Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020), pages 66–72
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

NTeALan Dictionaries Platforms: An Example Of Collaboration-Based Model

Elvis Mboning[1, 2], **Daniel Baleba**[1], **Jean Marc Bassahak**[1],**Ornella Wandji**[1], **Jules Assoumou**[1, 3]
NTeALan[1], ERTIM (INALCO)[2]
Tradex Makepe - Douala (Cameroon)[1] , 2 rue de Lille - Paris (France)[2]
elvis.mboning@inalco.fr[2], julesassoumou@yahoo.fr[1, 3]
{levismboning, daniel.baleba, bassahak, ornella.wandji}@ntealan.org[1]

Abstract

Nowadays the scarcity and dispersion of open-source NLP resources and tools in and for African languages make it difficult for researchers to truly fit these languages into current algorithms of artificial intelligence, resulting in the stagnation of these numerous languages, as far as technological progress is concerned. Created in 2017, with the aim of building communities of voluntary contributors around African native and/or national languages, cultures, NLP technologies and artificial intelligence, the NTeALan association has set up a series of web collaborative platforms intended to allow the aforementioned communities to create and manage their own lexicographic and linguistic resources. This paper aims at presenting the first versions of three lexicographic platforms that we developed in and for African languages: the REST/GraphQL API for saving lexicographic resources, the dictionary management platform and the collaborative dictionary platform. We also describe the data representation format used for these resources. After experimenting with a few dictionaries and looking at users feedback, we are convinced that only collaboration-based approaches and platforms can effectively respond to challenges of producing quality resources in and for African native and/or national languages.

Keywords: African languages, NLP platforms, resources, XML serialisation, collaboration-based model, dictionaries, lexicography, open-source

1. Introduction

For several years now, artificial intelligence technologies, including those of NLP, have greatly contributed to the economic and scientific emergence of poorly endowed languages in northern countries, thanks to the availability of lexicographic and terminography resources in sufficient quantity. African languages benefit very little from these intelligent tools because of the scarcity of available structured data and collaborative platforms for building linguistic and cultural knowledge bases. In order to meet this need and complement the initiatives already present on the continent ((De Pauw et al., 2009), (Mboning, 2016), (Vydrin, Valentin and Rovenchak, Andrij and Maslinsky, Kirill, 2016), (Abate et al., 2018), (Mboning, Elvis and NTeALan contributors, 2017), (Mangeot and Enguehard, 2011), (De Schryver, 2010), Afrilex association (Ruthven, 2005)), and also those from other African, European and American research centers, NTeALan (New Technologies for African Languages), specialized in the development of NLP / NLU tools for teaching African languages and cultures, has set up a collaborative and open-source platform for building lexical resources for African native and/or national languages.

Our paper focuses on the description of NTeALan's architectures platform and its lexicographic data format (African linguistics and cultural resources), component of our collaborative language resource platform, which is an important starting point for the technological step forward of each African language. This platform is divided into three components: the open-source dictionary REST API (back-end), the dictionary management platform and the collaborative dictionary platform (fronts-end).

2. Context of the work

2.1. NTeALan project

Created in 2017[1] and managed by academics and the African Learned Society, NTeALan is an Association that works for the implementation of intelligent technological tools necessary for the development, promotion and teaching of African native and/or national languages. Our goals are to digitize, safeguard and promote these poorly endowed languages through digital tools and Artificial Intelligence. By doing so, we would like to encourage and help young Africans, who are willing to learn and/or teach their mother tongues, and build a new generation of Africans aware of the importance and challenges of appropriating the languages and cultures of the continent. Another purpose of NTeALan's work is to provide local researchers and companies with data which could help them improve the quality of their services and work, hence building open-source African languages resources is one of our core projects.

2.2. NTeALan approach: a collaboration-based model

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 tongues, under the supervision of specialists and academics of African languages. Our model involves setting up several communities: a community of speakers of these languages, a community of native specialists

[1]Namely by Elvis Mboning (NLP Research Engineer at INALCO) and Jean Marc Bassahak (Contractor, Web designer and developer), who were later on joined by Jules Assoumou, Head of Department of Linguistics and African Literature at the University of Douala.

(guarantors of traditional, cultural and linguistic knowledge), a community of academics specialized in African linguistics technologies and a community of social and/or institutional and/or public partners. Grouped by languages, these communities work together with the same purpose: building as much linguistic and cultural resources as possible, required for research, education and technology needs.

