

Enabling Linguistics Diversity and Multilingualism Worldwide

NTeALan - AI/NLP/NLU Platforms For Sharing and Leveraging African Language Resources For Education In Africa

Elvis MBONING¹, Jean-Marc BASSAHAK¹, Juanita Fopa¹, Jules ASSOUMOU², Damien NOUVEL³

NTeALan

Makepe, parcours vita - Douala, Cameroon
 {levismboning, bassahak, juanita.fopa}@ntealan.org¹
 julesassoumou@yahoo.fr², damien.nouvel@inalco.fr³

Abstract

Among the challenges, the African continent faces the issue of safeguarding and enhancing its cultural and linguistic inheritance. Created in 2017, the NTeALan project aims at setting up intelligent tools for the digitization, development and teaching of African languages. Since 2019, our association is supported by 20 volunteers from various fields, several partnerships were signed (universities, institutions, startups, association) and we have implemented several prototypical applications: chatbots, collaborative dictionary, linguistic map, REST APIs for African language resources. With its multimedia center dedicated to NLP, NTeALan aims at making African languages a cornerstone of Africa's cultural, linguistics and technological development.

Keywords: African languages, AI, NLP, NLU, Lexicography, Education, Resources, API, Culture

Résumé (Bassa'a language spoken in Cameroon)

i kété mítíik mí mám má má nlámá hólós áfríkà í màṅgégé máná dī gwě, màhòl má má ntágbéné í máhóp més nì bífòṅól gwés, íṅṅù hàlà péné ntealan tòhàlá kíí à má sál ṅgàndàk mú í ndzél í, à ṅgí sálák ní láná léé: à ntí bínóṅól bí nòndó bí bí níhóla lèè dí nííga, nì hólós màhóp més lóṅní ndzì bífòṅól bí mòndó. náánó í ṅwíí dikóó díbàà nì mbògí dzóm nì bòó, í sí màtíṅmá «nílòṅú ntealan», í léí móó màà (mòò màà) má fót má lólàk í mítèn mí bífòlò ṅwómísó, bá úsál ntóṅwádá nì fès íṅṅù màhòl má lítúṅgá lí áfríkà. àndàk màtíṅmá yé ñsàngè, màpamá kèfá nsàngè lóṅnì mìnén mí bífòlò ṅgàndàk: bísúkùlù nì míníṅbíkújí. ṅgàndàk í màṃ í èfòṅà, í màṃ má gáhóla lítúṅgá lí áfríkà dzó líso, dínlà símá: api íṅṅù tèèdà mítèn mí bífòṅól gwés, bíkàat bí bífúk, bíkàat bí máhóp...). ndàp í fóló í nlp/nlu í yè í hálá (càmàlùn). ntealan à ñsómólò fòṅlèè màhóp má áfríkà má bá íṅgém ú màhòl má bílòṅ gwés gwó bísómá.

1. Overview on NTeALan project

Among the challenges, the African continent faces the issue of safeguarding and enhancing its cultural and linguistic inheritance. Created in 2017¹, and managed by academics and the African Learned Society, NTeALan (New Technologies for African Languages) is an Association that works for the implementation of intelligent technological tools for the promotion, development and teaching of African national languages. Our goals are to digitize, safeguard and promote African national languages through digital tools and Artificial Intelligence to build a new generation of young Africans aware of the challenges of appropriating the languages and cultures of the continent. In this paper, we want to present our first major activities realised between 2017 and 2019 with the NTeALan's teams. We will continue with the difficulties encountered and the future challenges for the upcoming years.

1.1. Why is this project a necessity for Africa ?

Language plays an important role in defining the identity and humanity of individuals. As Tunde Opeibi (Tunde,

2012, p.272) said "In Africa, evidence shows that language has become a very strong factor for ethno-national identity, with the ethnic loyalty overriding the national interest". To date, the African continent has more than 3000 languages, more than two thirds of which are poorly endowed. Among the reasons justifying this observation, we can list:

- The lack of a strong linguistic policy in favor of these languages,
- The lack of specialists in NLP / AI / NLU trained on the continent and specialists in these languages,
- The lack of linguistic resources (textual and oral mainly oral tradition) and NLP/NLU tools available for most of these languages,
- A virtual absence of these languages in the digital space (social networks, online platform, etc.) and in the educational system.
- No African scientific community dedicated to technological issues related to the written and oral transmission of knowledge in African languages
- Few standardized African languages on a vast ensemble and their gradual disappearance over the years

¹Mainly by Elvis MBONING (NLP Research Engineer at IN-ALCO) and Jean Marc BASSAHAK (Contractor, Web designer and developer). Jules Assoumou, Head of the Department of Linguistics and African Literature at University of douala, joined us later.

