestion of additional text corpora, which is particularly useful for training large-vocabulary speech recognition systems. For some endangered languages, text corpora such as Bible translations or Wikipedia datasets may be available to augment low-volume data sets coming in from language documentation work. Of course, the content of these sources may not be ideal for use in training conversational systems, but when faced with low-data situations, even out-of-domain data tends to help. As more text data becomes available, it would also be good to sweep language model parameters (such as the n-gram order) automatically on a development set to achieve the best possible result.

4.2.3 Audio Quality Enhancement

In terms of audio quality, it's worth pointing out that audio recorded during fieldwork is often made in noisy conditions, with intrusive animal and bird noises, environmental sounds, or air-conditioners whirring. Noisy corpora may become easier to process if automatic denoising and speech enhancement techniques are integrated into the pipeline. Loizou (2013) provides a recent overview of this particular area.

4.2.4 Multilingual Models

To further improve the quality of the output produced by the automatic speech recognition

(ASR) systems within toolkits such as Elpis and Persephone, multilingual acoustic models can be trained on data sets from multiple languages (Besacier et al., 2014; Toshniwal et al., 2018). This could allow so-called zero-shot model usage, meaning the application of these acoustic models to new languages, without any training data at all in the target language. Of course, the accuracy of such an approach would depend on how similar the target language is to the languages that were included in the training data for the multilingual model. Another approach is simply to reuse a monolingual acoustic model from a similar language. Either way, these models can be tailored towards the target language as training data becomes available through annotation.

4.2.5 Automatic Creation of Part-of-Speech Tags and Interlinear Glosses

Beyond automatically generating hypotheses for phonemic and orthographic transcriptions, it would technically also be relatively straightforward to produce candidate part-of-speech tags or interlinear glosses automatically. Both types of tags tend to use a limited set of labels, as described by e.g. the Universal Part-of-Speech Tagset documentation (Petrov et al., 2012) and the Leipzig Glossing Rules (Comrie et al., 2008). Where words occur without a part-of-speech tag or an interlinear gloss in the corpus (for example, because a linguist provided an orthographic transcription but no tag, or because the orthographic transcription is an automatically generated candidate), a tag could be proposed automatically.

Most straightforwardly, this could be done by re-using an existing tag for that word, if the word was already tagged in another context. Where multiple pre-existing interlinear glosses occur for the same surface form, it would be possible to contextually resolve these homographs (Mazovetskiy et al., 2018). For part-of-speech tagging, many mature libraries exist, such as the Stanford Part-of-Speech Tagger (Toutanova et al., 2003). However, in both cases, given the limited amount of training data available within small data sets, accuracy is likely to be low, and it may be preferable to simply highlight the existing set of tags for a given word for a linguist to resolve. Surfacing these cases to a linguist may even bring up an annotation quality problem that can be fixed (if, in fact, there should have been only one tag for this form).

Deriving interlinear glosses or part-of-speech tags for new surface forms that have not previously been observed could be significantly harder, depending on the target language. Achieving reasonable accuracy levels will require the use of morphological analyzers in many languages, e.g. as in (Arkhangeskiy et al., 2012). In polysynthetic languages in particular, development of such analyzers would be a challenging task (Mager et al., 2018; Littell et al., 2018), though see (Haspelmath, 2018) for a cautionary note on the term *polysynthetic*.

4.2.6 Automatic Machine Translation (MT)

Many language documentation workflows involve the translation of sentences from the target-language corpus into some other language, such as English, to make the content more broadly accessible. In effect, this creates a parallel corpus which can be used for training automatic machine translation (MT) systems. These MT systems could then, in turn, be used to propose translation candidates for parts of the corpus that are still lacking these translations: this may be because the linguist did not yet have time to translate their transcriptions, or because only machine-generated hypotheses are available for the transcriptions (in which case the cascading-error effect typically causes accuracy of translations to be lower). Such candidate translations would still require human post-editing, but for limited-domain applications, machine translations may be sufficiently accurate to accelerate the annotation process. Of course, where other parallel corpora exist (e.g. bilingual story books, or religious material like the Bible), these can also be used as training data. In addition, any existing bilingual word lexicons (like dictionaries) can also help in building machine translation systems, as in (Klementiev et al., 2012).

