Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 306–315
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Phonemic Transcription of Low-Resource Languages: To What Extent can Preprocessing be Automated?

Guillaume Wisniewski[1], **Alexis Michaud**[2], **Séverine Guillaume**[2]
(1) Laboratoire de Linguistique Formelle (LLF), CNRS – Université Paris-Diderot, Case 7031,
5 rue Thomas Mann, 75013 Paris, France
(2) Langues et Civilisations à Tradition Orale (LACITO), CNRS – Université Sorbonne Nouvelle – INALCO,
7 rue Guy Môquet, 94800 Villejuif, France
guillaume.wisniewski@u-paris.fr, alexis.michaud@cnrs.fr, severine.guillaume@cnrs.fr

Abstract

Automatic Speech Recognition for low-resource languages has been an active field of research for more than a decade. It holds promise for facilitating the urgent task of documenting the world's dwindling linguistic diversity. Various methodological hurdles are encountered in the course of this exciting development, however. A well-identified difficulty is that data preprocessing is not at all trivial: data collected in classical fieldwork are usually tailored to the needs of the linguist who collects them, and there is baffling diversity in formats and annotation schema, even among fieldworkers who use the same software package (such as ELAN). The tests reported here (on Yongning Na and other languages from the Pangloss Collection, an open archive of endangered languages) explore some possibilities for automating the process of data preprocessing: assessing to what extent it is possible to bypass the involvement of language experts for menial tasks of data preparation for Natural Language Processing (NLP) purposes. What is at stake is the accessibility of language archive data for a range of NLP tasks and beyond.

Keywords: Endangered Languages, Speech Recognition/Understanding, Speech Resource/Database

1. Introduction

1.1. Making Language Archive Data Tractable to Automatic Speech Processing: Why Preprocessing is a Key Issue

Towards Computational Language Documentation Automatic speech recognition (ASR) tools have potential for facilitating the urgent task of documenting the world's dwindling linguistic diversity (Besacier et al., 2014; Thieberger, 2017; Littell et al., 2018; van Esch et al., 2019). Encouraging results for automatic phoneme recognition for low-resource languages were published two years ago (Adams et al., 2018), and prospects of widespread deployment of the technology look extremely hopeful.

Why Preprocessing is a Major Hurdle Various methodological hurdles are encountered in the course of this exciting development, however. A well-identified difficulty is that data preprocessing is not at all trivial. In classical linguistic fieldwork (Bouquiaux and Thomas, 1971; Newman and Ratliff, 2001; Dixon, 2007), "good corpus production is ongoing, distributed, and opportunistic" (Woodbury, 2003, 47), and thus unlike scenarios in which data acquisition is tailored to meet the requirements of speech processing tasks. Because fieldwork data are not collected specifically for the purpose of ASR, data sets from language archives are highly diverse in a number of respects. Not only is there a wide range of tools for creating linguistic annotations, each with its own format (see the conversion tools TEIconvert `http://ct3.ortolang.fr/teiconvert/` and Multitool `https://github.com/DoReCo/multitool`): there is also diversity in the formats allowed by one and the same software package. Thus, ELAN, a commonly used software package (Brugman and Russel, 2004), allows users to define their own document structures: ELAN supports creation of multiple tiers and tier hierarchies, so that there is, in practice, no such thing as a unique "ELAN format". It would be desirable for a common format to be adopted in the mid run, such as the standard proposed as part of the Text Encoding Initiative (Schmidt, 2011; Liégeois et al., 2016), but convergence is not in sight yet. In the current situation, fieldwork data make up "eclectic data collections" rather than "systematically annotated corpora" (Gerstenberger et al., 2017, 26). Preprocessing typically involves retrieving pieces of information that are not encoded according to widely shared computational standards.

Preprocessing tasks are not just time-consuming: they require familiarity with the target language, and with the specific corpus. This is asking a lot from Natural Language Processing people who wish to try their hand at the data. An example (preprocessing transcriptions of Yongning Na, a Sino-Tibetan language, for training an acoustic model using the `Persephone` toolkit) is documented in some detail in an article that aims to explain to an audience of linguists (i) the way the automatic transcription toolkit `Persephone` operates and (ii) how the process of collaborating with natural language processing specialists was initiated and developed (Michaud et al., 2018). Trying to summarize the 37-page article in one sentence, it seems fair to say that without a sustained dialogue with the linguist who created the transcriptions, the pitfalls of preprocessing would probably have been enough friction to turn computing people off.

Adapting Data Collection Methods for Easier Application of Natural Language Processing Tools? One possible way to go would be to get linguists and Natural Language Processing experts to modify their usual workflows, and to work hand in hand designing and applying tools together. Computer scientists would take the time to find out about the implicit structure of the data sets, and also absorb as much information as possible about the linguistic structure of the target languages. Field linguists would anticipate the requirements of a range of Natural Language Processing tools from the early stages of data collection in the field. It has even been suggested that field linguists should modify their practice so as to assist the task of machine learning: for instance, "making multiple parallel or semi-parallel recordings, so as to have a robust envelope of phonetic variation across speakers that assists in generalizing sound-transcription matching from one speaker to another" (Seifart et al., 2018).

But an issue with this approch is that it adds to the workload of fieldworkers and computer scientists. Speech data acquisition has numerous challenges of its own (Niebuhr and Michaud, 2015), which linguists need to prioritize in their work. Thus, although *respeaking* is known to be a possible way to improve the performance of Automatic Speech Recognition (Sperber et al., 2013), the limited amounts of time that the language consultants and the linguist can spend together are best devoted to recording additional original materials and discussing linguistic issues, rather than to the mechanical task of going through a set of audio files and repeating each sentence. Moreover, tailoring speech data acquisition to cater to the needs of machine learning algorithms is problematic given how rapidly the technology evolves. There is a potential conflict between the traditional perspective of creating a reasonably thorough and balanced record for posterity, on the one hand, and on the other hand, the requirement to put together data sets that lend themselves easily to Natural Language Processing.

