Maninka Reference Corpus: A Presentation

Valentin Vydrin, Andrij Rovenchak, Kirill Maslinsky

Maninka Reference Corpus: A Presentation

Valentin Vydrin[1, 2, 3], Andrij Rovenchak[4], Kirill Maslinsky[5]
(1) INALCO, Paris, France
(2) LLACAN-CNRS (UMR-8135), Villejuif, France
(3) St. Petersburg State University, St. Petersburg, Russia
(4) Ivan Franko National University of Lviv, Lviv, Ukraine
(5) National research university Higher school of econimic, St. Petersburg, Russia
vydrine@gmail.com, andrij.rovenchak@gmail.com, maslinych@gmail.com

RÉSUMÉ

Présentation du Corpus Maninka de Référence
Un corpus annoté du maninka de Guinée, Corpus Maninka de Référence (CMR), a été publié en avril 2016. Il comporte deux sous-corpus : l'un contient des textes créés originalement en orthographe latine (792 778 mots), l'autre est composé des textes en alphabet N'ko (3 105 879 mots). La recherche peut être effectuée dans les deux sous-corpus en utilisant soit l'orthographe latine, soit le N'ko. L'outillage utilisé pour le CMR est représenté d'abord par le paquet de logiciel Daba (développé initialement pour le Corpus Bambara de Référence). Le logiciel NoSketchEngine est utilisé comme le moteur de recherche; nous avons adapté ce logiciel au script N'ko, qui s'écrit de droite à gauche. Tous les textes en N'ko ont été obtenu sous format électronique qu'il a fallu normaliser (utilisation de polices pré-Unicode). L'annotation morphologique est basée sur le dictionnaire électronique Malidaba qui est actuellement à une stade itermédiaire d'élaboration; il faut encore beaucoup d'efforts pour l'amener à un état acceptable.

ABSTRACT

An annotated corpus of Guinean Maninka, Corpus Maninka de Référence (CMR), was published in April 2016. It includes two subcorpora: one contains texts originally written in Latin-based graphics (792,778 words), and the other one is composed of texts in N'ko alphabet (3,105,879 words). Both subcorpora are searchable in both Latin-based graphics and in N'ko. In the building CMR, the Daba software package (earlier developed for the Corpus Bambara de Référence) has been used. As the search tool, NoSketchEngine has been used, it was adapted to the right-to-left direction of the N'ko writing. All texts in N'ko were obtained in electronic format, most of them were converted from pre-Unicode fonts. The morphological annotation is based on the Malidaba electronic dictionary which is in an intermediary stage of compillation; much efforts is needed to bring it to a minimally acceptable state.

MOTS-CLÉS : Corpus Maninka de Référence, N'ko, Malidaba, constitution de corpus écrits.
KEYWORDS: Corpus Maninka de Référence, N'ko, Malidaba, corpus building.

[illegible]

[illegible]

[illegible] NoSketchEngine. [illegible] "UNICODE" [illegible]

[illegible]: [illegible]

1. Corpus Maninka de Référence : general information

In Valentin Vydrin's presentation at TALAf-2014, it was said that a preliminary version of the Maninka Reference Coprus (Corpus Maninka de Référence, CMR) had been put online, although at that time we avoided to publicize it because of its numerous shortcomings. Two years later (more precisely, in April 2016), we finally decided to open it for the public: although still very far from perfection (and much less elaborated in comparison with its "elder brother" Bambara Reference Corpus), the CMR has reached the minimal requirements and can be used. The CMR site is available in two versions, French http://cormand.huma-num.fr/cormani/, and Russian, http://maslinsky.spb.ru/cormani/.

In what follows, we present main characteristics of CMR and the difficulties we had to overcome in the course of its development.

CMR consists of two subcorpora: Corpus Maninka (texts originally created in Latin alphabet; further on, “Latin Subcorpus”) which has now reached 792,778 words, and Corpus N'ko (texts originally written in the N'ko writing) of 3,105,879 words. It is provided with information about the project, with documentation necessary for its correct use (lists of glosses and of POS markers; explanations concerning the tonal notation and the principles of file naming for the texts included into CMR) and with the user guide. The entire corpus is automatically annotated for POS and French glosses. So far, no disambiguated subcorpora are available (for the reasons presented below).