Many high-quality open-source machine translation toolkits already exist, e.g. SockEye (Hieber et al., 2017), TensorFlow (Luong et al., 2017), and Moses (Koehn et al., 2007). Phrase-based machine translation still tends to yield better results than neural machine translation (NMT) for small data sets (Östling and Tiedemann, 2017), but multilingual NMT models seem promising for cases where similar languages exist that do have large amounts of training data (Johnson et al., 2017; Gu et al., 2018). Recent advances even allow the creation of speech-to-translated-text models (Bansal et al., 2018) and enable the use of existing translations

to enhance ASR quality (Anastasopoulos and Chiang, 2018).

5 Putting structured data and models to work more broadly

A toolkit designed to facilitate transcription and annotation efforts can benefit language workers and language communities beyond just the annotation output. Structured data sets used in these toolkits are ripe for conversion into formats needed for various other tasks, such as dictionary creation, e.g. Electronic Text Encoding and Interchange (TEI, 2018), or for dictionary publishing pipelines, as in Yinarlingi (San and Luk, 2018). Exported data could be formatted for import into a range of dictionary/publishing applications such as lexicon apps for mobile devices or ePub format (Gavrilis et al., 2013) for publishing on a wide range of devices such as smartphones, tablets, computers, or e-readers. Automatically generated pronunciations, word definitions, part-of-speech information and example sentences could easily be included. For difficult work such as automatic lemmatization, the output could be presented in a simple interface for human verification/correction before publication (Liang et al., 2014).

For languages to thrive in the digital domain, some technologies are considered essential: a standardized encoding; a standardized orthography; and some basic digital language resources, such as a corpus and a spell checker (Soria, 2018). If toolkits make it easy to create structured data sets and models for these languages, then these resources can also be applied outside the fields of language documentation, lexicography, and other linguistic research fields.

For example, Australia has a vibrant community-owned Indigenous media broadcasting sector. For these users, there is potential to re-use the ASR models to generate closed captions (subtitles) for the broadcast material. ASR transcription and translation technologies could be used to enrich the output of these Indigenous community media broadcasters, which would yield a significant improvement in the accessibility of broadcast content. Another option would be to facilitate the creation of smartphone keyboards with predictive text in the target languages, using toolkits like Keyman (SIL).

Language archives like PARADISEC could also benefit by applying these models, once created, on existing content for the same language that may have been contributed by other linguists, or that may have come in through other avenues. It may even be possible for these archives to offer web-based services to language communities, e.g. to allow them to use machine translation models for a given language in a web-based interface. Linguists could archive their models alongside their corpus; for some collections, the published models may inherit the same access permissions as the training data used in their creation, while some models may be able to be published under less restrictive conditions.

5.1 Data Set Availability

In general, to support the development of language technologies like the ones we described earlier, it is critical that software engineers are aware of, and have access to data sets reflecting the diversity of endangered languages. Already, data sets are available for a range of languages, and in formats suitable for ASR, text-to-speech synthesis (TTS) and other tasks on the Open Speech and Language Resources website (Povey, 2018). The addition of ML-ready data sets for more endangered languages would enable software engineers, whose primary concern may not be the language documentation process, to be involved in developing and refining speech tools and algorithms. However, access protocols and licenses in place for existing data sets can prohibit experimental development. Recording corpora specifically for the purposes of enabling ML experiments would avoid potentially lengthy negotiations of access rights.

6 Reaching more languages

Extending the reach of a toolkit beyond the (typically) few languages which are used when building and testing is critical. Applying technology to more diverse use cases encourages a tool to become more robust. With more use cases, a community of support can grow. A community of users, contributors and developers around a toolkit is important to encourage potential participants (Foley, 2015), and to reduce the burden of support on individual developers.