The concept of community is not a trivial choice in our case. Indeed, African sociology is built on the community model, that is, a set of social groups and sub-groups sharing the same language, the same culture and the same geographic space. In such groups, solidarity is created and social actions emerge for the interest of all: this is the case with villages cultural associations and representations in urban cities, collaborative meetings and cultural events. This concept clearly shows the strong cultural link that unites each citizen with his community, even before that of his country. This is precisely the reason why we have chosen this approach and we apply it to all NTeALan internal projects, especially to the development of language resources, their platforms, as well as their data representation.

3. NTeALan language resources platforms

Our language resources platforms are divided into three parts: one independent architecture and two dependent architectures. The independent architecture serves not only the two other architectures but all NTeALan projects, as illustrated in figure 1.

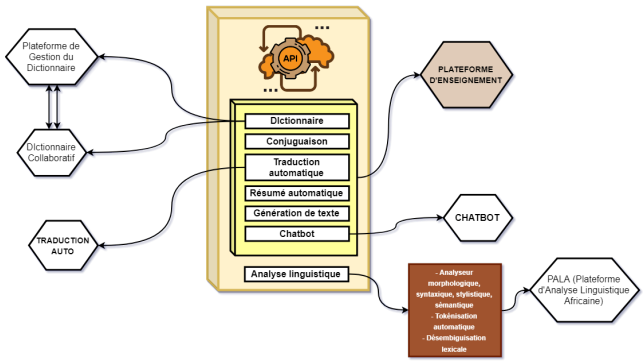

Figure 1: NTeALan REST/GraphQL APIs and services infrastructures

The three architectures are the fruit of two upstream processes depending on the input type (PDF files or images). The first process involves digitization and the second serialization:

- **digitization**: dictionaries in paper or digital format like PDF, TIFF, PNG by OCR (Optical Character Recognition) are digitized with Deep learning (Breuel, 2008); we annotate them to improve the OCR (see figure 2); each article constituents (featured word, translation, contextualization, conjugation, dialect variant, etc.) are automatically detected, extracted and xmlized in XND (NTeALan dictionary XML format) afterwards.
- **serialization**: dictionaries in an external format (toolbox, XML, TEI, LMF) are automatically serialized in XND format, using our internal NLP tools[2]. Reversed processes can also be done from XND to XML TEI, LMF, SIL Toobox.

In both cases, we start with a paper or digital dictionary and end up with a XML dictionary in XND format (see figure 6). The latter is the unique data entry format for our three architectures. It should be noted that the two processes described above are controlled by NTeALan linguists only. In future work, they will be opened to non-member contributors.

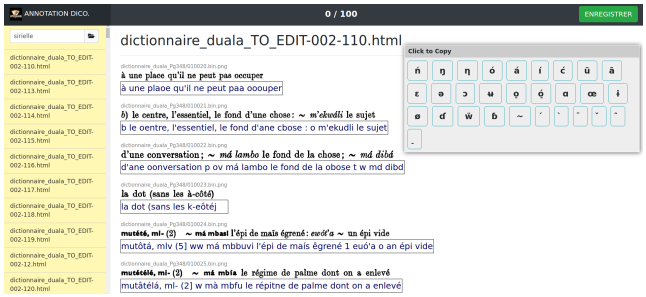

Figure 2: NTeALan dictionaries annotation platform based on the Ocropy tool and used to train Deep learning model for OCR. This platform is under license on Creative Commons BY-NC-SA 3.0 license: (`http://dico-edit.ntealan.net`)

Figure 2 shows an example of annotation (from the bilingual Duala-French dictionary) performed by NTeALan's members.

3.1. Independent architecture

The independent platform can also be called lexicographical resources management database. This architecture has two consultation interfaces : the web-based REST and the GraphQL APIs platform[3]. Built to be simple and accessible, this web application stores and distributes all lexicographic resources resulting from the collaborative work of NTeALan's communities members and external contributors.

The independent architecture uses our internal NLP tools to manage the XND file format in order to give users easy access to their contributions (see section 4.). The operations listed in table 3.1.1. are allowed in open access for each type of user.

3.1.1. Web-based REST API interface

This interface structures lexicographic resources into REST resources ranging from general to specific. It proceeds

[2]These include tokenizers, lemmatizers, text parsers and lexical disambiguation tools used for processing noisy lexicographic corpora.

