Breaking the Unwritten Language Barrier: The BULB Project

Article *in* Procedia Computer Science · December 2016

DOI: 10.1016/j.procs.2016.04.023

Sebastian Stüker
Karlsruhe Institute of Technology

Martine Adda-Decker
Sorbonne Nouvelle University

Laurent Besacier
Grenoble Alpes University

Helene Bonneau-Maynard
University of Paris-Saclay

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.

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[8]. 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 leads to almost the same performance as the baseline approach, while being applicable to any unwritten language.

[9] 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[10,11] were then learned on the resulting parallel corpus consisting of foreign phone sequences and their corresponding English translation.[12] combined this approach with the monolingual approach above and also did contrastive comparisons.[13] and[14] 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.

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,[15,16] are building on Bayesian monolingual segmentation models introduced by[17] and further expanded in[18]. 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.[19] 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[20,21], 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[22], 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

[23] 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[24], was used. Re-speaking involves listening to an original recording and repeating what was heard carefully and slowly.

In[25], 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[26]. 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 re-speaking 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[27], 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, runs on various mobile phones and tablets and offers a range of speech collection modes (i.e. recording, re-speaking, translation and elicitation). Lig-Aikuma's improved features also include a smart generation and handling of speaker metadata as well as re-speaking and parallel audio data mapping. It was already used for field data collections (see Section 3.3). More details on Lig-Aikuma can be found in the companion paper submitted to this conference[28]. The Lig-Aikuma app has been put on a *forge* and can be downloaded from a direct link `https://forge.imag.fr/frs/download.php/706/MainActivity.apk`.

3. Documentation of three Bantu Languages

3.1. Bantu languages

In BULB, three typologically diverse northwestern Bantu languages were selected, which stem from different Guthrie zones (areal-genetic groupings,[29]): Basaa (A43, Cameroon), Myene (B10, Gabon) and Embosi (C25, Congo-Brazzaville). The Bantu family is one of the largest 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 chosen Bantu languages however have in common that they are relatively well described, as there are also 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.

3.2. 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[30] and the first Basaa-French dictionary was developed over half a century ago[31]. Several dissertations have focused on various aspects of Basaa[32,33] and the language also benefits from recent and ongoing linguistic studies[34,35,36].

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[37]. 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[38] is available, as well as a few articles on aspects of the phonology, morphology and syntax of Myene ([39] 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[40] is available and, just like Basaa and Myene, the language benefits from recent linguistic studies[41,42].

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 and 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[43], although other tonal aspects of our northwestern Bantu languages will require the development of specific approaches.

3.3. 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 is 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 of which 20 hours were recorded before the project) 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.[2].

4. Project perspective and methodology

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[44].

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.

Acknowledgements

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

References

1. Nettle, D., Romaine, S.. *Vanishing Voices*. New York, NY, USA: Oxford University Press Inc.; 2000. ISBN 0195136241.
2. Crystal, D.. *Language Death*. Cambridge University Press; 2002. ISBN 9781139871549. URL: http://dx.doi.org/10.1017/CBO9781139871549; cambridge Books Online.

