Projet des corpus écrits des langues manding : le bambara, le maninka[1]

Valentin Vydrin
LLACAN, CNRS UMR-8135, 7 rue Guy Môquet - BP 8, 94801 Villejuif Cedex
INALCO, 65 rue des Grands Moulins, CS21351, 75214-PARIS cedex 13
vydrine@gmail.com

Résumé. Le projet des corpus électroniques de textes en langues mandingues a démarré à St. Petersbourg en 2009. Aujourd'hui, il est effectué par une équipe internationale avec l'implication des spécialistes en langues manding des pays différents. L'outillage tenant compte des caractéristiques spécifiques des langues manding (mais adaptable aux autres langues) a été développé. Le Corpus Bambara de Référence est mis en ligne en 2012, suivi par un corpus maninka (en écriture N'ko et latine) en février 2014. Un correcteur automatique d'orthographe bambara et un logiciel du ROC pour le bambara a été développé sur la base de l'outillage du CBR. L'utilisation expérimentale du CBR dans l'enseignement universitaire du bambara et dans les études linguistiques a montré son efficacité. L'expérience accumulée peut être facilement étendue sur les autres variétés manding (le dioula de RCI, le dioula de Burkina Faso), mais aussi sur d'autres langues africaines.

Abstract. The project of electronic corpora for Manding languages was launched in St. Petersburg in 2009. By now, it is carried out by an international team with an assistance by specialists in Manding languages from different countries. Tools have been developed taking into account the specifics of Manding languages (and adaptable to other languages). The Bamana Reference Corpus was put on line in 2012, it was followed by a Maninka corpus (in both Roman and N'ko writing) in February 2014. An orthography corrector for Bamana and a software for the Bamana OCR has been developed on the basis of the Bamana Reference Corpus tools. An experimental use of the Bamana Corpus in the Bamana teaching in universities and in linguistic studies has proved its effectiveness. The experience accumulated in the framework of this project can be relatively easily extended to other Manding varieties (Jula of Côte d'Ivoire, Jula of Burkina Faso), and, if necessary, to other African languages.

Mots-clés. Corpus Bambara de Référence, Bambara, Manding, Maninka, Malinké

Keywords. Bambara Reference Corpus, Bambara, Bamanankan, Manding, Maninka, Malinké

[1] Ce travail a bénéficié d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme Investissements d'Avenir portant la référence ANR-10-LABX-0083

1. Introduction

Le travail sur le Corpus Bambara de Référence (CBR) a démarré à St. Petersbourg en 2009 par un groupe de trois linguistes russes, spécialistes en langues mandé, assistés par un linguiste et informaticien Kirill Maslinsky. Actuellement, il s'agit d'une équipe internationale semi-formelle des linguistes et informaticiens russes, français, ukrainiens, avec une participation active des collègues africains. Les résultats du travail de notre équipe est un Corpus Bambara de Référence et un Corpus Maninka de Référence disponibles en ligne (libre accès), et certains autres outils créés sur la base de ces corpus. (Dans les deux cas, il s'agit des corpus des textes écrits.)

Dans cette communication, je présenterai brièvement les deux corpus et l'outillage développé dans le cadre de notre projet. Je traiterai surtout des aspects linguistiques de notre travail (les détails techniques seront présentés dans la communication de Kirill Maslinsky).

Avant de présenter les résultats de travail de notre équipe, il faut mentionner quelques particularités des langues manding s'avérant pertinentes pour ce travail.

Les langues manding sont isolantes, ayant très peu de morphologie flexionnelle. D'un part, cela facilite le développement d'un analyseur automatique morphologique, d'autre, cela le rend peu puissant. La forme du mot fournit très peu d'information sur son appartenance aux classes de lexèmes (annotation POS), ce qui est aggravé par une conversion très fréquente dans les langues manding.

Ces langues sont tonales (deux tons unis ; un morphème grammatical tonal, l'article), mais dans les publications dans ces langues, les tons ne sont presque jamais notés (sauf dans les publications académiques fait par linguistes).

