Innovative technologies for under-resourced language documentation: The BULB Project

Gilles Adda, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Héì Ene Bonneau-Maynard, Elodie Gauthier, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, et al.

Innovative technologies for under-resourced language documentation: The BULB Project

Gilles Adda[0,a], **Martine Adda-Decker**[a,b], **Odette Ambouroue**[c], **Laurent Besacier**[d], **David Blachon**[d], **Hélène Bonneau-Maynard**[a], **Elodie Gauthier**[d], **Pierre Godard**[a], **Fatima Hamlaoui**[e], **Dmitry Idiatov**[c], **Guy-Noël Kouarata**[b], **Lori Lamel**[a], **Emmanuel-Moselly Makasso**[e], **Annie Rialland**[b], **Sebastian Stüker**[f], **Mark Van de Velde**[c], **François Yvon**[a], **Sabine Zerbian**[g]

(a) LIMSI, CNRS, Université Paris-Saclay, France
(b) LPP, CNRS-Paris 3/Sorbonne Nouvelle, France
(c) Langage, Langues et Cultures d'Afrique Noire Laboratory (LLACAN), France
(d) Laboratoire d'Informatique de Grenoble (LIG)/GETALP group, France
(e) Zentrum für Allgemeine Sprachwissenschaft (ZAS), Germany
(f) Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, 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 will 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. 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. The data collection has begun for the three languages. For this we use standard mobile devices and a dedicated software—**LIG-AIKUMA**, which proposes a range of different speech collection modes (recording, respeaking, translation and elicitation). LIG-AIKUMA 's improved features include a smart generation and handling of speaker metadata as well as respeaking and parallel audio data mapping.

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

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. When speech is less normative (due to age, spontaneity, speaker's origin, pathology,. . .) performances drop significantly, with multiplicative effects in case of multiple factors(Gerosa and Giuliani, 2008); this also reflects weaknesses in modelisation. The technological divide is even wider 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.[1]

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 (Nettle and Romaine, 2000; Crystal, 2002; Janson, 2003). Even with the upsurge of documentary linguistics (Himmelmann and universität Bochum, 2002; Woodbury, 2011), 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" (Woodbury, 2011).

In this article, we present the French-German ANR-DFG project *Breaking the Unwritten Language Barrier* (BULB), whose goal it is to develop 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). 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 will therefore conduct the necessary research to obtain the technol-

[0] the names are in alphabetical order

[1] We include in these languages ethnolects as well as sociolects such as many regional varieties of Arabic, Shanghainese, slang . . .

ogy 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**. LIG-AIKUMA proposes a range of different speech collection modes (recording, respeaking, translation and elicitation). LIG-AIKUMA 's improved features include a smart generation and handling of speaker metadata as well as respeaking and parallel audio data mapping (see (Blachon et al., 2016) for further details). 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. The collected oral data (Bantu originals and French translations) contain the necessary information to document the studied languages. Phonetic alignments are highly valuable for large scale acoustic-phonetic studies, phonological and prosodic data mining and dialectal variations studies; cross-language alignments may also prove very useful for morphological studies, vocabulary and pronunciation elaboration.

3. **Tool development**. Tools will be built upon all these data and alignments. 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.

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 (Kohler, 1996). 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. By pooling phoneme sets and data from multiple languages one achieves two effects. First, the number of phonemes covered by a multilingual model is in general larger than that of a monolingual model. Second, by pooling data from multiple languages the models can become more robust to slight variations in the pronunciation of phonemes in different languages that are nonetheless denoted by the same symbol. 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 (Stüker et al., 2003). 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.