[2] Evaluations and Language resources Distribution Agency http://www.elda.org

3. Janson, T.. *Speak: A Short History of Languages*. Oxford University Press; 2003. ISBN 9780199263417. URL: `https://books.google.fr/books?id=tSBaF\_oFUDYC`.
4. Himmelmann, N.P., universität Bochum, R.. 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*. 2002.
5. Woodbury, A.C.. Language documentation. In: Austin, P.K., Sallabank, J., editors. *The Cambridge Handbook of Endangered Languages*; Cambridge Handbooks in Language and Linguistics. Cambridge: Cambridge University Press; 2011, p. 159–186.
6. Kohler, J.. Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. In: *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*; vol. 4. 1996, p. 2195–2198 vol.4. doi:`10.1109/ICSLP.1996.607240`.
7. Stüker, S., Schultz, T., Metze, F., Waibel, A.. Multilingual articulatory features. In: *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*; vol. 1. IEEE; 2003, p. I–144.
8. Besacier, L., Zhou, B., Gao, Y.. Towards speech translation of non written languages. In: Gilbert, M., Ney, H., editors. *SLT*. IEEE. ISBN 1-4244-0873-3; 2006, p. 222–225. URL: `http://dblp.uni-trier.de/db/conf/slt/slt2006.html#BesacierZG06`.
9. Stüker, S.. 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; 2008.
10. Brown, P.F., Pietra, S.A.D., Pietra, V.J.D., Mercer, R.L.. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 1993;**19**(2):263–311.
11. Och, F.J., Ney, H.. A systematic comparison of various statistical alignment models. *Comput Linguist* 2003;**29**(1):19–51. URL: `http://dx.doi.org/10.1162/089120103321337421`. doi:`10.1162/089120103321337421`.
12. Stüker, S., Besacier, L., Waibel, A.. Human Translations Guided Language Discovery for ASR Systems. In: *10th International Conference on Speech Science and Speech Technology (InterSpeech 2009)*. Brighton (UK): Eurasip; 2009, p. 1–4. URL: `https://hal.archives-ouvertes.fr/hal-00959225`.
13. Stahlberg, F., Schlippe, T., Vogel, S., Schultz, T.. Word segmentation through cross-lingual word-to-phoneme alignment. In: *SLT*. IEEE. ISBN 978-1-4673-5125-6; 2012, p. 85–90. URL: `http://dblp.uni-trier.de/db/conf/slt/slt2012.html#StahlbergSVS12`.
14. Stahlberg, F., Schlippe, T., Vogel, S., Schultz, T.. Pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment. In: *The 1st International Conference on Statistical Language and Speech Processing*. 2013, URL: `http://csl.uni-bremen.de/cms/images/documents/publications/SLSP2013-StahlbergSchlippe_PronunciationExtraction.pdf`; sLSP 2013.
15. Xu, J., Gao, J., Toutanova, K., Ney, H.. Bayesian semi-supervised Chinese word segmentation for statistical machine translation. In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, UK: Coling 2008 Organizing Committee; 2008, p. 1017–1024. URL: `http://www.aclweb.org/anthology/C08-1128`.
16. Nguyen, T., Vogel, S., Smith, N.A.. Nonparametric word segmentation for machine translation. In: *Proceedings of the 23rd International Conference on Computational Linguistics*; COLING '10. Stroudsburg, PA, USA: Association for Computational Linguistics; 2010, p. 815–823. URL: `http://dl.acm.org/citation.cfm?id=1873781.1873873`.
17. Goldwater, S., Griffiths, T.L., Johnson, M.. 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*. Sydney, Australia: Association for Computational Linguistics; 2006, p. 673–680. URL: `http://www.aclweb.org/anthology/P06-1085`. doi:`10.3115/1220175.1220260`.
18. Mochihashi, D., Yamada, T., Ueda, N.. 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*. Association for Computational Linguistics; 2009, p. 100–108. URL: `http://dl.acm.org/citation.cfm?id=1687894`.
19. Adams, O., Neubig, G., Cohn, T., Bird, S.. Inducing Bilingual Lexicons from Small Quantities of Sentence-Aligned Phonemic Transcriptions. In: *12th International Workshop on Spoken Language Translation (IWSLT)*. Da Nang, Vietnam; 2015.
20. Johnson, M., Griffiths, T.L., Goldwater, S.. Adaptor grammars: a framework for specifying compositional nonparametric bayesian models. In: Schölkopf, B., Platt, J., Hoffman, T., editors. *Advances in Neural Information Processing Systems 19*. Cambridge, MA: MIT Press; 2007, p. 641–648.
21. Johnson, M.. Unsupervised word segmentation for Sesotho using adaptor grammars. In: *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*. Columbus, Ohio: Association for Computational Linguistics; 2008, p. 20–27. URL: `http://www.aclweb.org/anthology/W/W08/W08-0704`.
22. Neubig, G., Mimura, M., Mori, S., Kawahara, T.. Learning a language model from continuous speech. In: *INTERSPEECH*. Citeseer; 2010, p. 1053–1056. URL: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.174.7798\&rep=rep1\&type=pdf`.
23. Bird, S.. 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. Berlin, Heidelberg: Springer-Verlag. ISBN 3-642-13653-2, 978-3-642-13653-5; 2010, p. 5–14. URL: `http://dl.acm.org/citation.cfm?id=1875689.1875692`.
24. Woodbury, A.C.. Defining documentary linguistics. In: Austin, P.K., editor. *Language Documentation and Description*; vol. 1. London; 2003, p. 35–51.
25. Bird, S., Chiang, D.. 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*. 2012, p. 125–134. URL: `http://aclweb.org/anthology/C/C12/C12-2013.pdf`.
26. Hanke, F.R., Bird, S.. Large-scale text collection for unwritten languages. In: *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*. 2013, p. 1134–1138. URL: `http://aclweb.org/anthology/I/I13/I13-1161.pdf`.
27. Kempton, T., Moore, R.K.. Discovering the phoneme inventory of an unwritten language: A machine-assisted approach. *Speech Communication* 2014;**56**:152–166. URL: `http://dx.doi.org/10.1016/j.specom.2013.02.006`. doi:`10.1016/j.specom.2013.02.006`.