Faced with this, specialists from various fields (computer developer, NLP engineer, academics specialized in questions of linguistics, cultures, didactic and African pedagogy) gathered around the world to create an association which was to make it possible to set up NLP and NLU tools based on the current state of the art of AI in order to create collaborative environments for the creation of sharing resources and tools around African languages intended to serve the scientific community, companies, social networks and all other public or private institutions.

We are convinced, as Tunde Opeibi (Tunde, 2012, p.289) already said so well that "the linguistic diversity in Africa can still become the catalyst that will promote cultural, socio-economic, political, and technological development, as well as sustainable growth and good governance in Africa."

1.2. Our strategies

Our approach is exclusively based on the collaboration model (Holtzblatt and Beyer, 2017). We want to allow African people to contribute to the development of their own mother tongue, supervised by specialists and academics of African languages. Our model involves setting up several communities: the community of speakers of these languages, the community of native specialists (guarantors of traditional, cultural and linguistic knowledge), the community of academics specialized in African linguistics technologies and the community of social / institutional / public partners. The chart 1 below summarizes this strategy.

Collaboration	Research and Tools
Create a community of professionals specialized in sociolinguistics and technological issues (NLP / NLU / IA) for Africa	Create open source platform for academics specialists in other to give them more means for their research activities
Create a community of volunteer contributors around collaborative platforms for building common language resources (text, image, video)	Systematically equip all the African national languages by language family and encourage their use by young people in Africa
Collaboratively use created resources to set up an autonomous language teaching platform	Help public/private institutions and companies by integrating our technologies in the education system, in their own platforms, and others.

Table 1: Strategy adopted by NTeALan

1.3. The main NTeALan's projects

During our first years², we initiated some projects mainly centered on building collaborative, multilingual and dis-

tributed resources for African languages and cultures. Our team has just put in place:

- The multilingual conversational agents platform [NTeABot] to teach young African students their mother tongues
- The collaborative dictionaries (lexicon, audio, picture and video) platform for African national languages and cultures [<https://ntealan.net/dictionnaires>],
- REST and Websocket APIs for sharing African language resources [<https://apis.ntealan.net>],
- The platform for the management of lexical and terminological resources in African languages [<https://ntealan.net/dictionaries-platform>],
- The dictionary platform illustration [<http://up-files.ntealan.org/koken>],
- The tool for digitizing documents in African languages
- The dictionaries annotation platform (component of the numerisation tool) [<http://dico-edit.ntealan.net>],
- Scientific research activities in Natural Language Processing (NLP) and Natural Language Understanding (NLU) for African languages,
- The management of the NTeALan center at Makepe (Douala) and many others internally.

Essentially based on REST and Websocket APIs technologies, these first initiated projects are still in the testing phase in a few languages. Indeed, we started from a few pilot languages (essentially Bantu and semi-Bantu languages)³ already having available resources in low quantity. Our objective was to apply the first versions of our NLP/NLU tools (morphological, syntactical and semantic analysis, NER, automatic conjugation and POS tagging) on this sample in order to analyze the results and see if these could be generalized on others in the same linguistic family. The figure 1 below show the general structure of our actual system.

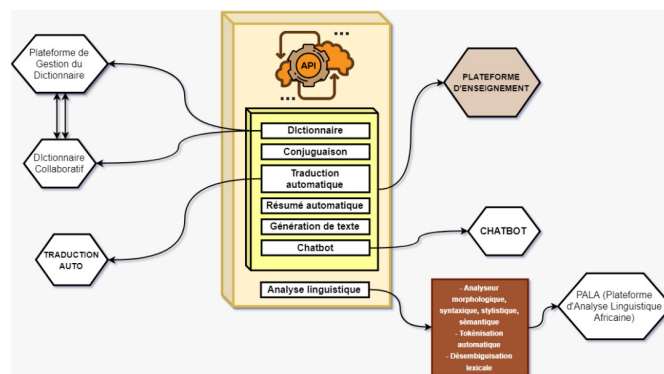


Figure 1: NTeALan APIs and services infrastructures

²We started at the end of 2017 till now.