Le résultat en est qu'une analyse automatique des textes mandingues s'appuyant sur une base lexicale (un dictionnaire électronique) et un analyseur morphologique produit une homonymie très élevée : moyennement, environ 70% de tous les mots d'un texte ont deux variantes d'analyse ou plus (ce qu'on peut comparer avec environ 30% pour une langue comme le russe).

2. Le logiciel

Il s'est avéré que les logiciels disponibles pour l'étiquetage automatique des textes sont difficilement adaptables aux langues manding se caractérisant de la quasi-absence de la morphologie flexionnelle. Un paquet de programmes « Daba » a été développé par Kirill Maslinsky sur le plateforme Python ; ces logiciels sont constamment améliorés en tenant compte du feed-back. Le paquet comporte les logiciels suivants :

- des convertisseurs orthographiques : l'ancienne orthographe bambara vers la nouvelle orthographe ; l'orthographe tonale de Charles Bailleul vers l'orthographe tonalisée standardisée ; plus tard, Andrij Rovenchak (Lviv, Ukraine) a élaboré le convertisseur de l'écriture N'ko en orthographe standard latine tonalisé pour le maninka, et ce convertisseur a été intégré dans le Daba ;
- un analyseur morphologique sur la base d'un dictionnaire bambara et tenant compte des règles combinatoires des morphèmes flexionnels et dérivationnels. Cet analyseur produit un texte bambara annoté (POS, les lemmes, les gloses françaises). Tout récemment, cet analyseur a été adapté à la langue maninka, ce qui montre sa flexibilité ;
- une interface graphique pour l'introduction des métadonnées (le modèle basé sur les recommandations de l'EAGLES (Sinclair Ball 1996)
- une interface graphique pour la désambiguïsation semi-automatique des textes annotés automatiquement.

3. Dictionnaires

Le dictionnaire électronique bambara-français Bamadaba a été développée dans le cadre du projet sur la base du dictionnaire de Charles Bailleul (2007), qui a été sérieusement réarrangé et uniformisé : des nombreux doublets ont été éliminés, la présentation des variantes phonétiques a été standardisée, les étiquettes POS homogénisées, les mots composés et dérivés ont été dotés des renvois aux composantes. La partie la plus difficile de ce travail a concerné la présentation des équivalents français : il fallait clairement distinguer entre les équivalents et les définitions (ce qui n'a pas été souvent le cas originellement) ; choisir un seul équivalent (parmi tous les équivalents qu'on attribue à un lexème polysémique) qui pourrait servir d'une glose ; créer des gloses pour des lexèmes sans équivalents ; standardiser la présentation de la polysémie ; faire la liste des gloses standards pour les mots et morphèmes auxiliaires.

Depuis sa création en 2010, la base lexicale électronique Bamadaba est en évolution permanente : au cours du travail de désambiguïsation, des nouveaux lexèmes sont rajoutés (le Bamadaba comporte maintenant environ 5% plus d'entrées que dans la première version), des erreurs sont corrigés, des équivalents peu convenables sont remplacés par d'autres, etc. Le perfectionnement du Bamadaba se passe en consultation permanente, par le moyen d'une liste de discussion, avec les meilleurs spécialistes en langues mandingues (Charles Bailleul, Gérard Dumestre, Denis Creissels, Kalilou Téra, Aby Sangaré, Boubacar Diarra participent dans cette discussion régulièrement, et d'autres le font épisodiquement). D'autres dictionnaires bambara, surtout (Dumestre 2011) et (Vydrine 1999), sont largement utilisés

comme des sources d'enrichissement du Bamadaba. Dès janvier 2014, le Bamadaba est affiché (sous format Lexique-Pro) sur le site du Corpus Bambara de Référence ; il est prévu que ses versions mises à jour y seront publiées régulièrement.