[3]This architecture is close to the Kosh APIs for dictionaries `https://cceh.github.io/kosh/`, as well as the ELEXIS Dictionary Service `https://github.com/elexis-eu/dictionary-service`

from a dictionary in an African language to access its lexicographic components: *dictionaries* > *articles* > *entry* > *dialect_variant* or *dictionaries* > *articles* > *entry* > *translation* > *language* or *dictionaries* > *articles* > *entry* > *conjugation* (for more precision, see table 3.1.1.).

Actions	URL path (root is */dictionaries*)
get metadata of dictionary	/metadata/{dictionary_id}
get article	/articles/{dictionary_id}/{article_id}
get entry of article	/articles/{dictionary_id}/{article_id}?entry
get translation of article	/articles/{dictionary_id}/{article_id}?trans=en
get comments of article	/comments/{dictionary_id}/{article_id}

Table 1: Sample of a REST API structure for our lexicographic resources

The documentation[4] for this interface is accessible under the Creative Commons BY-NC-SA 3.0 license. The access privileges, for each type of user, is described in table 3.1.1.

Operations	NTeALan's users	Native communities	Scientific experts
manage dictionary	yes	no	yes
manage article	yes	yes	yes
data validation	no	yes	yes
cultural media	yes	yes	no
comments	yes	yes	yes

Table 2: User's privileges for each operation in NTeALan's REST API

3.1.2. Web-based GraphQL API interface

The resources available in our lexicographic database can also be consulted using a GraphQL query language associated to the system (through a GraphQL API interface)[5]. This API is also required for all data parallel to our lexicographic resources, namely the comments from dictionary users, the dictionary metadata and articles.

The GraphQL API interface uses the request system of the Python Graphene library, which render the exploration process of our resources data easier. An external GraphQL clients can also easily be linked to the GraphQL server to extract the information sought. Unlike the REST API, this interface cannot add, modify or delete data.

3.2. Dependent architectures

Dependent architectures are single web page platforms[6] which use the data stored in the common REST API database (Independent platform) and enriched by contributors. They can also perform the operations described in table 3.1.1. through their web interface.

3.2.1. Dictionaries management platform

As a web platform, the dictionaries management platform is a graphical management version of the REST API platform. It allows NTeALan members (users) to manage dictionaries, articles, users, users comments, access requests and cultural resources.

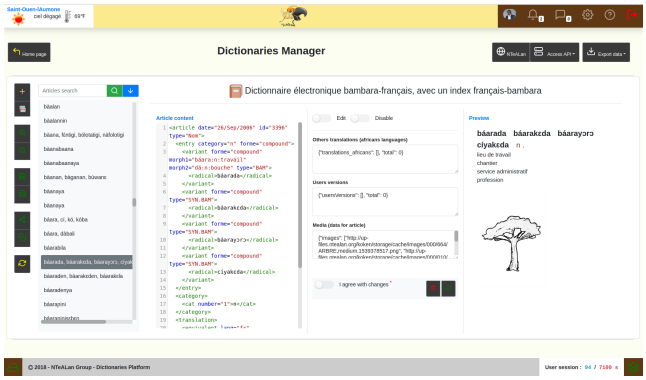

Figure 3: Dictionaries management platform for multi-modal and multilingual lexicographical resources for African languages. This platform is under NTeALan's license: (`https://ntealan.net/dictionaries-platform`)

Unlike the other above-mentioned platforms, this is not an open-source platform. It can be used strictly by NTeALan communities, as part of a direct collaboration between the linguistics team and other association members.

3.2.2. Collaborative dictionary platform

Collaborative dictionary platform[7] is also a web platform (see figure 4) which enhances the lexicographical resources from the REST API. It connects and gives native speakers and African languages experts (NTeALan communities as described in section 2.2.) the opportunity to build, in a collaborative approach, resources like lexicons[8], illustrations

[4] `https://apis.ntealan.net/ntealan/dictionaries`

[5] `https://apis.ntealan.net/ntealan/graphql`

[6] For these platforms, we have recourse to the latest front-end technologies (React Js and Angular Js), in priority the single web applications (SPA) for their simplicity, speed and robustness.

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

[8] To this aim, we built another platform to manage lexicographic resources: [https://ntealan.net/dictionaries-platform].

of cultural phenomena, sounds and videos (recording process) based on semantic information extracted from articles written in their native languages. These shared resources are stored and freely available for all contributors through our APIs.