³We have yemba, bassa'a, duala, ngiembon (spoken in Cameroon) and bambara (spoken in Mali)

2. Description of our current major projects

Our main projects, on which others depend, are the collaborative dictionaries for African languages resources and tools (cf. figure 2), the African linguistic map and the conversational agent platform [NTeABot] for teaching African languages.

2.1. Open source collaborative dictionaries, NLP/NLU tools and their REST API

For this first main project⁴, we give access to native speakers and experts who have an expertise in African language to build collaboratively resources like lexicon⁵, illustration of cultural phenomenon, sound and videos (recording process) based on semantic information on article in their native language. These shared resources are freely available for all contributors through our REST API hosted at [https://apis.nteanan.net/nteanan/dictionaries].

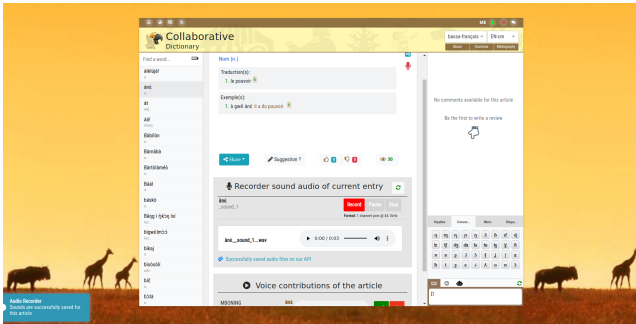


Figure 2: Collaborative dictionaries for sharing multi-modal and multilingual resources in African languages. This platform is under license on Creative Commons BY-SA: [https://nteanan.net]

AI being essential today for the construction of good quality linguistic resources and tools, we lead internally with our academic partners (language and African literature department of the University of Douala, the ERTIM team of the INALCO) numerous research activities in Artificial Intelligence, NLP, and NLU in order to contribute to the industrialization of these languages. The results of these studies help our applications and can be use by others researchers: this include data (in different common formats like TEI (Benoit and Turcan, 2006), LMF, XMLAF⁶) and tools.

2.2. Multilingual conversational agent platform and its REST API

For the second main project (cf. figure 3), we want NTeABot platform to teach young African students in Africa

⁴This project was born following the research work of Elvis MBONING at the University of Douala and University of Lille 3 (Master thesis): (MBONING, 2016) and (MBONING, 2017). We can cite other related works in this fields like (Assoumou, 2010), (Mangeot and Enguehard, 2011), (Vydrin et al., 2016), (Maslinsky, 2014), (Nouvel et al., 2016), etc.

⁵For this work, we also build another platform to manage lexicographic resource: [https://nteanan.net/dictionaries-platform].

⁶NTeALan codification format of dictionary for African bantu and semi-bantu languages

or at the diaspora in their different languages in the same time during course teaching with teachers at school and with their parents at home. With NTeABot platform we can also build other competences (applications) such as information on time, definition on wikipedia, informations on NTeALan dictionaries, on NTeALan project and some others. The test version is available on our official website [https://nteanan.org]



Figure 3: Sample of discussion about NTeALan dictionaries with NTeABot agent. Tested on [https://nteanan.org]

2.3. African linguistic and cultural map

We realised a detailed linguistic and cultural map on each country in Africa (cf. figure 4). For each of these countries, we provided a number of spoken languages, dialects, their classification, development status and the place where they are used. We want here to enumerate all the living languages in Africa in other to link them with their resources and NLP/NLU tools. The actual hosted version only contains the Cameroonian languages. We will add for others African countries in our next deliveries next year depending on the availability of language resources.

3. Problems encountered and futures challenges

The implementation of these first projects enabled us to note certain important problems. In upcoming years, it will be a question of filling them up and making them more mature for future deadlines. First of all, let's start with the problems encountered.

3.1. Problems encountered

We are currently facing two main problems in the NTeALan association:

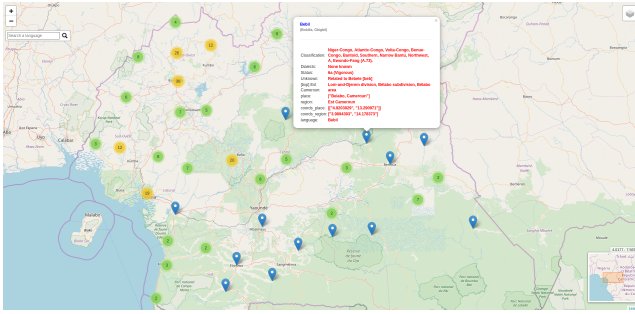


Figure 4: Linguistic description map of African languages. The figure show precisely Cameroon area with focus on bebil language spoken in the city of Belabo at the East region of Cameroon

- the first is the number of our researchers and insufficient IT resources. For the staff, we do not yet have all the specialists (NLP, NLU, African languages) that we wish to treat. Regarding IT resources, we do not have enough robust IT infrastructures (servers, field tools) for our research work on these languages.
- the second is the lack of funding to fill our research and development activities for the NLP and NLU tools concerned. Our funding mainly come from the members of the association, which is insufficient in light of our current ambitions.