Following the general practice, the texts included to the CMR are not available for the users in their integrity (otherwise, serious copyright problems could arise). Sentences containing a searched word are presented in a concordance, and the context can be extended to a sentence to the right and to the left.

The Latin Subcorpus is available for search in both Latin-based and N'ko alphabets. The N'ko version has been generated automatically (on the basis of a Latin to N'ko converter developed by Andrij Rovechnak). As far as the original texts in Latin script has no tonal marks, a "default" tonal notation is used in the converted N'ko version (all tones are marked as high, except on the long ɔɔ vowel, where the more frequent low tone is marked).[1]

The N'ko subcorpus is also searchable in both N'ko and Latin versions. In original texts written in N'ko, tonal marking is obligatory on every vowel. The automatically generated Latin version is also tonalized throughout. The standard model of tonal notation in the Latin version (identical to the model used in the Corpus Bambara de Référence) allows a search in a tonalized text disregarding contextual modifications of tones, so that all occurrences of a lexeme in the Corpus can be found in one search (which is often impossible for a search in the N'ko graphics; each context-conditioned tonal form is to be searched separately).

In the building of CMR, three major components can be singled out: adaptation of the software instruments to the specifics of Maninka texts; collecting and preparation of texts; building of an electronic lexical database. Let us consider these tasks one by one.

2. The instruments

For the CMR, it has been decided to use the Daba software package, initially developed for the Bambara Corpus (Maslinsky 2014). As a search tool, the NoSketchEngine (a version already adapted to the specifics of the Bambara language) has been chosen. The reason is that Bambara and Maninka languages are closely related, therefore, the adaptation of the instruments could be kept to a minimum. The main difficulty was represented by the right-to-left direction of the N'ko writing and its compatibilization with the annotation lines represented in Latin graphics (the Latin version of the Maninka text, the POS-tags, the French glosses, and several auxiliary lines of annotation).

As shown in (Maslinsky 2014), the tools of the Daba package include:

- an interface for adding metadata to the texts (Metaeditor);
- an interface for the automatic morphological analysis (gparser);
- an interface for the manual disambiguation (gdisamb);
- a list of auxiliary words and bound morphemes; a file representing rules of allowed combinations of grammatical morpemes within a word form.

These tools have been adapted to the specifics of Maninka by Kirill Maslinsky. The only adaptation needed for the morphological parsing of Maninka text was to develop a set of parsing rules in the Daba syntax. It took much more effort to adapt the NoSketchEngine interface. The major task was to allow for the transparent selection of the main writing direction (right-to-left or left-to-right) in the concordance lines. Standard version of the NoSketchEngine has a rudimentary support for the right-to-left script. Right-to-left may be set as a global option to the corpus so that right and left contexts are swapped in the concordance. This simplistic approach proves inefficient when applied to the contexts where N'ko source texts are set back to back with Latin annotation. Situation is further complicated by multitude of options for displaying annotation supported by NoSketchEngine (for key token only, for all tokens, options to select multiple lines of annotation

[1] So far, we haven't been able to discover any reliable statistical or pattern-based rules for tone marking. A breakthrough in this domain is expected as the dictionary is ready.

etc.) When standard right-to-left option in NoSketchEngine was used, it resulted in many cases in unnatural ordering of right and left contexts and tokens inside contexts both for right-to-left and left-to-right text. To solve this problem default concordance display templates of the NoSketchEngine were modified to adaptively select writing direction for each line in a concordance based on the prevailing script in the sentence context: N'ko or Latin.

Unicode convertors for all pre-unicode N'ko fonts have been developed by Andrij Rovenchak. For convenience of N'ko users, the online convertor from pre-Unicode fonts was created, http://nkoconvert.ho.ua/. Andrij Rovenchak has also developed a convertor from N'ko to Latin-based Maninka tonal orthography and a convertor from Latin-based orthography to N'ko (non-tonal).

Work on the Malidaba lexical database is described in section 4.