En janvier 2014, un premier pas a été fait vers le développement d'un dictionnaire électronique du maninka guinéen, « Malidaba », où touts les mots sont présentées en caractères latins et en N'ko. Comme aucun dictionnaire maninka-français n'est disponible, nous avons décidé de le développer à partir d'une concordance des mots-formes d'un corpus des textes maninka comportant environ 2 millions de mots (à ce sujet, cf. ci-dessous). Cette concordance a été rangée dans l'ordre décroissant des fréquences, ce qui nous permet de doter des équivalents d'abord les mots les plus fréquents de la langue. Ainsi, la création d'un dictionnaire maninka sera véritablement « corpus-driven ». Une participation active des linguistes guinéens et des activistes du mouvement du N'ko dans ce projet est prévue, l'obstacle principal étant la faible connexion à l'Internet en Guinée.

4. Développement des corpus

Les premières versions du logiciel Daba ont été testées sur le corpus électronique de textes bambara de 102 000 mots très gentiment mis en notre disposition par Gérard Dumestre. Au moment où la première version du Corpus Bambara de Référence a été mise en ligne en avril 2012, elle comportait environ 1 000 000 mots, dont 20 000 dans les textes désambiguïsés. En avril 2014, le volume du Corpus a atteint presque 1 681 000 mots, dont presque 229 000 dans le sous-corpus désambiguïsé. Nous faisons un effort pour que le Corpus représente les genres principaux du bambara écrit : journaux en bambara, belles-lettres (la prose et la poésie), littérature orale (épopées, contes populaires, devinettes…), livres d'alphabétisation fonctionnelle, documents juridiques et politiques, textes religieux… D'autre part, nous cherchons à équilibrer le Corpus du point de vue diachronique.

Un grand obstacle dans ce travail est une très faible présence du bambara dans l'Internet, ce qui ne nous laisse pas d'accès à des grands massifs des textes numérisés. Certes, une partie du Corpus consiste en textes qu'on nous a fournis sous format électronique (la traduction de l'Ancien Testament ; les numéros du mensuel Jɛkabaara des derniers années ; des collections de textes de Gérard Dumestre et de Charles Bailleul), mais actuellement la croissance du Corpus continue surtout par la numérisation manuelle des textes disponibles sur le papier.

Jusqu'ici, la numérisation est effectuée par une double saisie manuelle (par deux personnes différentes ne maîtrisant pas le bambara), suivie par le collationnement des deux versions (de préférence, par une personne maîtrisant le bambara). Cela nous permet d'atteindre une identité parfaite du texte numérisé avec l'original. Ensuite, le coordinateur du projet introduit les méta-données, il fait passer le texte par l'analyseur automatique et le met dans le Corpus. Certains textes sont envoyés aux opérateurs de désambiguïsation (maîtrisant bien le bambara).

En mars 2014, un logiciel pour l'OCR a été adapté au bambara par Jean Jacques Méric. Nous espérons qu'à l'avenir, nous pourrons passer à une procédure mixte : on comparera une version de chaque texte saisi manuellement avec le même texte OCRisé, ce qui permettra d'accélérer la croissance du volume du Corpus et de rendre ce processus moins coûteux.

Pour le maninka de Guinée, nous avons une situation tout différente. Grâce au dynamisme du mouvement culturel N'ko, un grand nombre de textes (pratiquement tous, en graphie N'ko) sont disponibles sous format électronique, le plus souvent sous format PDF convertis du Word, et parfois sous format Word. Nous avons téléchargé un grand nombre des textes de l'Internet (surtout les périodiques), et un massif important de textes nous a été fourni par Ibrahima Sory 2 Condé, le Secrétaire Scientifique de l'Académie N'ko (*Ńko Dúnbu*). En février 2014, un corpus maninka « semi-annoté » de 3 millions mots a été mis en ligne. Dans le Corpus Maninka, les textes sont accessibles en versions N'ko et latine, mais ils ne sont pas encore dotés de méta-données, et une partie des mots seulement (surtout les mots grammaticaux et les plus fréquents) sont annotés.

5. Le site du Corpus Bambara de Référence

Le Corpus Bambara de Référence a été mis en ligne en avril 2012, cf. http://cormand.tge-adonis.fr/. Pour le moteur de recherche, le NoSketchEngine a été choisi. Ce logiciel permet une recherche par la forme du mot dans le texte original, par sa lemme, par la glose, par le POS; on peut combiner des paramètres de recherche. Une recherche dans le corpus entier ou dans le sous-corpus désambiguïsé est prévue, on peut chercher en tenant compte des tons ou en les ignorant.