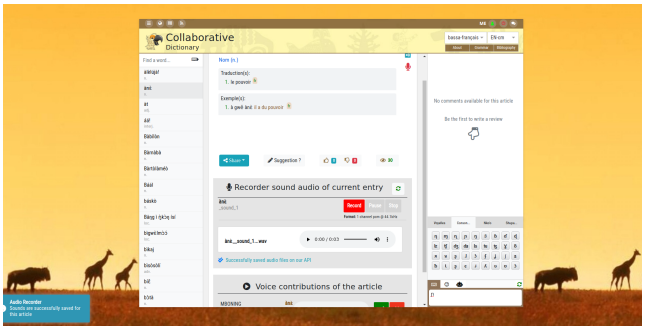

Figure 4: Collaborative dictionaries for sharing multimodal and multilingual lexicographical resources in African languages. This platform is under Creative Commons BY-NC-SA 3.0 license: (`https://ntealan.net`)

4. NTeALan language resources and representation

Most of our dictionaries resources are old bilingual dictionaries (from the work of linguists) found on the web as open-source or under Creative Commons BY-NC-SA 3.0 license. The links to the original sources and to the NTeALan's versions are provided on all our platforms from where they can also be consulted.

4.1. African language resource dictionaries

We currently host and share 7 dictionaries on our APIs. Although the number of articles entries to date is still relatively limited (from 0 to 11,500 entries), a growing community is participating daily in their filling. Table 4.1. shows the current statistics on the resources managed by our API.

Language resources	Entries	Entries contrib.	Media contrib.
Bambara-French	11487	1	1
Yemba-French	3031	2	90
Bassa-French	427	5	5
Duala-French	191	5	0
Ghomala-French	9	1	0
Ngiemboon-French	3	2	1
Fulfulde-French	0	0	0

Table 3: State of the art of NTeALan language resources currently saved in the REST API

Even if the current resources are insufficient and cover only 6 African languages, we are nevertheless satisfied with the craze that is beginning to appear within the communities of users behind our platforms in just one year of existence. We would like to determine whether our different infrastructures fit with the resources produced, the load of connected users and their needs. Once we have completed the tests on the platform, the next steps will be generalizing the model to all others African languages.

4.2. Description of NTeALan's XML format

Each lexical resource management platform has its own model for structuring and presenting data, it is the case for (Mangeot, 2006), Kosh, ELEXIS Dictionary Service and (Benoit and Turcan, 2006). The XML format (mainly the TEI and LMF standards) is today a reference choice for structuring linguistic, lexicographic and terminographic data. We can also mention the TEI Lex-0 (Romary and Tasovac, 2018) and Lexicog (OntoLex Lemon Lexicography from W3C), which are frequently used to codify lexicographic resources. Unfortunately these standards are not often adapted to represent and describe some morpho-syntactic particularities of African languages. Indeed, several linguistic phenomena, such as the concept of nominal class, the notion of clicks and the management of the translation and localisation of dialect variants of the article entry, are not explicitly treated, despite all the needs expressed with regard to the matter[9].

By analyzing the structure of a Bantu language (`Yemba` spoken in West region in Cameroon), we decided to define a proprietary XML structuring model, whose structure was inspired by the 4 major families of African languages, namely: the Afro-Asian family, the Niger-Kordofan family, Nilo-Saharan family and the Koisan family. Three principles guided our choice: representation, simplification and extensibility:

- **representation**: with this principle, we describe the language data at the smallest morpho-syntactic level i.e. word components (prefix+root+suffix) and phrase components like class accord (1/2, 3/4, 5/7, etc.).
- **simplification**: here we choose XML tag names and international languages that are easily understandable for the research communities. Also, we chose to use a linear XML representation, with less parents and more children in the same parent node.
- **extensibility**: we give external contributors the possibility to extend our main XML structures by adding new nodes (children or parent nodes), depending on the element to be represented.

We design our *core-node* lexicographic data with a root node called `<ntealan_dictionary>`, which is divided into two subnodes: `<ntealan_paratexte>` and `<ntealan_articles>`. `<ntealan_paratexte>` describes the metadata around the version(s) of the document (context of the dictionaries production, source

[9]Note that it is nonetheless possible in these standards to add new formalisms (tags and attributes) in addition to existing classes.

description of the original authors and target description of the XML VERSION). <ntealan_articles> describes all the dictionary articles (<article>).