From the point of view of Natural Language Processing engineers and computer scientists, the requirement to become familiar with the linguistic structure of the data sets likewise appears too steep, given the number of different languages (and of different data sets) that Natural Language Processing researchers handle in their work. The workflow in the first experiments on the `Persephone` toolkit (Adams et al., 2018) benefited from hands-on participation from the linguists who produced the transcriptions used as input data. Clearly, it is unrealistic to assume that as much 'insider' information will be available for all languages. Seen in this light, it becomes clear that what is at stake in preprocessing is no less than the availability of language archive data for language processing purposes.

The issue of facilitating preprocessing for Natural Language Processing is part of a broader topic which could be referred to as *interdisciplinary user design*: removing hurdles in the way of interdisciplinary collaborations. The expected benefit for language archives is that they can become accessible to an increased number of users, from a wider ranger of backgrounds. To date, data from language archives remain little-used, not only in Natural Language Processing but also in experimental phonetic research, for example (Whalen and McDonough, 2019).

1.2. Goals

The tests reported in the present paper aim at investigating possibilities for automating the process of data preprocessing: assessing to what extent it is possible to bypass the need for thorny and time-consuming expert tasks. We use the same data set from the Yongning Na language as was used in a previous study (Adams et al., 2018) to investigate which properties in the input transcription are conducive to best results in the recognition task. Second, we extend the tests to new languages.

1.3. Relevance to Natural Language Processing Research

In addition to the goal of achieving practical usefulness for field linguists, phonemic transcription for low-resource languages raises several interesting methodological challenges for Natural Language Processing (NLP).

- The amount of training data (transcribed audio) is limited: for data collected in linguistic fieldwork, ten hours counts as a large corpus. Corpus size can be less than one hour.

- Languages differ greatly from one another along various dimensions: phonemic inventories, phonotactic combinations, word structure, not to mention morphology, syntax and pragmatics. As a result, experiments over fieldwork data lead to encounters with a host of linguistic phenomena that differ from those commonly observed in the most widely spoken languages. Designing NLP methods to deal with this diversity of languages is a good way to explore the limits of state-of-the-art models and better understand how (and when) they are working.

- The sheer number of languages to be addressed suggests that attempts at a language-independent acoustic model may be a fruitful avenue to explore. (The world's 30 most widely spoken languages only represent about 1% of the world's linguistic diversity – on the order of 6,000 languages.)

2. Method

2.1. Phonemic Transcription Model

Prediction Model Our work aims at developing a phonemic transcription model which, given an audio signal represented by a sequence of `fbank` vectors,[1] predicts the corresponding sequence of phonemes and tones.

We use the implementation of a long short-term memory (LSTM) recurrent neural network provided by the `Persephone` toolkit (Adams et al., 2018). In all our experiments, we have considered a network made of 3 hidden layers with 250 hidden units. Our experiments show that, as pointed out in Adams et al. (2018), these parameters consistently achieve 'good' performances.

We use the Connectionist Temporal Classification (CTC) loss function (Graves et al., 2006) as a training criterion. This loss function allows us to learn the mapping between an audio signal and a sequence of phonemes without explicitly knowing the alignment between each phoneme and the corresponding audio frames.

[1] We consider the 41 usual `fbank` features as well as their first and second derivative.

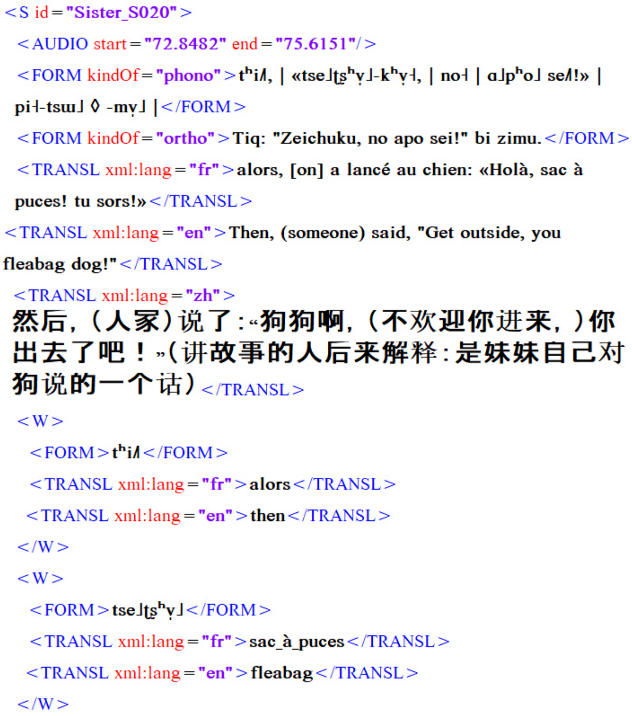
We trained our model using the Adam optimizer (Kingma and Ba, 2015), stopping after 100 epochs or when the loss on the validation set stopped improving for 10 consecutive epochs.

2.2. Extracting Labels

Extracting labels from the training data is a crucial step of the acoustic model creation workflow. The tests carried out here aim at automating the process, with as little reliance as possible on hand-crafted rules based on interactions with the linguist who produced the transcriptions. This is important for wider use of the tool in real-world application.

In the phonemic transcription task we consider, labels are sequences of phonemes. However, when field linguists annotate their recordings, phoneme boundaries are not encoded as such. Linguistic fieldwork data typically consist of transcriptions that are time-aligned with the audio (and, increasingly, video) signal at the level of larger units, such as sentences or intonation units, but rarely at the word level, and even more rarely at the level of the phoneme, which is not even encoded as such. For instance, in the XML format of the Pangloss Collection (Michailovsky et al., 2014),[2] texts are divided into sentences (S), themselves divided into words (W), divided into morphemes (M). Phonemes are not encoded as a level of their own. Instead, there is an implicit convention that transcriptions (the FORM at each of the levels) consist of strings of phonemes. Thus, no information about phonemes is readily available.[3] An example is shown in Figure 1.