28. Blachon, D., Gauthier, E., Besacier, L., Kouarata, G.N., Adda-Decker, M., Rialland, A.. Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app; 2016. Submitted to SLTU 2016.
29. Guthrie, M.. *The classification of the Bantu languages*. Oxford University Press for the International African Institute; 1948.
30. Rosenhuber, S.. Die Basa-Sprache. *MSOS* 1908;**11**:219–306.
31. Lemb, P., de Gastines, F.. *Dictionnaire Basaá-Français*. Collge Libermann; Douala; 1973.
32. Bot ba Njock, H.M.. *Nexus et nominaux en bàsàa*. Ph.D. thesis; Université Paris 3 Sorbonne Nouvelle; 1970.
33. Makasso, E.M.. *Intonation et mélismes dans le discours oral spontané en bàsàa*. Ph.D. thesis; Université de Provence (Aix-Marseille 1); 2008.
34. Dimmendaal, G.. *Aspects du basaa*. Peeters/SELAF; 1988. [translated by Luc Bouquiaux].
35. Hyman, L.. Basaá (A43). In: Nurse, D., Philippson, G., editors. *The Bantu languages*. Routledge; 2003, p. 257–282.
36. Hamlaoui, F., Makasso, E.M.. Focus marking and the unavailability of inversion structures in the Bantu language Bàsàá. *Lingua* 2015; **154**:35–64.
37. Lewis, P.M., Simons, G.F., Fennig, C.D., editors. *Ethnologue: Languages of the World*. Dallas, Texas: SIL International; seventeeth ed.; 2013.
38. Ambouroue, O.. *Eléments de description de l'orungu, langue bantu du Gabon (B11b)*. Ph.D. thesis; Université Libre de Bruxelles; 2007.
39. Van de Velde, M., Ambouroue, O.. The grammar of Orungu proper names. *Journal of African Languages and Linguistics* 2011;**23**:113–141.
40. Kouarata, G.N.. *Dictionnaire Mbochi - Français*. Brazzaville: SIL-Congo; 2000.
41. Amboulou, C.. *Le Mbochi: langue bantoue du Congo Brazzaville (Zone C, groupe C20)*. Ph.D. thesis; INALCO; Paris; 1998.
42. Embanga Aborobongui, G.M.. *Processus segmentaux et tonals en Mbondzi – (variété de la langue embosi C25)*. Ph.D. thesis; Université Paris 3 Sorbonne Nouvelle; 2013.
43. Barnard, E., Zerbian, S.. From Tone to Pitch in Sepedi. In: *Proceedings of the Workshop on Spoken Languages Technologies for Under-Resourced Languages (SLTU10)*. 2010.
44. Rialland, A., Embanga Aborobongui, G.M., Adda-Decker, M., Lamel, L.. Dropping of the class-prefix consonant, vowel elision and automatic phonological mining in Embosi. In: *Proceedings of the 44th ACAL meeting*. Somerville: Cascadilla; 2015, p. 221–230.