Each article has its own subnodes: <entry> (dialect variant currently processed), <category> (grammatical categor(y/ies) links to the dialect variants), <translations> (translations associated to the dialect), <examples> (contextualisation of dialect variants). Figure 5 illustrates this data representation.

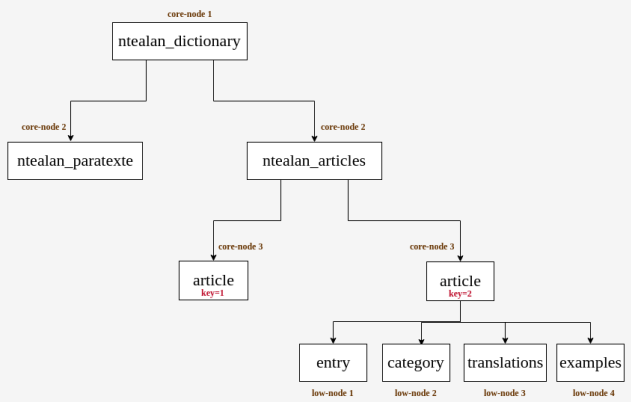

Figure 5: NTeALan dictionaries XML data model

The extension of the article structure by contributors is only possible in *low-node*, as shown in figures 5, 6 and 7, which means that the article model can be updated in each of its nodes, referred to by an id.

```
<article type="Nom">
 <entry forme="simple">
   <variant type="YN" forme="simple">
     <prefix>m</prefix>
     <radical>bā</radical>
  </variant>
   <variant type="YS" forme="simple">
     <radical>mba-nné</radical>
  </variant>
 </entry>
 <category>
   <cat number="1">n</cat>
 </category>
 <classe_d_accords>
   <cl_sing number="1">9</cl_sing>
   <cl_plur number="1">10</cl_plur>
 </classe_d_accords>
 <translations>
   <equivalent lang="fr" number="1">foureau</equivalent>
 </translations>
</article>
```

Figure 6: Sample of Xmlisation of noun article *mbā* extracted from the Yemba-French dictionary

Our XND format is not intended to be standardized to serve as a reference. On the contrary, it is used as intermediate format, required by our internal NLP tools and by well-known standardized formats. Indeed once the external formats are serialized in XND, we have the possibility to convert the data into other formats such as those of the TEI and LMF dictionaries. These features will be available at the API level soon.

```
<article type="Verbe">
  <entry forme="simple">
    <variant type="YN" forme="simple">
      <prefix>le</prefix>
      <radical>baka</radical>
   </variant>
    <variant type="YS" forme="simple">
      <prefix>li</prefix>
      <radical>cu'o</radical>
   </variant>
 </entry>
  <category>
    <cat number="1">v</cat>
 </category>
  <conjugation>
    <conj_variant type="YN">
      <forme_conj type="2-F_infinitive">ḿbáká</forme_conj>
      <forme_conj type="imperative">báká</forme_conj>
   </conj_variant>
    <conj_variant type="YS">
      <forme_conj type="2-F_infinitive">ncú'ó</forme_conj>
      <forme_conj type="imperative">cú'o</forme_conj>
   </conj_variant>
 </conjugation>
  <translations>
    <equivalent lang="fr" number="1" emprunt_En="pack">
        entasser, accumuler
    </equivalent>
 </translations>
</article>
```

Figure 7: Sample of Xmlisation of verb article *lebaka* extracted from the Yemba-French dictionary.

5. Problems encountered and further challenges

The implementation of these first platforms enabled us to take note of the type of challenges that can arise in such a project. We are currently focused on these issues, trying to improve and enrich our platforms.

5.1. Problems encountered

At the moment, we are facing two main difficulties with the NTeALan platforms:

- the first is the low number of contributors and the insufficient IT resources. The staff do not have all the specialists needed (in NLP, NLU, African languages) to reach the targeted goals and great ambitions. The current work is mainly carried out by 4 active members of the association. Regarding IT resources, we do not have enough robust IT infrastructures (servers, field tools) as required by our research work for African languages.

- the second is the lack of funding to carry out our research activities, more precisely for the development of NLP and NLU tools in and for African native languages. Our funding mainly comes from the contributions of the association members, which is not enough in the light of our current ambitions.