2.2. Word Discovery by Word-to-Phoneme-Alignment

The feasibility of automatically discovering word units (as well as their pronunciations) in an unknown (and unwritten) language without any supervision was examined by (Besacier et al., 2006). This goal was achieved by unsupervised aggregation of phonetic strings into word forms from a continuous flow of phonemes (or from a speech signal) using a monolingual algorithm based on cross-entropy. This approach lead to almost the same performance as the baseline approach, while being applicable to any unwritten language.
(Stüker, 2008) introduced a phone-based speech translation approach that made use of cross-lingual supervision. This approach works on a scenario in which a human translates the audio recordings of the unwritten language into a written language. Alignment models as used in machine translation (Brown et al., 1993; Och and Ney, 2003) were then learned on the resulting parallel corpus consisting of foreign phone sequences and their corresponding English translation. (Stüker et al., 2009) combined this approach with the monolingual approach above and also did contrastive comparisons. (Stahlberg et al., 2012) and (Stahlberg et al., 2013) then continued to work on this approach by enhancing alignment model for the task and examined the impact of the choice of written language to which the phoneme sequence is aligned.
The proposed phone-based approach lead to almost the same performance as the baseline approach, while being theoretically applicable to any unwritten language. In this work, the unsupervised aggregation of phonemes into pseudo-words was performed using a monolingual algorithm based on cross-entropy. It is however appealing to also make use of bilingual resources, such as parallel corpora, to extract these pseudo-words inventories. Such corpora can be obtained in a scenario in which a human translator produces utterances in the (unwritten) target language from English prompts. In such a setting, it is possible to add the English source to help the word discovery process, e.g. by using statistical word alignment models such as (Brown et al., 1993; Och and Ney, 2003). An overview of various techniques for pseudo-word discovery using monolingual or bilingual information is presented in (Stüker et al., 2009), where the effectiveness of this approach was demonstrated. More recently, in (Stahlberg et al., 2012), an extended version of the alignment model IBM Model3 (called Model3P) was proposed to improve the aggregation of the phoneme strings. (Stahlberg et al., 2013) examined the choice of the concrete (well-resourced) language to help the pseudo-word discovery process. Phonetic transcriptions of target language words using Model3P were deduced and then introduced in the pronunciation dictionary.

Working with a similar goal in mind, and using bilingual information in order to jointly learn the segmentation of a target string of characters (or phonemes) and their alignment to a source sequence of words, (Xu et al., 2008; Nguyen et al., 2010) are building on Bayesian monolingual segmentation models introduced by (Goldwater et al., 2006) and further expanded in (Mochihashi et al., 2009). This trend of research has become increasingly active in the past years, moving from strategies using segmentation as a preprocessing to the alignment steps, to models aiming at jointly learning relevant segmentation and alignment. (Adams et al., 2015) reports performance improvements for the latter approach on a bilingual lexicon induction task, with the additional benefit of achieving high precision even on a very small corpus, which is of particular interest in the context of BULB.

Many questions still need to be addressed. Implicit choices are usually made through the way data are specified and represented. Taking, for example, tones into account, prosodic markers, or even a partial bilingual dictionary, would require different kinds of input data, and the development of models able to take advantage of this additional information.

A second observation is that most attempts to learn segmentation and alignments need to inject some prior knowledge about the desired form of the linguistic units which should be extracted. This is because most machine learning schemes deployed in the literature tend to otherwise produce degenerated and trivial (over-segmented or conversely under-segmented) solutions. The additional constraints necessary to control such phenomena are likely to greatly impact the nature of the units that are identified. Supporting the documentation of endangered languages within the framework of BULB should lead us to consequently question as systematically as possible the linguistic validity of those constraints and the results they produce. The Adaptor Grammar framework (Johnson et al., 2007; Johnson, 2008), which enables the specification of high-level linguistic hypotheses appears to be of particular interest in our context. Another important aspect of the endeavor we are facing lies in the noisy nature of the input produced by the phonemicization of the unwritten language. Processing a phoneme lattice instead of a phonemic transcription, following the work of (Neubig et al., 2010), seems to be a promising strategy here.

More generally, a careful inventory of priors derived from the linguistic knowledge at our disposal should be undertaken. This is especially true regarding cross-lingual priors we can postulate about French on the one hand, and Basaa, Myene and Embosi on the other hand: for lack of taking such priors into account, it is dubious that general purpose unsupervised learning techniques will succeed in delivering any usable linguistic information.

