Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 8 – 14

Procedia
Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Breaking the Unwritten Language Barrier: The BULB Project

Gilles Adda[a,*], Sebastian Stüker[b,1], Martine Adda-Decker[a,c], Odette Ambouroue[d], Laurent Besacier[e], David Blachon[e], Hélène Bonneau-Maynard[a], Pierre Godard[a], Fatima Hamlaoui[f], Dmitry Idiatov[d], Guy-Noël Kouarata[c], Lori Lamel[a], Emmanuel-Moselly Makasso[f], Annie Rialland[c], Mark Van de Velde[d], François Yvon[a], Sabine Zerbian[g]

[a]LIMSI, CNRS, Université Paris-Saclay, France
[b]Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany
[c]LPP, CNRS-Paris 3/Sorbonne Nouvelle, France
[d]Langage, Langues et Cultures d'Afrique Noire Laboratory (LLACAN), France
[e]Laboratoire d'Informatique de Grenoble (LIG)/GETALP group, France
[f]Zentrum für Allgemeine Sprachwissenschaft (ZAS), Germany
[g]Universität Stuttgart/Institut für Linguistik, Germany

Abstract

The project *Breaking the Unwritten Language Barrier* (BULB), which brings together linguists and computer scientists, aims at supporting linguists in documenting unwritten languages. In order to achieve this we develop tools tailored to the needs of documentary linguists by building upon technology and expertise from the area of natural language processing, most prominently automatic speech recognition and machine translation. As a development and test bed for this we have chosen three less-resourced African languages from the Bantu family: Basaa, Myene and Embosi. Work within the project is divided into three main steps:

1) **Collection** of a large corpus of speech (100h per language) at a reasonable cost. For this we use standard mobile devices and a dedicated software—*Lig-Aikuma*. After initial recording, the data is re-spoken by a reference speaker to enhance the signal quality and orally translated into French.

2) **Automatic transcription** of the Bantu languages at phoneme level and the French translation at word level. The recognized Bantu phonemes and French words will then be automatically aligned.

3) **Tool development**. In close cooperation and discussion with the linguists, the speech and language technologists will design and implement tools that will support the linguists in their work, taking into account the linguists' needs and technology's capabilities.

Keywords: Language documentation, automatic phonetic transcription, unwritten languages, automatic alignment

[*] Corresponding author. Tel.: +33-169858180 ; fax: +33-169858080.

[1] apart from the first two authors, the names are in alphabetical order

E-mail address: Gilles.Adda@limsi.fr

doi:10.1016/j.procs.2016.04.023

1. Introduction

It is well known that only a very limited proportion of the languages spoken in the world is covered by technology or by scientific knowledge. For technology, only normative productions of very few languages in very few situations are mastered. The technological divide is wide considering the languages spoken: we have a minimally adequate quantity of data for less than 1% of the world's 7000 languages. Most of the world's everyday life speech stems from languages which are essentially unwritten and we include in these languages ethnolects as well as sociolects such as many regional varieties of Arabic, Shanghainese, slang ... There are thousands of endangered languages for which hardly any documentation exists and time is running out before they disappear: some linguists estimate that half of the presently living languages will become extinct in the course of this century[1,2,3]. Even with the upsurge of documentary linguistics[4,5], it is not realistic to expect that the documentary linguistics community will be able to document all these languages before they disappear without the help of automatic processing—given the number of languages involved and the amount of human effort required for the "creation, annotation, preservation, and dissemination of transparent records of a language"[5].

In this article, we present the French-German ANR-DFG project *Breaking the Unwritten Language Barrier* (BULB `http://www.bulb-project.org/`), whose goal is to develop within three years a methodology and corresponding processing tools to achieve efficient automatic processing of unwritten languages, with a first application on three mostly unwritten African languages of the Bantu family (Basaa, Myene and Embosi, see Section 3.1 for more detail on the choice of languages). Among the languages in danger of disappearing, many of those that have not yet been properly documented are non-written languages. The lack of a writing system makes these languages a challenge for both documentary linguists and natural language processing (NLP) technology. In the present project, we therefore conduct the necessary research to obtain the technology that is presently missing to efficiently document unwritten languages. Work within the project is divided into three main steps:

1. **Collection** of a large corpus of speech (100h per language) at a reasonable cost. For this we use standard mobile devices and a dedicated software called Lig-Aikuma. After initial recording, the data is re-spoken by a reference speaker to enhance the signal quality, and orally translated into French.
2. **Automatic transcription** of the Bantu languages at phoneme level and the French translation at word level, followed by the **automatic alignment** of the recognized Bantu phonemes and the French words.
3. **Tool development**. In close cooperation and discussion with the linguists, the speech and language technologists will design and implement tools that will support the linguists in their work, taking into account the linguists' needs and technology's capabilities.

At this stage of the project (end of first year) we have focused on the data acquisition, and began to work on automatic transcription and alignment using the data available (see section 3.3).

2. NLP Technology for Language Documentation

2.1. Language Independent Phoneme and Articulatory Feature Recognition

Systems for language independent phoneme recognition often utilize multilingual models[6]. The idea behind this approach is to identify phonemes that are common to multiple languages, e.g., by using global phoneme sets, such as the International Phonetic Alphabet (IPA). Models for phonemes that are common to multiple languages share all the training material from those languages. A multilingual model can be applied to any new language that was not originally included in the training languages. Phonemes in the new language that are not covered by the multilingual model need to be mapped appropriately.

Alternatively to phonemes, methods exist to recognize articulatory features across languages, either with monolingual models from many languages or with multilingual models trained on many languages[7]. The advantage of multilingual models for articulatory features is that the coverage of the model for the articulatory features in a new language is generally higher than it is for phonemes and that they can be recognized more robustly across languages.