For the benefit of readers who are not thoroughly familiar with XML, let us spell out explanations on the data structure. (Most readers can safely skip the present paragraph.) The identifier of the sentence (indicating simply that this is the twentieth sentence) is followed by an AUDIO element containing the time codes for the entire sentence (in this instance, from the 72nd to the 75th second on the audio recording), then by sentence-level transcriptions, coded as the sentence's FORM. There can be different types of transcription: for instance, in this example, there is a phonemic transcription, tagged as "phono" (the value assigned to the attribute kindOf associated with the transcription), and an orthographic transcription, tagged as "ortho". The choice of using orthographic or phonological (phonemic) transcription is up to the contributor. Differences in performance of automatic transcription for phonological vs. orthographic input will be returned to below. Translations at any level (the sentence, as shown here, or the entire text, or a word or morpheme) appear as TRANSL elements, with a tag indicating the translation language using two-letter codes: "fr", "en" and "zh" for French, English and Chinese, respectively. Word-level information likewise contains FORM elements and TRANSL elements. Note that orthographic representation and Chinese translations are only offered at the level of the sentence, not for each word. Most pieces of information are optional: word-level glosses are not mandatory, any more than translation into any specific language. Linguists who contribute data to the language archive deposit their documents *as is*, with the levels of annotation that they chose to produce for the sake of their research purposes.



```
<S id="Sister_S020">
  <AUDIO start="72.8482" end="75.6151"/>
  <FORM kindOf="phono">tʰi˧˥, | «tse˩ʈʂʰv̩˩-kʰv̩˧, | no˧ | ɑ˩pʰo˩ se˧˥!» |
 pi˧-tsɯ˩ ◊ -mv̩˩ |</FORM>
  <FORM kindOf="ortho">Tiq: "Zeichuku, no apo sei!" bi zimu.</FORM>
  <TRANSL xml:lang="fr">alors, [on] a lancé au chien: «Holà, sac à
 puces! tu sors!»</TRANSL>
<TRANSL xml:lang="en">Then, (someone) said, "Get outside, you
fleabag dog!"</TRANSL>
 <TRANSL xml:lang="zh">
 然后，（人家）说了：«狗狗啊，（不欢迎你进来，）你
 出去了吧！»（讲故事的人后来解释：是妹妹自己对
 狗说的一个话）</TRANSL>
 <W>
   <FORM>tʰi˧˥</FORM>
   <TRANSL xml:lang="fr">alors</TRANSL>
  <TRANSL xml:lang="en">then</TRANSL>
 </W>
 <W>
   <FORM>tse˩ʈʂʰv̩˩</FORM>
   <TRANSL xml:lang="fr">sac_à_puces</TRANSL>
  <TRANSL xml:lang="en">fleabag</TRANSL>
 </W>
```

Figure 1: Sample of XML code: beginning of sentence 20 of the narrative "The sister's wedding" (https://doi.org/10.24397/pangloss-0004342#S20).

Building an automatic phonemic transcription system therefore requires to, first, segment the transcriptions into sequences of phonemes. Transcriptions often contains annotations or comments about the audio content: e.g. to indicate the presence of drum rolls in the transcription of a song, or to point out that the annotator is not sure about what they have heard. See, for instance, the first sentence (S-unit) of the epic "Rani Raut 2"[4], whose transcription contains the indication "Dedicatory chant", instead of a transcription of the chant itself. Another example is sentence 24 of the Yongning Na narrative "How the Lake was created",[5] which contains a critical apparatus encoded through conventions (square brackets for additions, and angle brackets for deletions) that are not explained within the XML file itself.

It is also important to 'clean' the transcription to remove information that cannot be directly predicted. This cleanup ensures a direct mapping between the transcription and the audio, therefore making both learning and prediction easier.

In the 2018 LREC paper (Adams et al., 2018), transcrip-

[2] All our experiments are based on corpora freely available from the Pangloss Collection, an open archive of (mostly) endangered languages. See § 2.3. for details.

[3] There are possibilities for adding word-level and phoneme-level time codes to linguistic fieldwork documents using forced alignment (Strunk et al., 2014). In the current state of language archives, phoneme-level alignment remains a rarity, however.

[4] https://doi.org/10.24397/pangloss-0004315#S1

[5] https://doi.org/10.24397/pangloss-0004349#S24

tions extracted from the Yongning Na documents in the Pangloss Collection were segmented into sequences of phonemes by means of a set of hand-crafted rules. These rules are based on two kinds of information:

- a list of all the phonemes that can appear in Na;
- the explicit knowledge of the convention used by the field linguist that has annotated the data, e.g. that text between square brackets corresponds to additions and comments and should be removed to obtain a direct mapping of the sound signal to the transcription.

The use of these two kinds of information made it possible to extract labels of very high quality in which the sequence of phonemes was a faithful transcription of the audio signal. The process is documented in detail in §3.2 of Michaud et al. (2018). This workflow assumes hands-on participation from the linguist who produced the transcriptions used as input data. By contrast, in the present work, a much simpler approach is chosen to extract the sequence of phonemes from transcriptions.

Tools for Automatic Normalization of Unicode Labels[6] Following the recommendations of the *Unicode Cookbook for Linguists* (Moran and Cysouw, 2018), we carry out a segmentation into *grapheme clusters* in which each letter (as identified by its Unicode category) is grouped with all the modifiers (again identified by their Unicode category). Characters from other Unicode categories (e.g. modifier letter small h, ʰ) are considered as 'standard' letters. Table 1 shows an example of a sequence of labels segmented with this method, as compared with the output of the hand-crafted segmentation method of Adams et al. (2018).

①	si˧dzi˩-ʈʂʰɯ˩, | ʈʰææ̃˧ | ʈʂʰɯ˧-bv̩˧ɻ̍˧ | dɑ˧-kv̩˥-mæ˩! |
②	s␣i␣˧␣dz␣i␣˩␣ʈsʰ␣ɯ␣˩␣ʈʰ␣æ␣æ̃␣˧␣ʈʂʰ␣ɯ␣˧␣b␣v̩␣˧␣ɻ̍ ␣˧␣d␣ɑ␣˧␣k␣v̩␣˥-m␣æ␣˩
③	s␣i␣˧␣d␣z␣i␣˩␣- ␣ʈ␣ʂ␣ʰ␣ɯ␣˩␣ʈ␣ʰ␣æ␣æ̃␣˧␣ʈ␣ʂ␣ʰ␣ɯ␣˧␣b␣v̩␣˧␣ɻ̍ ␣˧␣d␣ɑ␣˧␣-␣k␣v̩␣˥-m␣æ␣˩

Table 1: Example of a transcription of Na (①) and two segmentations into label sequences: in ②, phonemes are separated by whitespaces using the rules of Adams et al. (2018); in ③, whitespaces identify grapheme clusters. Note that, in both segmentations, punctuations marks are removed: in the current setup, no attempts were made at predicting them.