2.3. Preservation of Unwritten Languages by Advanced Technologies

(Bird, 2010) described the model of "Basic Oral Language Documentation", as adapted for use in remote village locations, which are "far from digital archives but close to endangered languages and cultures". Speakers of a small Papuan language were trained and observed during a six weeks period. A technique called re-speaking, initially introduced by (Woodbury, 2003), was used. Re-speaking involves listening to an original recording and repeating what was heard carefully and slowly. This results in a secondary recording that is much easier to transcribe later on (transcription by a linguist or by a machine). The reason is that the initial speech may be too fast, the recording level may be too low, and background noise may degrade the content. For instance, in the context of recording traditional narratives, elderly speakers are often required (and they may have a weak voice, few teeth, etc.) compromising the clarity of the recording (Hanke and Bird, 2013). In (Bird and Chiang, 2012), the use of statistical machine translation is presented as a way to support the task of documenting the world's endangered languages. An analogy is made between the primary resource of statistical translation models – bilingual aligned text – and the primary artefact collected in documentary linguistics – recordings of the language of interest, together with their translation. The authors suggest exploiting this similarity to improve the quantity and quality of documentation for a language. Details on the mobile application (called AIKUMA) are given in (Hanke and Bird, 2013). AIKUMA is an Android application that supports the recording of audio sources, along with phrase-by-phrase oral translation. In their paper, the concept of respeaking was extended to produce oral translations of the initial recorded material. Oral translation was performed by listening to a segment of audio in a source language and spontaneously producing a spoken translation in a second language.

Finally, it is also worth mentioning the work of (Kempton and Moore, 2014), who suggest the use of advanced speech technologies to help field linguists in their work. More precisely, they proposed a machine-assisted approach for phonemic analysis of under-resourced and under-documented languages. Several procedures were investigated (phonetic similarity, complementary distribution, and minimal pairs) and compared.

During the first year of BULB, features were added to the original AIKUMA app to facilitate the collection of parallel speech data required in the project. The resulting app, called LIG-AIKUMA , is described in section 3..

3. LIG-AIKUMA recording application

3.1. Motivations and specifications

Within the BULB project, the use of LIG-AIKUMA is associated to a set of use cases, identified by a series of operations to perform. The first one is a basic audio recording. The next ones consist of respeaking and translation. The last one concerns elicitation of speech after the display of text, image or video.

The use of LIG-AIKUMA is driven by those objectives, so one of the changes to be performed to the application AIKUMA is about the user interface that should identify the modes associated and focus on them. More generally, several requirements were made for the application to be quick and easy to use: save and load the metadata of the latest recording for saving the time of filling them in the form