3. The collection of the texts

The specifics of the situation with written texts in Guinean Maninka lies in coexistence of two different graphic systems, Latin-based and N'ko. The Latin-based (non-tonal) orthography was broadly used during the period of the 1st Republic (1958-1984) in the school and university education, in written press, in literacy campaigns. Unfortunately, libraries are quasi-inexistant in Guinea, and archives of services responsible for the publication of literature in national languages are in deplorable state. The great majority of publications of that period has disappeared without leaving a trace, and the amount of Maninka texts available in old Latin orthography is rather low; so far, it hardly exceeds 100,000 words (we still hope to find more texts in personal collections and in European and North American university libraries). Under the 2nd Republic, an orthographic reform took place (1989) in order to bring the orthographies of Guinean languages closer to those used in neighbouring countries. At the same time, the language policy radically changed: national languages were removed from the curricula, state-sponsored periodicals in national languages disappeared, and the amount of literature published harshly decreased. Still, books are published from time to time by Institut de Recherches Linguistics Appliquées (IRLA), by NGOs and by Protestant missionaries; the most voluminous text available is a Bible translation (so far incomplete).

The N'ko alphabet was created in 1949, and it has never been officialized in Guinea or in any other country of the Manding language area. The lack of official recognition is compensated by the dynamism of the broad grass-root movement: there is a dense network of informal and semi-formal N'ko schools all over the Manding language area and in the numerous Manding diaspora; there are several periodicals, and the number of books appearing in N'ko every year exceeds by far the number of publications in Manding languages in Latin script.[2] It should be also mentioned that the N'ko activists are very responsive to the technical innovations, which explains a tangible presence of N'ko in the Internet.

We started collecting texts for the corpus in 2009. Unlike in the Corpus Bambara de Référence project, manual typesetting represents a minor part of the job: it concerns only texts written in Latin script, and among these, the Bible translation (which by far exceeds in volume the totality of all other texts) was put to our disposal by missionaries of the Pioneers' Bible Translators.

All the N'ko texts were received in various electronic formats. One of the major sources is the web-site of Mamadi Baba Jaane, an outstanding N'ko promoter living in Cairo,

[2] N'ko is widely spread not only in Guinea, but it is this country that its position is the strongest. On the dynamism of the N'ko writing, see in particular (Vydrine 2011).

http://www.kanjamadi.com. Numerous texts in the electronic form were obtained from the N'ko Academy (*Ńkó Dúnbu*) via Ibrahima Sory 2 Condé, who is Secretary for Scientific Research of the *Ńkó Dúnbu* and whose active stance has very much contributed to the success of the CMR project.

The N'ko alphabet became a part of the Unicode standard in July 2006 (version 5.0) and such an encoding is gradually becoming the most widely-used one. However, a large number of available texts were created using pre-Unicode fonts, with N'ko characters placed at the codepoints corresponding to Arabic letters. Two ways to map the N'ko script onto the Arabic script are represented by the following typefaces:

1) *BOURAMA-KANTE*, *Nko Manding1 Cote D'Ivoiredf1*, *N'ko2 Manden-1L*, *N'ko2 Manden-1sa*, *Solomana Kante;*

2) *A Manding BATEKA, A Manding IT BAMA*, *Fofona*, *Karifala Berete*, *Mading*, *Mamudu Bamba*, *Manding N'ko Sigui*, *Nko Africa*, *Nko Kouroukan Fouwa*, *Nko Koutoub Sano*, *Nko Kwame Nkurumah.*

In one typeface, *NkoHeb*, the N'ko script was mapped onto the Hebrew script. Presently, we possess only one text with such an encoding.

A large number of documents are available as PDF files. At least four encoding schemes were discovered while trying to extract texts from various PDF files containing N'ko text. The final output text is encoded in UTF-8.

A large share of texts in the corpus (about 700 thousand words) come from periodicals, in particular from the following titles: *Dàlu Kɛ́ndɛ* (the periodical issued by the N'ko Academy being a source for more than 50 per cent of texts in this group)*, Yèreyá fɔɔbé, Sínjiya fɔɔbé, Jànsán, Yélen, Sékutureya Kìbaro, Màndenká*. This material was a subject of a recent quantitative linguistical studies (Rovenchak 2015).