The method used here, ③, has the advantage of being language-independent and of not relying on any knowledge of the data. It also comes with several drawbacks. First, it increases the number of possible labels, which makes both training and prediction slower. More importantly, it places higher demands on the statistical model, which could make prediction less successful. If a phoneme is made of two symbols (e.g. the digraph /dʑ/ for a voiced alveolo-palatal affricate), then these will be considered as two independent symbols and the transcription system will have to learn from the statistical distribution of these symbols that `d` when followed by `ʑ` may correspond to fairly different acoustic states than when followed by a vowel (in which case `d` constitutes a consonant on its own). The difference could have been made explicit by forcing segmentation as /dʑ/ in the one case and /d/ /ʑ/ in the other.

2.3. Workflow for Applying `Persephone` to Data Sets from the Pangloss Collection

A Command Line Interface between `Persephone` and the Pangloss Collection To test the `Persephone` toolkit for various languages, we have developed a simple command line interface between `Persephone` and the Pangloss Collection, a digital library whose objective is to store and facilitate access to audio recordings in endangered languages of the world (Michailovsky et al., 2014). Our tool[7] provides two commands. The first command allows a user to download, from the Pangloss Collection, all the audio recordings matching a language and/or a specific speaker (or set of speakers) and to organize the data so that they can be readily used by the `Persephone` toolkit. The second command can be used to train and test a phonemic transcription system.

The goal of this tool is twofold. First, it aims at allowing NLP practitioners to easily access datasets of great interest (or to say the least, with diverse and unusual characteristics) without having to spend time understanding how the data are organized in the Pangloss Collection. Second, it will (hopefully) help field linguists to train their own transcription models without having to convert their recordings and annotations into yet another format (the format required by `Persephone`).

Choice of Languages Out of the 170 languages currently hosted by the Pangloss Collection, we singled out seven for tests on automatic transcription. We chose data sets that had sentence-level alignment with the audio, a prerequisite for using `Persephone`. We also favoured languages for which substantial amounts of transcribed data are available: earlier tests suggest that when the training set is less than 20 minutes long, the model does not even converge, or error rates are extremely high. This criterion brushes aside no less than 112 languages: twenty minutes or more of transcribed data are currently available for only 58 languages.

Table 2 provides the main characteristics of the data sets we used in our experiments: language names, three-letter ISO codes from the Ethnologue inventory of languages (Simons and Fennig, 2017), duration of the training set, nature of the labels, and number of labels. In all our experiments we consider only a single speaker setting.

For the sake of reproducibility (Borgman, 2015; Maurel, 2016; Lust et al., 2019), a preprocessed version of all the data used in our experiments (i.e. the audio file for each sentence and the corresponding sequence of labels) orga-

[6] The implementation of this method is freely available in our extension of the `Persephone` toolkit, at `https://github.com/gw17/sltu_corpora`; see § 2.3. for details.

[7] Our tool is freely available in a fork of the `Persephone` toolkit: `https://bitbucket.org/gwisniewski/pangloss-persephone/src/pangloss/`.

nized according to the format expected by Persephone is available at https://github.com/gw17/sltu_corpora.

3. Experimental Results

We conducted three series of experiments to assess:

- the impact of using the label segmentation method described in Section 2.2. rather than hand-crafted rules tailored to the language at hand;
- the impact of considering different languages.

In all our experiments, we evaluate the performance achieved by our phonemic transcription system by computing the average edit distance between the predicted and gold labels of the test set (i.e. the Label Error Rate). This metric is a crude estimation of the effort required by an annotator to correct the prediction of an automatic transcription system.

3.1. Impact of Label Segmentation

Table 3 reports the performance achieved by Persephone when different segmentation methods are applied (see Section 2.2. for details). To start with, it is reassuring to note that we were able to reproduce the results reported in the study that we use as a point of reference (Adams et al., 2018): using the same rules to clean the transcription and identify phonemes, the prediction performance we achieved is very close to the earlier results.

As for segmentation methods, it also appears that using a generic segmentation method rather than a method tailored specifically for the target language hardly impacts prediction performance at all. Our interpretation is that Persephone is able to match polygraphs with phonemes (multiple character sequences, such as such as $\mathrm{ts^h}$, used to denote a phoneme), even when the components of these polygraphs, taken individually, refer to other phonemes (in Na, /t/ and /s/ are phonemes, as are /ts/ and /$\mathrm{ts^h}$/). This result does not come as a huge surprise, since the machine learning architecture is known to perform well in extracting patterns such as those described by phonotactics. We nonetheless see this as a very important observation from a practical point of view, because it suggests that it is possible to develop transcription systems with no knowledge of the language (in particular, without a list of phonemes drawn up by an expert linguist).

3.2. Evaluation on a Wider Array of Languages

Table 4 reports the performance achieved by Persephone on the selection of languages from the Pangloss Collection shown in Table 2. It appears that, for most languages, Persephone, when used as a black-box tool, performs very poorly. As shown by the learning curve (Figure 2), for four of the seven languages the system does not even seem to be able to memorize the training data. Increasing the number of parameters (i.e. the number of hidden units and/or of hidden layers) does not improve performance (neither on the validation set nor on the training set).

Several reasons can explain these disappointing results.

Audio Qxuality First, there appears to be a minimum threshold in terms of quality of the audio data. Some recordings may be of insufficient audio quality for automatic transcription given the current state of the art. For instance, the Dotyal data set consists of epics that contain singing, drums and bells: the successive sentences are sung or chanted, rather than spoken. Listening to the data,[8] it does not come as a surprise that automatic transcription as currently offered by Persephone does not work. Automatic Speech Recognition for such materials, if possible at all, will have to rely on much more elaborate processing.