3.2. Next challenges to be met

Our ambitions are great and will require more staff and financial resources.

- Above all, we want to make the greatest number of specialists in African languages and cultures in each African country and in the world, to join our association to further challenges.
- Find funding from private and public institutions, businessmen, companies who can support our research work and the continuous development of our applications for the teaching of these poorly endowed languages.
- Improve and enrich all existing platforms and open them up more to the scientific community and speakers of these languages. We will focus on : the autonomous platform for teaching languages and cultures, the conversational Agent Assistant for Language Teaching and Virtual cultural museum for the safeguarding of the African socio-cultural inheritance, etc.
- Strengthen our partnerships with African social and cultural institutions, universities, research laboratories and companies specializing in our research areas. The aim is to create a community of experts in linguistics, technological and cultural issues across the continent

4. Conclusion

All in all, NTeALan is: 2 years of existence (2017-2019), 3 institutional and private partners, more than 5 students

trained in computational linguistics, 1 equipped research and development center based in Douala (Cameroon), more than 15 members worldwide, 3 tooled african languages, more than 10 internal projects and 5 open source and collaborative resources and tools. We are motivated to continue to grow up and to build an active environment around data, researchers, african people and technological companies for these languages. You are free to support or join us: contact@ntealan.org | dons@ntealan.org

5. Acknowledgements

This project is actually supported by the ministry of post and telecommunication of Cameroon, Department of linguistics and African literature of the University of Douala (Cameroon), Research teams ERTIM of INALCO (France) and Fractals system (France). We can also cite: Daniel BALEBA, Juanita FOPA, Merci Christian BONOGBILAP, Marcel TOMI BANOU, NTOMB David, NTOMB Nicolas, Théophile KENGNE, Parfait DIMONO, Yves Bertrand DISSAKE, Daniel NOUVEL and many other contributors.

6. References

- Assoumou, J. (2010). *Enseignement oral des langues et cultures africaines à l'école primaire*. Éditions Clé, Yaoundé, Cameroun, 1st edition.
- Benoit, J.-L. and Turcan, I. (2006). La TEI au service de la transmission documentaire ou de la valorisation des richesses patrimoniales : le cas difficile des dictionnaires anciens.
- Holtzblatt, K. and Beyer, H. (2017). 7 - Building Experience Models. In Karen Holtzblatt et al., editors, *Contextual Design (Second Edition)*, Interactive Technologies, pages 147–206. Morgan Kaufmann, Boston, January.
- Mangeot, M. and Enguehard, C. (2011). Informatisation de dictionnaires langues africaines-français. In *Journées LTT 2011*, page 11.
- Maslinsky, K. (2014). *Daba: a model and tools for Manding corpora*.
- MBONING, E. (2016). De l'analyse du dictionnaire yémba-français à la conception de sa DTD et de sa réédition sur support numérique. Mémoire Master 1, Université de Lille 3.
- MBONING, E. (2017). Vers une métalexigraphie outillée : conception d'un outil pour le métalexigraphe et application aux dictionnaires Larousse de 1856 à 1966. Mémoire Master 2, Université de Lille 3.
- Nouvel, D., Donandt, K., Auffret, D., Maslinsky, K., Chiarcos, C., and Vydrin, V. (2016). Resources and Experiments for a Bambara POS Tagger. *Intra Speech*, page 14.
- Tunde, O. (2012). Investigating the Language Situation in Africa. In *Language and Law*, Language rights, pages 272–293. Oxford Handbooks in Linguistics, Great Clarendon street.
- Vydrin, V., Rovenchak, A., and Maslinsky, K. (2016). Maninka Reference Corpus: A Presentation. In *TALAF 2016 : Traitement automatique des langues africaines (écrit et parole)*. Atelier JEP-TALN-RECITAL 2016 - Paris le, Paris, France, July.