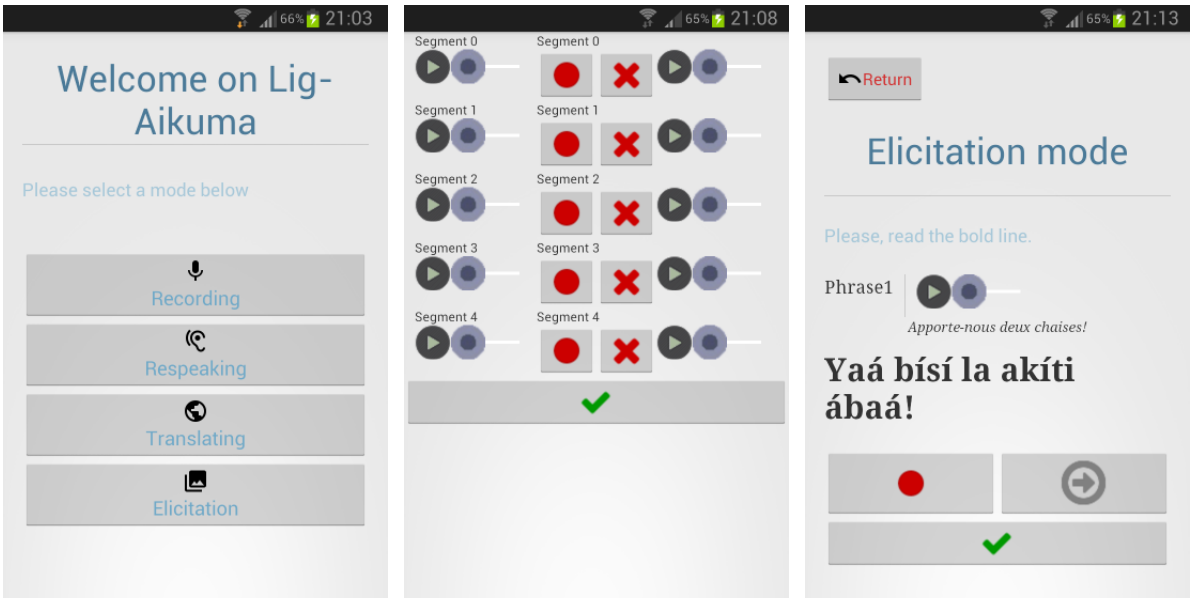

Figure 1: Screenshots from the LIG-AIKUMA application: from left to right, i) the home view ; ii) the summary view after respeaking is done, the speaker may play and edit every segment; iii) the elicitation mode

again; give a better feedback on the respeaking once it is done, etc.

3.2. Recording modes

The core features of the initial AIKUMA for recording, respeaking and translation have been kept, along with the storage of metadata about the speaker; also, some parts of the interface have been reused.
On top of that, new developments have focused on the setup of 4 modes, dedicated to specific tasks of speech recording. The home view is illustrated on Figure 1 (left). As one can see, the following four modes are identified:

- Free recording of spontaneous speech,
- Respeaking a recording (previously recorded with the app or loaded from a wav file): the respeaking allows now to listen (optionnally) to the latest recording segment so as to check it and respeak it if needed, before going to the next segment. Also, once the respeaking is finished, a summary view displays the new segments and their corresponding original segments and allows to (optionnally) listen to or respeak any of them before finishing the session. On Figure 1 (middle), one can see that original segments are aligned with respoken ones. Both can be played while the latter can also be recorded if necessary, which is useful for double check and error correction,
- Translating a recording (previously recorded or loaded): same features as for the respeaking mode except that the source and target languages must be different,
- Elicitating speech from a text file (image and video media will follow very soon): the user loads a text file within the app, then reads the sentence, speaks, listens to the recording for checking and goes to the next sentence, etc. This mode was specifically required for the data collection which took place in Congo-Brazaville during summer 2015. Figure 1 (right) illustrates the text elicitation mode.

3.3. Current state of development

The interface has been adapted for the large screens of tablets (10 inches), so the app works both on Android powered smartphones and tablets. Apart from the specifications, based on multiple discussions with linguists colleagues, this new version was developed in approximately 3 man/months and generated 5000+ lines of code. All the new code has been put on the LIG forge and is accessible open source[2] for use or development on demand. The application LIG-AIKUMA has been successfully tested on different devices (including Samsung Galaxy SIII, Google Nexus 6, HTC Desire 820 smartphones and a Galaxy Tab 4 tablet). Figure 2 illustrates the use of the LIG-AIKUMA tablet to collect Mboshi verb conjugations (left) or more free conversations including several speakers (right).
Users who just want to use the app without access to the code can download it directly from the forge direct link[3].

4. Documentation of three Bantu Languages

4.1. Bantu languages

In BULB, three typologically diverse northwestern Bantu languages were selected, which stem from different Guthrie zones (areal-genetic groupings, (Guthrie, 1948)): Basaa (A43, Cameroon), Myene (B10, Gabon) and Embosi (C25, Congo-Brazzaville). The Bantu family is one of the largest

[2] https://forge.imag.fr/projects/lig-aikuma/
[3] https://forge.imag.fr/frs/download.php/706/MainActivity.apk

Figure 2: Examples of the use of Aikuma on Android tablets for data collection: elicited verb conjugations spoken by a native Mboshi woman (left) and free conversations involving several speakers (right).

genera in the world and most of the genetic and typological diversity within this family can be found in the northwestern part of the domain, closest to the Bantu homeland. As northwestern Bantu languages are spoken in the so-called *fragmentation belt*, – a zone of extreme linguistic diversity – they differ from their eastern and southern Bantu relatives such as Swahili, Sotho or Zulu in that they are much less studied, protected and resourced.