5.2. Further challenges

As already explained, our ambitions are great and will require more staff (language specialists) and financial resources. We would like to:

- Above all, encourage the greatest number of specialists in African languages and cultures from various African countries and in the world, to join our association. Together we are more powerful to meet the 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 industrialization and teaching of poorly endowed African languages.

- Improve and enrich all our existing platforms and open them up more to the scientific community and to the speakers of these languages. We mainly focus on : the autonomous platform for teaching languages and cultures, the conversational assistant for language teaching and the virtual cultural museum for safeguarding of the African socio-cultural inheritance.

- Strengthen our partnerships with African social and cultural institutions, universities, research laboratories and companies specialized in our research areas. The aim is to enlarge our already existing communities of experts in linguistics, technological and cultural issues throughout the continent, so that we can keep on working hand in hand for the development of African native languages.

6. Conclusion

In this paper, we described three NTeALan lexicographic platforms and the XND data format used, and we showed how essential an association is nowadays, for the construction of quality linguistic and lexicographic resources and tools for poorly endowed African native and/or national languages. We lead, internally with our academic partners (the 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 African languages. It is obvious that a lot remains to be done, however the first results of our study have proven to be very useful for our applications (conversational agent NTeABot, learning platform, translation platform, etc.) and for the users as well. These results can be used by other researchers: they include data (in different common formats like XML TEI, TEI Lex-0, LMF, XND) and tools. We are convinced, as Tunde Opeibi (Tunde, 2012, p.289) already said, 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."

7. Acknowledgements

These platforms were developed by Elvis Mboning (REST API and Dictionaries Management platform), Daniel Baleba (Collaborative dictionaries) and Jean-Marc Bassahak (Web Design and interfaces). NTeALan's projects are actually supported by NTeALan Language Research, the Ministry of Post and Telecommunication of Cameroon, the Department of linguistics and African literature of the University of Douala (Cameroon), the Research teams ERTIM of INALCO (France) and Fractals system (France). We can also cite: Professor Jules Assoumou, Merci Christian Bonog Bilap, Marcel Tomi Banou, Ntomb David, Ntomb Nicolas, Théophile Kengne, Yves Bertrand Dissake, Juanita Fopa, Damien Nouvel and all our contributors.

8. Bibliographical References

Abate, S. T., Melese, M., Tachbelie, M. Y., Meshesha, M., Atinafu, S., Mulugeta, W., Assabie, Y., Abera, H., Ephrem, B., Abebe, T., Tsegaye, W., Lemma, A., Andargie, T., and Shifaw, S. (2018). Parallel Corpora for bi-lingual English-Ethiopian Languages Statistical Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3102–3111, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

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.

Breuel, T. M. (2008). The OCRopus open source OCR system. *Proc.SPIE*, 6815.

De Schryver, G.-M. (2010). State-of-the-Art Software to Support Intelligent Lexicography. *ResearchGate*, page 16.

Holtzblatt, K. and Beyer, H. (2017). 7 - Building Experience Models. pages 147–206, January.

Mangeot, M. and Enguehard, C. (2011). Informatisation de dictionnaires langues africaines-français. In *journées LTT 2011*, page 11.

Mangeot, M. (2006). Dictionary building with the jibiki platform. In Cristina Onesti Elisa Corino, Carla Marello, editor, *Proceedings of the 12th EURALEX International Congress*, pages 185–188, Torino, Italy, sep. Edizioni dell'Orso.

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 métalexicographie outillée : conception d'un outil pour le métalexicographe 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.

Romary, L. and Tasovac, T. (2018). TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources. In *TEI Conference and Members' Meeting*, Tokyo, Japan, September.

Ruthven, R. (2005). The African Association for Lexicography: After Ten Years. *Lexikos journal*, page 9.

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.

9. Language Resource References

De Pauw, Guy and Waiganjo Wagacha, Peter and de Schryver, Gilles-Maurice. (2009). *The SAWA corpus: a parallel corpus English - Swahili*.

Mboning, Elvis and NTeALan contributors. (2017). *NTeALan lexicographic African language resources: an open-source REST API*. NTeALan Project, distributed via NTeALan, Bantu resources, 1.0.

Vydrin, Valentin and Rovenchak, Andrij and Maslinsky, Kirill. (2016). *Maninka Reference Corpus: A Presentation*. Speecon Project, distributed via ELRA, Madingue resources, 1.0, ISLRN 613-489-674-355-0.