Duration of Audio Chunks The upper limit on the duration of audio chunks taken as input by Persephone is 10 seconds. This results in exclusion of any longer chunks from the training process. Thus, the document "Romanmangan, the fairy from the other world"[9] has a duration of 1,890 seconds, and is divided into 212 sentences. The distribution of sentence durations is shown in Figure 3. Seventy sentences, amounting to 1 032 seconds (more than half of the total duration of this substantial story), are above the 10-second limit, and thus not used in training. The total amount of data available for the language is down from 22 minutes to 16. This goes a long way towards explaining why training fails to converge: there is simply not enough data to train a statistical model.

This issue affects the real-life usefulness of Persephone, and needs to be addressed so as to make use of all the available data for training. A possibility would be to detect silence and non-silence (by Voice Activity Detection) and then trim the long waveform, removing silences, so as to arrive at a duration below 10 seconds. But removing silences comes at the cost of tampering with the audio signal, removing cues that may well be relevant for training. Pauses are part and parcel of intonational structure, and removing them can create acoustic 'monsters'. Instead, the way to go is to do forced alignment as an initial approach, then split the long sentences based on silence, and finally feed the chunks thus obtained into training. This work is considered as part of future improvements planned for Persephone. Within the 10-second limit on audio chunks, it is likely that shorter time-aligned chunks in the training set make for better scores, but this has not been tested empirically yet.[10]

Number of Labels There are large differences in the number of labels and the quantity of training data be-

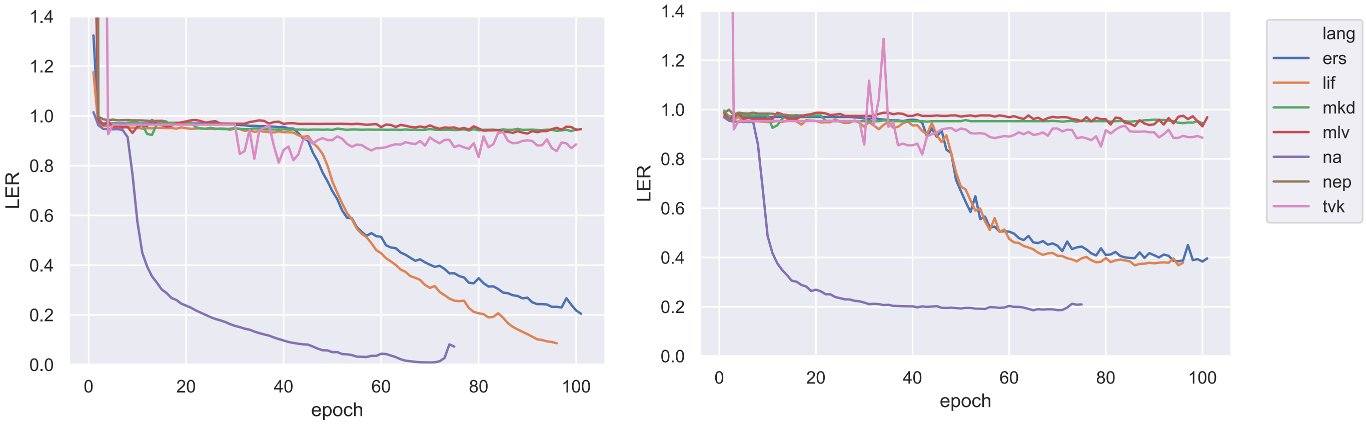
[8] For example: https://doi.org/10.24397/pangloss-0004091

[9] https://doi.org/10.24397/pangloss-0002300

[10] Remember that the level at which time alignment is generally provided in the Pangloss Collection's XML documents is the S level: the *sentence*, in a sense which contributors can interpret freely. A hypothesis to be tested empirically is that the average duration of the S-level units correlates with the field of specialization of the contributing linguist. Linguists with a strong interest in phonetics and phonology may tend to cut up speech into smaller units, whereas those with a stronger interest in syntax will tend to choose larger chunks, which constitute syntactically complete blocks. The Mwotlap corpus would be a case in point: the texts were collected by a specialist of syntax (François, 2003), and their relatively large chunks make good syntactic sense.

language	duration		IPA	# labels
	total	after filtering		
Dotyal (nep)	1h44mn	44mn	✗	366
Duoxu (ers)	32mn	32mn	✓	35
Mwotlap (mlv)	22mn	16mn	✗	39
Na (nru)	8h35mn	7h49mn	✓	80
Nashta (mkd)	25mn	23mn	✓	39
Limbu (lif)	1h50mn	1h34mn	✓	37
Vatlongos (tvk)	14mn	14mn	✗	20

Table 2: Languages from the Pangloss collection that were used in our experiments. The IPA column indicates whether the transcriptions are phonological (✓) or orthographic (✗). We report the size of each corpus ('total' column) as well as the size after utterances lasting more than 10s have been removing (see §3.2.).



(a) train

(b) validation

Figure 2: Learning curve (train & validation sets) for the different languages considered in our experiments

segmentation method	LER
Adams et al. (2018) (phonemes)	0.130
Adams et al. (2018) (phonemes + tones)	0.172
grapheme cluster (phonemes + tones)	0.186

Table 3: Prediction performance for different segmentation methods. LER = Label Error Rate.

Language	LER on train set	LER on test set
nru	0.016	0.186
lif	0.167	0.368
ers	0.218	0.383
tvk	0.822	0.818
mkd	0.926	0.926
mlv	0.944	0.932
nep	0.98	0.965

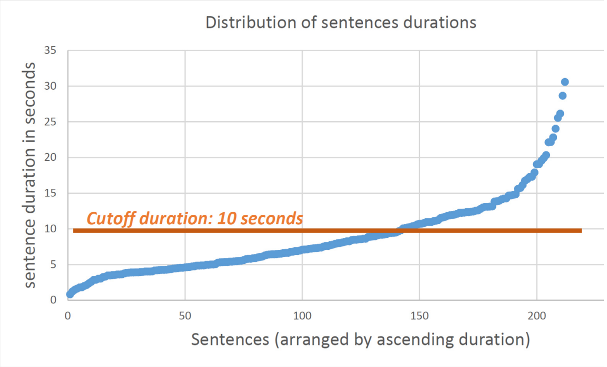
Table 4: Results (ordered from best to worse performance) achieved by the Persephone toolkit on different languages of the Pangloss collection. LER = Label Error Rate.

tween the different languages and there might not always be enough training data to properly estimate model parameters.