Our three Bantu languages however have in common that they are relatively well described, as there are competent native-speaker linguists working on each of them and, at least in the case of Myene, some basic electronic resources are already available (albeit in need of further development to make them suitable for corpus-based linguistic analyses). This was an important criterion in our choice of languages, as the available linguistic analyses will allow us to test the efficiency and improve the outcome of our new tools.

4.2. Three under resourced Bantu languages

Basaa, which is spoken by approximately 300 000 speakers (SIL, 2005) from the "Centre" and "Littoral" regions of Cameroon, is the best studied of our three languages. The earliest lexical and grammatical description of Basaa goes back to the beginning of the twentieth century (Rosenhuber, 1908) and the first Basaa-French dictionary was developed over half a century ago (Lemb and de Gastines, 1973). Several dissertations have focused on various aspects of Basaa (Bot ba Njock, 1970; Makasso, 2008) and the language also benefits from recent and ongoing linguistic studies (Dimmendaal, 1988; Hyman, 2003; Hamlaoui and Makasso, 2015).

Myene, a cluster of six mutually intelligible varieties (Adyumba, Enenga, Galwa, Mpongwe, Nkomi and Orungu), is spoken at the coastal areas and around the town of Lambarene in Gabon. The current number of Myene speakers is estimated at 46 000 (Lewis et al., 2013). The language is presently considered as having a "vigorous" status, but the fact that no children were found that could participate in a study on the acquisition of Myene suggests that the language is already endangered. A basic grammatical description of the Orungu variety (Ambouroue, 2007) is available, as well as a few articles on aspects of the phonology, morphology and syntax of Myene ((Van de Velde and Ambouroue, 2011) and references therein).

Our third and last language, **Embosi**, originates from the "Cuvette" region of the Republic of Congo and is also spoken in Brazzaville and in the diaspora. The number of Embosi speakers is estimated at 150 000 (Congo National Inst. of Statistics, 2009). A dictionary (Kouarata, 2000) is available and, just like Basaa and Myene, the language benefits from recent linguistic studies (Amboulou, 1998; Embanga Aborobongui, 2013).

From a linguistic perspective, the three languages display a number of features characteristic of the Bantu family: (i) a complex morphology (both nominal and verbal), (ii) challenging lexical an postlexical phonologies (with processes such as vowel elision and coalescence, which bring additional complexities in the recovery of individual words), and (iii) tones that serve establishing both lexical and grammatical contrasts. Regarding the latter feature, we will be able to build upon the expertise gained in the automatic annotation of the tonal systems of South African languages (Barnard and Zerbian, 2010), although other tonal aspects of our northwestern Bantu languages will require the development of specific approaches.

4.3. Recording of Bantu Languages

From our experience, we have evaluated the quantity of spoken data to be recorded , re-spoken and translated to 100 hours per language, in order to build reliable models for transcription and alignment, and extract some useful information from them. A part of this data will be transcribed, in order to evaluate the automatic transcription and alignment. At the moment of writing about 50 hours of Embosi have been recorded and partly re-spoken using LIG-AIKUMA while Myene (44 hours[4]) and Basaa (40 hours) have been recorded partly with LIG-AIKUMA and mobile devices, partly with traditional methods. The data collected within this project will be provided after the end of the project to the general scientific community via the ELDA agency.[5].

[4]20 hours were recorded before the project

[5]Evaluations and Language resources Distribution Agency Evaluations and Language resources Distribution Agency `http:`

5. Project perspective and methodology

The development of LIG-AIKUMA continues, with the experience gained from the the first successful fieldworks. We are developing the elicitation modes (with images, videos), as well as the different features needed to save the work done during the fieltrips.

BULB's success relies on a strong German-French cooperation between linguists and computer scientists. So far, cooperation has been fostered and strengthened by a series of meetings and courses benefiting the scientific community beyond the present consortium. During the courses, the linguists presented to the computer scientists the major steps to document an unknown language, and the computer scientists introduced their methods to process a "new" language and generate phonetic transcriptions and pseudo-word alignments.