Over 500 thousand words in the corpus come from the translation of Qur'an with tafsirs. Numerous texts of books of various genres, texts of public speeches, working documents of N'ko associations, etc. were collected by Ibrahima Sory 2 Condé (with the agreement of their authors).

The responsiveness of the N'ko community allowed to compile a corpus of more than 3 million words with much less effort than a smaller Bambara corpus (which counts by now slightly less than 3 million words).

4. The Malidaba dictionary

A crucial tool for the building of an annotated corpus is an electronic dictionary. In this relation, the CMR was from the beginning in a less favorable situation than the Bambara Corpus: there is no Maninka-French dictionary available in electronic format. There are however several lexicographical sources for Maninka which can be used for building such a dictionary, of which the major ones are the followig:

- a French-Maninka dictionary (Kántɛ 2012) containing about 4000 entries; an electronic version has been put at our disposal by the *Ńkó Dúnbu*;
- a monolingual Maninka dictionary (Kántɛ 2003) containing about 33,000 entries (an electronic version of this dictionary is also available);
- a Manding-English Dictionary (Vydrine 1999), or rather its updated (2015) version (in this version, some entries have French equivalents too).

The Malidaba[3] electronic dictionary is in construction. As a departure point, a concordance of the word forms aligned by their frequencies was generated on the basis of a 2-million word Maninka corpus, and about 300 most frequent words (together with their homonyms and quasi-homonyms) were integrated into Malidaba (with all necessary lexicographical information). This mini-dictionary assured glossing of about 50% of tokens in the Corpus.

At the next stage, the French-Maninka Dictionary (Kántɛ 2012) was automatically reversed and converted into the Toolbox format. The resulting database can by no means be regarded as a true Maninka-French dictionary, however, it can serve as raw material (with all necessary precautions) in the dictionary-building work.

Then, the Maninka data (with English or French equivalents) were automatically extracted, by Andrij Rovenchak and Kirill Maslinsky, from the Vydrin's Manding-English Dictionary (in this dictionary, both Maninka and Bambara data were included, and the extraction of the data of one language represented a nontrivial challenge requiring a complicated algorithm).

The three sources (the 300 most frequent words; the extract from the enlarged version of Vydrine 1999; Kántɛ 2012) were merged, which brought us to an electronic dictionary of about 8200 entries. This dictionary was extremely “dirty” (numerous duplicate entries; even more numerous entries with unfitting equivalents) and unbalanced (the vocabulary is very well covered for the initial letters of the alphabet, and much less so for the rest of the dictionary); in numerous entries imported from Vydrine 1999, English glosses appear instead of French ones. In this state, Malidaba has assured glossing of more than 90% of all the occurrences in the CMR, however, the quality of the glosses is currently below any standards.

The most urgent task today is a cleaning of Malidaba, a task which is being carried out by Valentin Vydrin (unfortunately, rather irregularly); from December 2015 to April 2016, the first 900 entries were cleaned (which also brought forth the elimination of about 250 duplicate entries). Before the first cleaning of Malidaba is completed, it is hardly expedient to start disambiguation of the CMR.

5. Perspectives

Even in its present state, CMR can be successfully used for various tasks, and the growing statistics of visits shows that it is being already used.[4] As has been already said, the most urgent task is to wrap up the preliminary cleaning of the Malidaba electronic dictionary. At the same time, collection, convertion and integration of further texts into CMR is going on. Another time-consuming task that could be carried out simultaneously with the cleaning of Malidaba is adding metatextual information to the files; this job can be entrusted to our Guinean partners from *Ńkó Dúnbu,*however, such a solution requires preliminary training of the workers (preferably in France).

---

[3] Abbreviation for Malinke DataBase (Malinke is the French word for Maninka). In Manding, *màli dába* can be also interpreted as 'hyppopotamus' big mouth'. Cf. the Bambara electronic dictionary Bamadaba (Bambara DataBase), where interpretation *bàma dába* 'crocodile's big mouth' is possible.