Being proactive in community discussions to publicize the abilities of tools, providing in-person access to training workshops, publishing online support material and tutorials, and ensuring tools have high-quality documentation for differ-

ent levels of users, are all important (albeit labor-intensive) methods of promoting and encouraging language technology to reach more languages.

7 Conclusion

Speech and language technology toolkits, designed specially for users without backgrounds in speech science, have the potential for significant impact for linguists and language communities globally. Existing toolkits can be enhanced with richer feature sets, and connection with workflow processes, helping language documentation workers. These toolkits enable language documentation workers to focus on deeper linguistic research questions, by presenting the results of automated systems in ways that let language experts easily verify and correct the hypotheses, yielding annotation speed-ups. At the same time, making these technologies and the structured data they help produce more widely available would benefit language communities in many ways.

Most of the technologies described here are readily available and easily implemented, while others are still highly experimental in their application and potential benefits for Indigenous languages. With language technology as a whole making rapid progress, and with an increasing amount of dialogue between fieldwork linguists and computer scientists, it is an exciting time to be working on computational methods for language documentation, with many advances that look to be within reach for the near future.

Acknowledgments

We would like to thank all the participants in the Transcription Acceleration Project (TAP), run by the Centre of Excellence for the Dynamics of Language, for many fruitful discussions on how technology could be applied in the language documentation context to assist fieldwork linguists. We would like to add a special word of thanks to Nick Thieberger, who provided the Nafsan example, and who offered valuable input on a draft version of this paper. We would also like to thank Zara Maxwell-Smith and Nicholas Lambourne for all of their insightful suggestions and contributions.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of LREC 2018*.

Antonis Anastasopoulos and David Chiang. 2018. Leveraging translations for speech transcription in low-resource settings. *CoRR*, abs/1803.08991.

Timofey Arkhangeskiy, Oleg Belyaev, and Arseniy Vydrin. 2012. The creation of large-scale annotated corpora of minority languages using uniparser and the eanc platform. In *Proceedings of COLING 2012: Posters*, pages 83–92. The COLING 2012 Organizing Committee.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. Low-resource speech-to-text translation. *CoRR*, abs/1803.09164.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Commun.*, 56:85–100.

Brigitte Bigi. 2015. SPPAS: Multi-lingual approaches to the automatic annotation of speech. *The Phonetician*, 111-112(ISSN:0741-6164):54–69.

Paul Boersma and David Weenink. 2001. Praat, a system for doing phonetics by computer. 5:341–345.

Hennie Brugman and Albert Russel. 2004. Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of LREC 2004*.

Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, Mark Liberman, and Neville Ryant. 2018. The First DIHARD Speech Diarization Challenge. https://coml.lscp.ens.fr/dihard/index.html.

CoEDL. 2018. Sydney Speaks Project. http://www.dynamicsoflanguage.edu.au/sydney-speaks.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. The Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses.

Chris Culy and Verena Lyding. 2010. Visualizations for exploratory corpus and text analysis. In *Proceedings of the 2nd International Conference on Corpus Linguistics CILC-10*, pages 257–268.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.

Aliya Deri and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 399–408.

DGA. 2014. TranscriberAG: a tool for segmenting, labeling and transcribing speech. http://transag.sourceforge.net.

Ewan Dunbar, Xuan-Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. 2017. The zero resource speech challenge 2017. *CoRR*, abs/1712.04313.

Gautier Durantin, Ben Foley, Nicholas Evans, and Janet Wiles. 2017. Transcription survey.

Daan van Esch and Richard Sproat. 2017. An expanded taxonomy of semiotic classes for text normalization. In *Proceedings of Interspeech 2017*.

Dominique Estival, Steve Cassidy, Felicity Cox, and Denis Burnham. 2014. Austalk: an audio-visual corpus of australian english. In *Proceedings of LREC 2014*.