Our three chosen languages, Basaa, Myene and Embosi, have in common a lack of stable orthographic conventions and a lack of texts. Their linguistic resources generally rely on a handful of speakers and none of them is corpus-based. The BULB project will also have the positive outcome of adding to the existing resources (100 hours per language with some transcription and translation) and will thus allow to address new questions with the help of new methodologies (Rialland et al., 2015).

What do endangered languages spoken by few individuals and other unwritten, major languages (e.g., Shanghainese, spoken by 77M people) have in common? They lack written material which drastically limits their access to language processing tools such as speech recognition or translation, not to mention other NLP tools.Our goal is to develop a methodology that can ultimately be applied to any mostly or completely unwritten language, even if it is not endangered.

6. Acknowledgements

This work was realized in the framework of the ANR-DFG project BULB (ANR-14-CE35-002).

Adams, O., Neubig, G., Cohn, T., and Bird, S. (2015). Inducing Bilingual Lexicons from Small Quantities of Sentence-Aligned Phonemic Transcriptions. In *12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam.

Amboulou, C. (1998). *Le Mbochi: langue bantoue du Congo Brazzaville (Zone C, groupe C20)*. Ph.D. thesis, INALCO, Paris.

Ambouroue, O. (2007). *Eléments de description de l'orungu, langue bantu du Gabon (B11b)*. Ph.D. thesis, Université Libre de Bruxelles.

Barnard, E. and Zerbian, S. (2010). From Tone to Pitch in Sepedi. In *Proceedings of the Workshop on Spoken Languages Technologies for Under-Resourced Languages (SLTU10)*.

Besacier, L., Zhou, B., and Gao, Y. (2006). Towards speech translation of non written languages. In Mazin Gilbert et al., editors, *SLT*, pages 222–225. IEEE.

//www.elda.org

Bird, S. and Chiang, D. (2012). Machine translation for language preservation. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Posters, 8-15 December 2012, Mumbai, India*, pages 125–134.

Bird, S. (2010). A scalable method for preserving oral literature from small languages. In *Proceedings of the Role of Digital Libraries in a Time of Global Change, and 12th International Conference on Asia-Pacific Digital Libraries*, ICADL'10, pages 5–14, Berlin, Heidelberg. Springer-Verlag.

Blachon, D., Gauthier, E., Besacier, L., Kouarata, G.-N., Adda-Decker, M., and Rialland, A. (2016). Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app. submitted to SLTU 2016.

Bot ba Njock, H.-M. (1970). *Nexus et nominaux en bàsàa*. Ph.D. thesis, Université Paris 3 Sorbonne Nouvelle.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Crystal, D. (2002). *Language Death*. Cambridge University Press. Cambridge Books Online.

Dimmendaal, G. (1988). *Aspects du basaa*. Peeters/SELAF. [translated by Luc Bouquiaux].

Embanga Aborobongui, G. M. (2013). *Processus segmentaux et tonals en Mbondzi – (variété de la langue embosi C25)*. Ph.D. thesis, Université Paris 3 Sorbonne Nouvelle.

Gerosa, M. and Giuliani, D. (2008). A comparison of read and spontaneous children^{2}s speech recognition. In *The 1st Workshop on Child, Computer and Interaction, WOCCI 2008, Chania, Crete, Greece, October 23, 2008*, page 5.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2006). Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 673–680, Sydney, Australia, July. Association for Computational Linguistics.

Guthrie, M. (1948). *The classification of the Bantu languages*. Oxford University Press for the International African Institute.

Hamlaoui, F. and Makasso, E.-M. (2015). Focus marking and the unavailability of inversion structures in the Bantu language Bàsàá. *Lingua*, 154:35–64.

Hanke, F. R. and Bird, S. (2013). Large-scale text collection for unwritten languages. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pages 1134–1138.

Himmelmann, N. P. and universität Bochum, R. (2002). Documentary and descriptive linguistics. In *In Osamu Sakiyama and Fubito Endo (eds.), Lectures on Endangered Languages 5, 37-83. Kyoto: Endangered Languages of the Pacific Rim*.

Hyman, L. (2003). Basaá (A43). In Derek Nurse et al., editors, *The Bantu languages*, pages 257–282. Routledge.