Degree of Phonetic/Phonological Transparency Last but not least, the use of a phonemic representation rather than an orthographic representation seems to result in better performances:[11] the transcriptions of all the languages in our test set for which Persephone was actually able to learn something use a phonemic representation in the International Phonetic Alphabet. At the time of creation of an alphabetic writing system, there is usually a good match between graphemes (orthographic units) and phonemes (phonological units). But as languages change, which they do constantly, they gradually diverge from the state reflected in the orthography. As a result, orthographic representations depart from phonemic structure to an extent that varies greatly across orthographic systems, depending partly on their time depth, and partly on features inherited from the orthographic systems that served as a reference when devising them. For instance, Vietnamese orthography contains peculiarities which originate in spelling conventions in various Romance languages (Haudricourt,

[11] This point was already noted by Niko Partanen on the basis of tests applying Persephone to orthographic data (personal communication, 2018).

Figure 3: Distribution of sentence durations in the Mwotlap narrative “Romanmangan, the fairy from the other world” (`https://doi.org/10.24397/pangloss-0002300`).

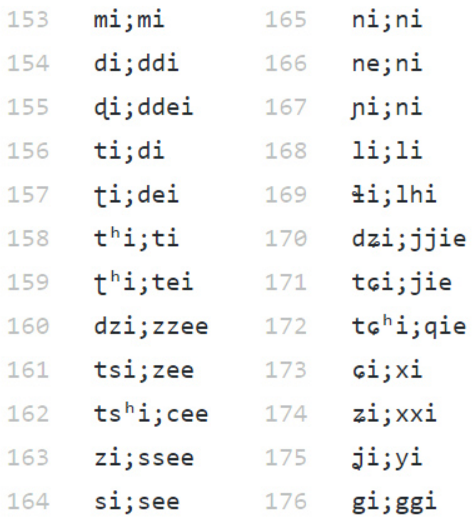


2010), and Na (Narua) orthography follows many conventions of the *Pinyin* romanization system for Standard (Beijing) Mandarin (which young speakers of Na learn at school). There can thus be a large distance between orthography and sound structure as manifested in the audio signal. This makes the training of an automatic transcription system, if not impossible, at least much more complicated.

3.3. Orthographic versus Phonemic Representations

To assess the difficulty of predicting an orthographic rather than a phonological transcription, we again turn to Na data. Roselle Dobbs and Xióng Yàn have developed an orthography for Yongning Na (Dobbs and Yàn, 2018). A complexity is that the orthography was devised as dialect-independent, so as to be an acceptable compromise between the various dialects of this highly diverse language area. As a result, some words are written in the orthography in a way that does not match their pronunciation in the dialect represented in the data set that we use here. For instance, ‘pretty’ and ‘pitiable’ are *nuxie* and *niggo*, respectively, in Na orthography, with different vowels in their first syllable, but the first syllable is phonologically identical in the dialect under consideration here. Such mismatches detract from the phonetic transparency of the transcriptions. Phonological transcriptions cannot be converted deterministically into orthographic transcriptions.

But these mismatches are absent from the orthographic transcriptions that we generated from IPA transcriptions. Phonological transcriptions in IPA (as available from the Pangloss Collection) can be readily converted into a simplified Na orthography by means of an algorithm that replaces IPA by orthography on a syllable-by-syllable basis.[12] A sample of the correspondences is shown in Figure 4.[13]



```
153   mi;mi        165   ni;ni
154   di;ddi       166   ne;ni
155   ɖi;ddei      167   ɲi;ni
156   ti;di        168   li;li
157   ʈi;dei       169   ɬi;lhi
158   tʰi;ti       170   dʑi;jjie
159   ʈʰi;tei      171   tɕi;jie
160   dzi;zzee     172   tɕʰi;qie
161   tsi;zee      173   ɕi;xi
162   tsʰi;cee     174   ʑi;xxi
163   zi;ssee      175   ʝi;yi
164   si;see       176   gi;ggi
```

Figure 4: Sample of the syllabic correspondences between IPA and orthography for Yongning Na.

In the real-life application of generating *bona fide* orthographic transcription for Na documents from IPA transcription, the automatically generated output needs to be improved manually to reflect the orthographic conventions for individual words, as provided in a dictionary of Yongning Na (Michaud, 2018). By contrast, in the tests conducted here, no such adjustments are performed. To distinguish the type of transcription that we generated from *bona fide* orthographic transcription, we will refer to the automatically converted transcriptions as ‘quasi-orthographic’ transcriptions. ‘Quasi-orthographic’ transcriptions have a relatively straightforward mapping to IPA – although it is not bijective, because some phonemic distinctions are not reflected in the orthography. For instance, as can be seen from Figure 4, three syllables, /ni/, /ne/ and /ɲi/, all correspond to *ni* in the orthography. The ‘quasi-orthographic’ transcriptions thus contain slightly fewer distinctions than the IPA notations.

With these caveats in mind, it is possible to compare the performance of a phonemic transcription system trained on the two kinds of transcriptions: phonemic and ‘quasi-orthographic’.

The two ways to transcribe data induce two different labels distributions: as shown in Figure 5, there are far more labels in phonological transcriptions, with a long tail. In orthographic transcriptions, the diversity of the phonemes is described by combinations of a small number of symbols and the model must discover and learn the structure of these combinations.

The results are clear: as shown in Table 5, while `Persephone` achieves very good results when predicting phonological transcriptions with a phoneme error rates of 13.0%, it cannot predict orthographic transcriptions of the same data (the validation error rate is above 90% even after 100 epochs).

These results suggest that orthographies, even with limited

[12] The code to convert phonetic transcription of Na into orthographic can be found at `https://github.com/alexis-michaud/na`.