Ben Foley. 2015. Angkety map digital resource report. Technical report.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, František Kratochvíl, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2018)*.

Dimitris Gavrilis, Stavros Angelis, and Ioannis Tsoulos. 2013. Building interactive books using EPUB and HTML5. In *Ambient Media and Systems*, pages 31–40, Cham. Springer International Publishing.

Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, and Hasim Sak. 2014. Automatic language identification using long short-term memory recurrent neural networks. In *Proceedings of Interspeech 2014*.

Kyle Gorman and Richard Sproat. 2016. Minimally supervised number normalization. *Transactions of the Association for Computational Linguistics*, 4:507–519.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354. Association for Computational Linguistics.

Martin Haspelmath. 2018. The last word on polysynthesis: A review article. 22:307–326.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *CoRR*, abs/1712.05690.

Martin Jansche. 2014. Computer-aided quality assurance of an icelandic pronunciation dictionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Alexandre Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2012. Toward statistical machine translation without parallel corpora. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 130–140, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *The Annual Conference of the Association for Computational Linguistics*.

Thomas Krause and Amir Zeldes. 2016. Annis3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118–139.

Y Liang, K Iwano, and Koichi Shinoda. 2014. Simple gesture-based error correction interface for smartphone speech recognition. pages 1194–1198.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. Indigenous language technologies in canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632. Association for Computational Linguistics.

Philipos C. Loizou. 2013. *Speech Enhancement: Theory and Practice*, 2nd edition. CRC Press, Inc., Boca Raton, FL, USA.

Minh-Thang Luong, Eugene Brevdo, and Rui Zhao. 2017. Neural machine translation (seq2seq) tutorial. *https://github.com/tensorflow/nmt*.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza. 2018. Challenges of language technologies for the indigenous languages of the Americas.

Gleb Mazovetskiy, Kyle Gorman, and Vitaly Nikolaev. 2018. Improving homograph disambiguation with supervised machine learning. In *LREC*.

Alexis Michaud, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit.

Robert Östling and Jörg Tiedemann. 2017. Neural machine translation for low-resource languages. *CoRR*, abs/1708.05729.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Dan Povey. 2018. Open speech and language resources. http://www.openslr.org.

The Language Archive Max Planck Institute for Psycholinguistics. ELAN. https://tla.mpi.nl/tools/tla-tools/elan/.

J. Ramirez, J. M., and J. C. 2007. Voice activity detection. fundamentals and speech recognition system robustness. In *Robust Speech Recognition and Understanding*. I-Tech Education and Publishing.

Erich R. Round. 2017. The AusPhon-Lexicon project: Two million normalized segments across 300 Australian languages.

Nay San and Ellison Luk. 2018. Yinarlingi: An R package for testing Warlpiri lexicon data. https://github.com/CoEDL/yinarlingi.

SIL. Keyman: Type to the world in your language. https://keyman.com/.

Claudia Soria. 2018. Digital Language Survival Kit. http://www.dldp.eu/en/content/digital-language-survival-kit.

TEI. 2018. Guidelines for electronic text encoding and interchange. http://www.tei-c.org/P5/.

Nick Thieberger. 2016. Language documentation tools and methods summit report. https://docs.google.com/document/d/1p2EZufVIm2fOQy4aIZ100jnZjfcuG-Bb9YKiXMggDRQ.

Shubham Toshniwal, Tara N. Sainath, Ron J. Weiss, Bo Li, Pedro J. Moreno, Eugene Weinstein, and Kanishka Rao. 2018. Multilingual speech recognition with a single end-to-end model. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4904–4908.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline. Technical report.

Martijn Wieling and John Nerbonne. 2015. Advances in dialectometry. *Annual Review of Linguistics*, 1(1):243–264.

John Wiseman. 2016. Python interface to the WebRTC Voice Activity Detector. https://github.com/wiseman/py-webrtcvad.