Phonemic transcription of low-resource languages: To what extent can preprocessing be automated?
Guillaume Wisniewski, Alexis Michaud, Séverine Guillaume

To cite this version:
Guillaume Wisniewski, Alexis Michaud, Séverine Guillaume. Phonemic transcription of low-resource languages: To what extent can preprocessing be automated?. 2020. hal-02513914v1

HAL Id: hal-02513914
https://shs.hal.science/hal-02513914v1
Preprint submitted on 26 Mar 2020 (v1), last revised 27 May 2020 (v3)
Phonemic Transcription of Low-Resource Languages:
To What Extent can Preprocessing be Automated?

Guillaume Wisniewski¹, Alexis Michaud², Séverine Guillaume²
(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

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. 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.

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 a precise speech processing task. 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. They constitute “eclectic data collections” rather than “systematically annotated corpora” [Gerstenberger et al., 2017, 26]. Preprocessing typically involves retrieving pieces of information that had not been encoded according to computational standards in the first place.

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 approach 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 range 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 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, 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. 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

\[1\] We consider the 41 usual fbank features as well as their first and second derivative.
Building an automatic phonemic transcription system therefore requires to, first, segment the transcriptions into sequences of phonemes. Transcriptions often contain 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" 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" 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, transcriptions extracted from the Yongning Na documents in the Pangloss Collection were segmented into sequences of phonemes by means of a set of handcrafted 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, 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.

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).

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 "phone" (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 and orthographic transcription is up to the contributor. Differences in performance of automatic transcription for phonological and orthographic transcription is up to the contributor. Differences in performance of automatic transcription for phonological and orthographic transcription is up to the contributor. Differences in performance of automatic transcription for phonological and orthographic transcription is up to the contributor. Differences in performance of automatic transcription for phonological
extension of the Persephone toolkit, at https://github.com/gw17/sltu_corpora 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 (out of a total 170 in the Pangloss Collection).

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.

3. Experimental Results

We conducted three series of experiments to assess:

- the impact of considering different languages.
- 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.

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), whitespaces identify grapheme clusters. Note that, in both segmentations, punctuation marks are removed: in the current setup, no attempts were made at predicting them.

| 1. si!dz!-[t]h!n!], | [t]a!r! | [t]h!ur!-by!t!| | da!- ky!-ma! | 
| 2. s!i!-!dz!-!i!-!t!h! | 3. s!j!-!dz!-!i!-!t!h! | 4. s!i!-!dz!-!i!-!t!h! |

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 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 (out of a total 170 in the Pangloss Collection).
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).

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

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.

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

<table>
<thead>
<tr>
<th>segmentation method</th>
<th>LER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams et al. (2018) (phonemes)</td>
<td>0.130</td>
</tr>
<tr>
<td>Adams et al. (2018) (phonemes + tones)</td>
<td>0.172</td>
</tr>
<tr>
<td>Grapheme cluster (phonemes + tones)</td>
<td>0.186</td>
</tr>
</tbody>
</table>

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 ts\(^6\), 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 /ts\(^3\)/). 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 quality** 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, 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” has a duration of 1,890 seconds, and is divided into 212 sentences. The dis-

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.

<table>
<thead>
<tr>
<th>Language</th>
<th>LER on train set</th>
<th>LER on test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>nru</td>
<td>0.016</td>
<td>0.186</td>
</tr>
<tr>
<td>lif</td>
<td>0.167</td>
<td>0.368</td>
</tr>
<tr>
<td>ers</td>
<td>0.218</td>
<td>0.383</td>
</tr>
<tr>
<td>tvk</td>
<td>0.822</td>
<td>0.818</td>
</tr>
<tr>
<td>mkd</td>
<td>0.926</td>
<td>0.926</td>
</tr>
<tr>
<td>mlv</td>
<td>0.944</td>
<td>0.932</td>
</tr>
<tr>
<td>nep</td>
<td>0.98</td>
<td>0.965</td>
</tr>
</tbody>
</table>

\(^{6}\)For example: [https://doi.org/10.24397/pangloss-0004091](https://doi.org/10.24397/pangloss-0004091)

\(^{7}\)For example: [https://doi.org/10.24397/pangloss-0002300](https://doi.org/10.24397/pangloss-0002300)
Figure 2: Learning curve (train & validation sets) for the different languages considered in our experiments

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

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.

Number of labels

There are large differences in the number of labels and the quantity of training data between 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, 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, 2010), and Na (Narua) orthography follows many con-
ventions 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 Xiòng Yán have developed an orthography for Yongning Na [Dobbs and Yán, 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 *nuxix* 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. A sample of the correspondences is shown in Figure 4.

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 /ñi/, all correspond to *ni* in the orthography. The ‘quasi-orthographic’ transcriptions thus contain slightly fewer distinctions than the IPA

![Figure 4: Sample of the syllabic correspondences between IPA and orthography for Yongning Na.](image)

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 complexity, offer a less suitable basis for training a tool for automatic transcription. The excellent results reported

![Figure 5: Distribution of the labels for the orthographic (orth curve) and phonological (phono curve) transcriptions.](image)
<table>
<thead>
<tr>
<th>labels</th>
<th>tone information</th>
<th>LER (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>phonetic</td>
<td>✗</td>
<td>0.130</td>
</tr>
<tr>
<td>phonetic</td>
<td>✓</td>
<td>0.172</td>
</tr>
<tr>
<td>orthographic</td>
<td>✗</td>
<td>0.933</td>
</tr>
</tbody>
</table>

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

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 for further work

Towards a Computational Language Documentation Cookbook The tests reported here only constitute a small series of steps 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 For a greater number of users to be able to carry out tests on various data sets, 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 Nasha 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). (Similar situations are not uncommon: language is a big component of social and ethnic identification, and hence a sensitive topic in many areas.) In addition to speaker identity (at the basic level of distinguishing different 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 (the linguist or anthropologist) is often privy to some extent: 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 partner 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 et al., 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. References


Niebuhr, O. and Michaud, A. (2015). Speech data acquisi-