[13] The syllables in Figure 4 do not carry tone. In view of the fact that tone varies greatly across Na dialects (Dobbs and La, 2016), the choice made in orthography development was to record only very limited tonal information. Automatic (rule-based) conversion currently disregards tone altogether. This topic is not relevant to the tests reported in this section, as those focus exclusively on vowels and consonants.

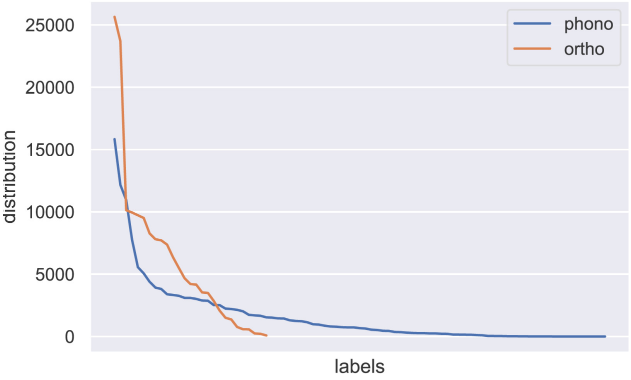


Figure 5: Distribution of the labels for the orthographic (`orth` curve) and phonological (`phono` curve) transcriptions.

labels	tone information	LER (test)
phonetic	✗	0.130
phonetic	✓	0.172
orthographic	✗	0.933

Table 5: Comparison of the performance achieved when using different type of transcriptions.

complexity, offer a less suitable basis for training a tool for automatic transcription. The excellent results reported about Tsuut'ina data transcribed in orthography (Michaud et al., 2019) are certainly due to the high degree of phonemic transparency of Tsuut'ina orthography. To labour the point: orthographic representations can be used: there is no technical gain in using International Phonetic Alphabet symbols rather than any other type of symbols as labels. The issue is not one of writing system *per se*: what matters is the degree of phonemic transparency of the transcription system. Successful application of `Persephone` in its current state requires a transcription that offers a high degree of phonemic transparency.

4. Conclusion and Perspectives

Towards a *Computational Language Documentation Cookbook* The tests reported here constitute a step towards a *Computational Language Documentation Cookbook* to determine which approaches are most appropriate to make the most of 'small' data sets for Automatic Speech Recognition tasks.

Perspectives for Multi-Speaker Tests and Transfer to other Languages Perspectives for further testing include attempting multi-speaker acoustic models (as against single-speaker setup as mostly studied so far) and model adaptation (pre-training a model on an extensive data set, then adapting it to another speaker, or even another language, on the basis of smaller amounts of data).

User interfaces for Natural Language Processing and Document Editing To empower a greater number of users to carry out tests, an easy-to-use interface is much wanted. Progress in this area is being made at a sustained pace (Foley et al., 2018; Foley et al., 2019). Plans for a general-purpose linguistic annotation backend (LAB) are also being carried out at Carnegie Mellon University (Neubig et al., 2018).

Perspectives for Collaboration with Language Archives: the Issue of Confidential Speaker Metadata One of the issues encountered in the course of the tests reported here is that available metadata are not as rich as one could wish from the point of view of computational tests. For instance, training an acoustic model in single-speaker mode requires knowledge of speaker identity, so as to be able to tease apart recordings from different speakers. But among the documents in the Nashta language available from the Pangloss Collection, seven are by "Anonymous woman" and eight are by "Anonymous man", and there is no telling, from the metadata, whether there is only one "Anonymous woman" or several. Some Romani and Slavic speakers from Greece choose "to remain anonymous due to the complexity of the political context in the country" (Adamou, 2016, v). (Language is a big component of social and ethnic identification, and hence a sensitive topic in many places.) In addition to speaker identity (at the basic level of distinguishing speakers from one another), the language consultants' age, linguistic history (proficiency in languages other than the one(s) that they use during the recording), and even their health record could be relevant parameters in combining documents into a training set. Those are pieces of information to which the investigator is to some extent privy: in the course of immersion fieldwork, one gets to learn a lot about the villages where one is staying. Such personal information must not be disclosed inconsiderately on the open Internet: one owes it to collaborators (language consultants) to protect their data. But destroying private information altogether is also a problem, as it detracts from the usefulness of the data. Use of data from language archives in Natural Language Processing (and in other areas of research) highlights the need for a more elaborate system for metadata management than is currently in place at the Pangloss Collection. In the same way as data can be kept private as long as necessary (the Pangloss Collection's host archive has provisions for keeping data offline for as long as fifty years for reasons of privacy, and as long as a century in the case of state secrets and documents deemed similarly sensitive), it would be a service to research if this archive would curate metadata that go beyond the Dublin Core and the metadata schema of the Open Language Archives Community (and manage the related access rights).

Phonemic Transcription beyond Phonemes: Leveraging the Full Extent of the Linguist's Annotations The research focus was placed here on the recognition of phonemes, but there is, technically, no notion of phoneme in the neural-network architecture, and labels that are not vowels, consonants, tones or other phonemic units can also be fed into the tool at training, and integrated to the acoustic model. Thus, tone-groupe boundaries, an important morpho-phonological landmark in Yongning Na (Michaud, 2017, 321-356), can be recognized by `Persephone` with good accuracy, and including tone-groupe boundaries improves overall performance.

5. Acknowledgments

We wish to express our gratitude to the language consultants who entrusted their living voices to visiting linguists, and the Pangloss Collection depositors, who generously chose to share the fruit of their hard work.

We are grateful to Oliver Adams, Laurent Besacier, Chris Cox, Hilaria Cruz, Benjamin Galliot, Nathan Hill, Michel Jacobson, Pat Littell, Niko Partanen, and Mandana Seyfeddinipur for useful discussions, to the three reviewers for detailed comments, and to Jesse Gates for proofreading. Errors and shortcomings are the authors' responsibility.

Financial support from grants ANR-10-LABX-0083 (*Laboratoire d'excellence* "Empirical Foundations of Linguistics", 2011-2024) and ANR-19-CE38-0015 ("Computational Language Documentation by 2025", 2019-2024) from *Agence Nationale de la Recherche* (France) is gratefully acknowledged.

6. Bibliographical References

Adamou, E. (2016). *A corpus-driven approach to language contact: Endangered languages in a comparative perspective*. Walter de Gruyter, Berlin.

Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., and Michaud, A. (2018). Evaluation Phonemic Transcription of Low-Resource Tonal Languages for Language Documentation. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Borgman, C. L. (2015). *Big data, little data, no data: scholarship in the networked world*. MIT Press, Cambridge, MA.

Bouquiaux, L. and Thomas, J. (1971). *Enquête et description des langues à tradition orale. Volume I : l'enquête de terrain et l'analyse grammaticale*. Société d'études linguistiques et anthropologiques de France, Paris. 3 volumes.

Brugman, H. and Russel, A. (2004). Annotating multimedia/multi-modal resources with ELAN. In *Proceedings of LREC 2004*.

Dixon, R. M. (2007). Field linguistics: A minor manual. *Sprachtypologie und Universalienforschung*, 60(1):12–31.

Dobbs, R. and La, M. (2016). The two-level tonal system of Lataddi Narua. *Linguistics of the Tibeto-Burman Area*, 39(1):67–104.

Dobbs, R. and Yàn, X. (2018). Yongning Narua orthography: users' guide and developers' notes. https://halshs.archives-ouvertes.fr/halshs-01956606/.

Foley, B., Arnold, J., Coto-Solano, R., Durantin, G., and Ellison, T. M. (2018). Building speech recognition systems for language documentation: the CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *Proceedings of the 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU), 29-31 August 2018*, pages 200–204, Gurugram, India. ISCA.

Foley, B., Rakhi, A., Lambourne, N., Buckeridge, N., and Wiles, J. (2019). Elpis, an accessible speech-to-text tool. In *Proceedings of Interspeech 2019*, pages 306–310, Graz.

François, A. (2003). *La sémantique du prédicat en mwotlap, Vanuatu*, volume 84. Peeters, Louvain.

Gerstenberger, C., Partanen, N., Rießler, M., and Wilbur, J. (2017). Instant annotations–Applying NLP methods to the annotation of spoken language documentation corpora. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*, pages 25–36, St. Petersburg, Russia.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Haudricourt, A.-G. (2010). The origin of the peculiarities of the Vietnamese alphabet. *Mon-Khmer Studies*, 39:89–104.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Yoshua Bengio et al., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Littell, P., Kazantseva, A., Kuhn, R., Pine, A., Arppe, A., Cox, C., and Junker, M.-O. (2018). Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632.

Liégeois, L., Etienne, C., Parisse, C., Benzitoun, C., and Chanard, C. (2016). Using the TEI as pivot format for oral and multimodal language corpora. *Journal of the Text Encoding Initiative*, 10.

Lust, B. C., Blume, M., Pareja-Lora, A., and Chiarcos, C. (2019). Development of Linguistic Linked Open Data resources for collaborative data-intensive research in the language sciences: An introduction. In *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*. MIT Press, Cambridge, MA.

Maurel, L. (2016). Quel statut pour les données de la recherche après la loi numérique ? https://scinfolex.com/2016/11/03/quel-statut-pour-les-donnees-de-la-recherche-apres-la-loi-numerique/.

Michailovsky, B., Mazaudon, M., Michaud, A., Guillaume, S., François, A., and Adamou, E. (2014). Documenting and researching endangered languages: The Pangloss Collection. *Language Documentation and Conservation*, 8:119–135.

Michaud, A., Adams, O., Cohn, T., Neubig, G., and Guillaume, S. (2018). Integrating automatic transcription into the language documentation workflow: experiments with Na data and the Persephone toolkit. *Language Documentation and Conservation*, 12:393–429.

Michaud, A., Adams, O., Cox, C., and Guillaume, S. (2019). Phonetic lessons from automatic phonemic transcription: preliminary reflections on Na (Sino-Tibetan) and Tsuut'ina (Dene) data. In *Proceedings of ICPhS XIX (19th International Congress of Phonetic Sciences)*, Melbourne.

Michaud, A. (2017). *Tone in Yongning Na: lexical tones and morphotonology*. Language Science Press, Berlin.

Michaud, A. (2018). *Na (Mosuo)-English-Chinese dictionary*. Lexica, Paris.

Moran, S. and Cysouw, M. (2018). *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Translation and Multilingual Natural Language Processing. Language Science Press, Feb.

Neubig, G., Littell, P., Chen, C.-Y., Lee, J., Li, Z., Lin, Y.-H., and Zhang, Y. (2018). Towards a general-purpose linguistic annotation backend. *arXiv:1812.05272 [cs]*, December. arXiv: 1812.05272.

Newman, P. and Ratliff, M. (2001). *Linguistic fieldwork*. Cambridge University Press, Cambridge.

Niebuhr, O. and Michaud, A. (2015). Speech data acquisition: The underestimated challenge. *KALIPHO - Kieler Arbeiten zur Linguistik und Phonetik*, 3:1–42.

Schmidt, T. (2011). A TEI-based approach to standardising spoken language transcription. *Journal of the Text Encoding Initiative*, (1).

Seifart, F., Evans, N., Hammarström, H., and Levinson, S. C. (2018). Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

Gary F. Simons et al., editors. (2017). *Ethnologue: languages of the world*. SIL International, Dallas, twentieth edition edition.

Sperber, M., Neubig, G., Fügen, C., Nakamura, S., and Waibel, A. (2013). Efficient speech transcription through respeaking. In *Proceedings of Interspeech 2013*, pages 1087–1091, Lyon.

Strunk, J., Schiel, F., and Seifart, F. (2014). Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 3940–3947, Reykjavik. European Language Resources Association (ELRA).

Thieberger, N. (2017). LD&C possibilities for the next decade. *Language Documentation and Conservation*, 11:1–4.

van Esch, D., Foley, B., and San, N. (2019). Future directions in technological support for language documentation. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, page 3, Honolulu, Hawai'i.

Whalen, D. H. and McDonough, J. (2019). Under-resourced languages: Phonetic results from language archives. In William F. Katz et al., editors, *The Routledge Handbook of Phonetics*.

Woodbury, T. (2003). Defining documentary linguistics. In Peter Austin, editor, *Language documentation and description*, volume 1, pages 35–51. School of African and Oriental Studies, London.