Janson, T. (2003). *Speak: A Short History of Languages*. Oxford University Press.

Johnson, M., Griffiths, T. L., and Goldwater, S. (2007). Adaptor grammars: a framework for specifying compositional nonparametric bayesian models. In B. Schölkopf, et al., editors, *Advances in Neural Information Processing Systems 19*, pages 641–648, Cambridge, MA. MIT Press.

Johnson, M. (2008). Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27, Columbus, Ohio, June. Association for Computational Linguistics.

Kempton, T. and Moore, R. K. (2014). Discovering the phoneme inventory of an unwritten language: A machine-assisted approach. *Speech Communication*, 56:152–166, January.

Kohler, J. (1996). Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 4, pages 2195–2198 vol.4, Oct.

Kouarata, G. N. (2000). *Dictionnaire Mbochi - Français*. SIL-Congo, Brazzaville.

Lemb, P. and de Gastines, F., (1973). *Dictionnaire Basaá-Français*. Collge Libermann, Douala.

Paul M Lewis, et al., editors. (2013). *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, seventeeth edition.

Makasso, E.-M. (2008). *Intonation et mélismes dans le discours oral spontané en bàsàa*. Ph.D. thesis, Université de Provence (Aix-Marseille 1).

Mochihashi, D., Yamada, T., and Ueda, N. (2009). Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 100–108. Association for Computational Linguistics.

Nettle, D. and Romaine, S. (2000). *Vanishing Voices*. Oxford University Press Inc., New York, NY, USA.

Neubig, G., Mimura, M., Mori, S., and Kawahara, T. (2010). Learning a language model from continuous speech. In *INTERSPEECH*, pages 1053–1056. Citeseer.

Nguyen, T., Vogel, S., and Smith, N. A. (2010). Nonparametric word segmentation for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 815–823, Stroudsburg, PA, USA. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March.

Rialland, A., Embanga Aborobongui, G. M., Adda-Decker, M., and Lamel, L. (2015). Dropping of the class-prefix consonant, vowel elision and automatic phonological mining in Embosi. In *Proceedings of the 44th ACAL meeting*, pages 221–230, Somerville. Cascadilla.

Rosenhuber, S. (1908). Die Basa-Sprache. *MSOS*, 11:219–306.

Stahlberg, F., Schlippe, T., Vogel, S., and Schultz, T. (2012). Word segmentation through cross-lingual word-to-phoneme alignment. In *SLT*, pages 85–90. IEEE.

Stahlberg, F., Schlippe, T., Vogel, S., and Schultz, T. (2013). Pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment. In *The 1st International Conference on Statistical Language and Speech Processing*. SLSP 2013.

Stüker, S., Schultz, T., Metze, F., and Waibel, A. (2003). Multilingual articulatory features. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–144. IEEE.

Stüker, S., Besacier, L., and Waibel, A. (2009). Human Translations Guided Language Discovery for ASR Systems. In *10th International Conference on Speech Science and Speech Technology (InterSpeech 2009)*, pages 1–4, Brighton (UK). Eurasip.

Stüker, S. (2008). Towards human translations guided language discovery for asr systems. In *Proceedings of the First International Workshop on Spoken Languages Technologies for Under-resourced languages (SLTU)*, Hanoi, Vietnam, May.

Van de Velde, M. and Ambouroue, O. (2011). The grammar of Orungu proper names. *Journal of African Languages and Linguistics*, 23:113–141.

Woodbury, A. C. (2003). Defining documentary linguistics. In Peter K. Austin, editor, *Language Documentation and Description*, volume 1, pages 35–51. London.

Woodbury, A. C. (2011). Language documentation. In Peter K. Austin et al., editors, *The Cambridge Handbook of Endangered Languages*, Cambridge Handbooks in Language and Linguistics, pages 159–186. Cambridge University Press, Cambridge.

Xu, J., Gao, J., Toutanova, K., and Ney, H. (2008). Bayesian semi-supervised Chinese word segmentation for statistical machine translation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1017–1024, Manchester, UK, August. Coling 2008 Organizing Committee.