[4] 10 visits by 9 unique visitors during the second half of April 2016; 27 visits by 21 unique visitors during the initial 8 days of May. It is very little if compared with the current statistics of visits of the Corpus Bambara de Référence (in 2016, more than 600 unique visitors every month), but it is comparable with the number of visits of the Bambara Corpus 3 years ago.

After the primary cleaning of Malidaba, a Maninka/N'ko spellchecker could be produced, following the model of the Bambara spellchecker (cf. Méric 2014; in fact, a preliminary version has been already created on the basis of the Maninka concordance by Jean-Jacques Méric); it will be perfectioned together with the improvement of the Malidaba dictionary.

If we follow the model already applied to the building of the Bambara Corpus, we should begin disambiguation of automatically annotated texts as soon as Malidaba attains certain degree of acceptability. An alternative (or a parallel) way could be an application of computer technologies for the corpora of underesourced languages which could allow to find numerous shortcuts.

Acknowledgments

This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the “Investissements d’Avenir” program (reference: ANR-10-LABX-0083).

The authors are grateful to Ibrahima Sory 2 Condé, the Secretary for Scientific Research of the *Ńkó Dúnbu*, for his ethusiasm, energy and assistance in collecting N'ko texts and other activities related to the building of the N'ko Corpus, and to Mamady Diané, the director of the Institut de Recherches Linguistiques Appliquées (Conakry) for his continuous support in the work of collection of Maninka texts in Latin-based orthography.

Références

Kántɛ, Sùlemáana. (2003) *Ńkó kɔ́dɔfɔlan Kànjamáadi Màndén fòdoba kán ní kèlennatɔ́ɔ kàn.* [*Dictionnaire "N'ko" Kanjamaadi linguistique et de noms propres.*]. Ed. by Bàba Jàanɛ, Caire.

Kántɛ, Sùlemáana. (2012). *Dictionnaire bilingue français-N’ko.* Conakry.

Maslinsky K. (2014). Daba: a model and tools for Manding corpora. In : Mathieu Mangeot, Fatiha Sadat (éd.). *Actes de l'atelier sur le traitement automatique des langues africaines TALAf 2014.* (*Actes des Ateliers TALN 2014.* Éd. par Brigitte Bigi) http://www.taln2014.org/proceedings/Ateliers/TALAf2014/Paper_TALAf-O.3.pdf

Méric J.-J. (2014). Un vérificateur orthographique pour la langue bambara. In : Mathieu Mangeot, Fatiha Sadat (éd.). *Actes de l'atelier sur le traitement automatique des langues africaines TALAf 2014.* (*Actes des Ateliers TALN 2014.* Éd. par Brigitte Bigi) http://talaf.imag.fr/2014/Actes/MERIC%20-%20Un%20verificateur%20orthographique%20pour%20la%20langue%20bambara.pdf

Rovenchak A. (2015). Quantitative studies in the corpus of Nko periodicals. *Recent Contributions to Quantitative Linguistics,* edited by Arjuna Tuzzi, Martina Benešová and Ján Mačutek (Berlin–Boston: Mouton de Gruyter), 125-138.

Vydrine V. (1999). *Manding-English Dictionary (Maninka, Bamana). Vol. 1.* St. Petersburg: Dmitry Bulanin Publishing House,
Vydrine V. (2011). L'alternative du N'ko : une langue écrite mandingue commune, est-elle possible ? In: Vold Lexander, Kristin; Lyche, Chantal, Moseng Knutsen, Anne (eds.). *Pluralité des langues pluralité des cultures : regards sur l'Afrique et au-delà* (Oslo: The Institute for Comparative Resarch in Human Culture), 195-204.

Vydrin V. (2014). Projet des corpus écrits des langues manding : le bambara, le maninka. In : Mathieu Mangeot, Fatiha Sadat (éd.). *Actes de l'atelier sur le traitement automatique des langues africaines TALAf 2014.* (*Actes des Ateliers TALN 2014.* Éd. par Brigitte Bigi) http://www.taln2014.org/site/actes-en-ligne/actes-en-ligne-ateliers/