A Scalable Method for Preserving Oral Literature from Small Languages

Steven Bird

Dept of Computer Science and Software Engineering, University of Melbourne
Linguistic Data Consortium, University of Pennsylvania

Abstract. Can the speakers of small languages, which may be remote, unwritten, and endangered, be trained to create an archival record of their oral literature, with only limited external support? This paper describes the model of “Basic Oral Language Documentation”, as adapted for use in remote village locations, far from digital archives but close to endangered languages and cultures. Speakers of a small Papuan language were trained and observed during a six week period. Linguistic performances were collected using digital voice recorders. Careful speech versions of selected items, together with spontaneous oral translations into a language of wider communication, were also recorded and curated. A smaller selection was transcribed. This paper describes the method, and shows how it is able to address linguistic, technological and sociological obstacles, and how it can be used to collect a sizeable corpus. We conclude that Basic Oral Language Documentation is a promising technique for expediting the task of preserving endangered linguistic heritage.

1 Introduction

Preserving the world’s endangered linguistic heritage is a daunting task, far exceeding the capacity of existing programs that sponsor the typical 2-5 year “language documentation” projects. In recent years, digital voice recorders have reached a sufficient level of audio quality, storage capacity, and ease of use, to be used by local speakers who want to record their own languages. This paper investigates the possibility of putting the language preservation task into the hands of the speech community. With suitable training, they can be equipped to record a variety of oral discourse genres from a broad cross-section of the speech community, and then provide additional content to permit the recordings to be interpreted by others who do not speak the language. The result is an audio collection with time-aligned translations and transcriptions, a substantial archival resource.

This paper describes a method for preserving oral discourse, originating in field recordings made by native speakers, and generating a variety of products including digitally archived collections. It addresses the problem of unwritten languages being omitted from various ongoing efforts to collect language resources for ever larger subsets of the world’s languages [1]. The starting point is Reiman’s work [2], modified and refined so that it uses appropriate technology.
for Papua New Guinea, and so that it can scale up easily. The method has been tested with Usarufa, a language of Papua New Guinea. Usarufa is spoken by about 1200 people, in a cluster of six villages in the Eastern Highlands Province, about 20km south of Kainantu (06°25’S, 145°39’E). There are probably no fluent speakers of Usarufa under the age of 25; only the oldest speakers retain the rich vocabulary for animal and plant species, and for a variety of cultural artefacts and traditional practices. Some texts including the New Testament and a grammar have been published in Usarufa [3]. However, only a handful of speakers are literate in the language.

2 Basic Oral Language Documentation

2.1 Audio Capture

The initial task in Basic Oral Language Documentation (BOLD) is audio capture. Collecting the primary text from individual speakers is straightforward. They press the record button, hold the voice recorder a few inches from their mouth, and begin by giving their name, the date, and location. The person operating the recorder may or may not be the speaker.

Collecting a dialogue involves two speakers plus someone to operate the voice recorder (who may be a dialogue participant). The operator can introduce the recording and hold the recorder in an appropriate position between the participants. The exchange shown in Fig. 1 involved a language worker (left), the author, and a village elder. The dialogue began with an extended monologue from the man on the left, explaining the purpose of the recording and asking the other man to recount a narrative, followed by some conversation for clarification, followed by an extended monologue from the man on the right. The voice recorder was moved closer to the speaker during these extended passages, but returned to the centre during conversational sections. In most cases, the person operating the recorder was a native speaker of Usarufa, and was also participating in the dialogue. The operator was instructed not to treat the recorder like a hand-held microphone, moved deliberately between an interviewer and interviewee to signal turns in the