On trouve sur le site du Corpus Bambara de Référence des informations nécessaires pour l'utilisateus : un guide d'utilisation du Corpus ; des listes des gloses standards des affixes et des mots auxiliaires ; les principes de la notation tonale dans le Corpus ; la liste des documents inclus dans le Corpus (séparément pour les sous-corpus désambiguïsé et non-désambiguïsé). Sur le même site nous avons mis le dictionnaire électronique Bamadaba sous format Lexique-Pro.

Le Corpus est en accès libre (apparemment, de nos jours, c'est le seul grand corpus d'une langue de l'Afrique au sud de Sahara librement accessible). Les mises à jour du Corpus Bambara de Référence et du dictionnaire Bamadaba se font normalement tous les trois mois.

Au moment où j'écris ce texte, le Corpus maninka n'est pas encore doté de l'outillage comparable à celui du Corpus Bambara. Il est affiché sur le site mandelang.org ; en attendant, nous ne faisons pas encore de la publicité de ce corpus, mais en principe, on peut l'utiliser (et on l'utilise déjà) dans les recherches.

6. L'utilisation du Corpus Bambara de Référence

6.1. Les corpus mandingues dans l'enseignement et la recherche

Depuis sa publication en ligne en 2012, le Corpus Bambara de Référence est de plus en plus utilisé dans l'enseignement du bambara à l'INALCO et à l'Université d'État de St. Petersbourg.[2] Cela se fait de façons suivantes :

- sélection (par l'enseignant) d'exemples illustratifs naturels pour les exercices de grammaire ;
- recherche dans le Corpus et l'analyse des occurrences des phénomènes grammaticaux étudiés dans le cadre du cours de grammaire bambara par les étudiants ;
- études individuelles par les étudiants sur les sujets ponctuels suggérés par l'enseignant. Ainsi, en 2013/2014, les étudiants de l'INALCO du niveau L2 ont fait des recherches sur la polysémie de certains verbes bambara (ce qui peut être vu comme une première ébauche d'un futur projet d'un dictionnaire « corpus-driven » du bambara) ;
- désambiguïsation des textes bambara par les étudiants. Ce travail s'avère un exercice excellent permettant aux étudiants d'atteindre très rapidement un niveau élevé de maîtrise de morphosyntaxe bambara.

Les premières tentatives des études grammaticales du Corpus Bambara de Référence (sur les adverbes préverbaux ; sur l'infinitif) ont confirmé le fait qu'il s'agit d'un outil très puissant permettant d'élever le niveau d'études linguistiques très considérablement. On peut pronostiquer que dans peu de temps, aucune étude sur la grammaire bambara faite sans recours au Corpus Bambara de Référence (ou un corpus bambara alternatif, si quelqu'un le développe) ne serait plus acceptable.

Le peu de temps passé depuis le lancement du Corpus Maninka (février 2014) ne nous permet pas encore d'évaluer son impact. Cependant, en tenant compte de l'attitude très active des membres du mouvement culturel N'ko et leur avidité des innovations techniques portant sur la langue maninka et l'écriture N'ko, on peut prédire que ce corpus sera en haute demande.

6.2. Des outils développés sur la base du Corpus Bambara de Référence

Le lancement du projet du Corpus Bambara de Référence a permis de développer assez facilement, sur la base de son outillage, de quelques applications pratiques :

- un correcteur automatique de l'orthographe bambara pour le Libre Office (et quelques autres logiciels). La première version a été développée par Andrij Rovenchak, sur la base du Bamadaba et de la présentation formalisée de la grammaire bambara du Corpus Bambara de Référence. Le développement de ce logiciel a été continué par Jean Jacques Méric (qui présentera ses résultats dans sa communication, ce qui me délivre de l'obligation d'en parler en détail) ;
- un logiciel pour le ROC des textes bambara (avec un correcteur automatique d'orthographe), développé par le même Jean Jacques Méric.

Des premières tentatives par J. J. Méric du développement des outils analogues pour le maninka en écriture N'ko ont donné des résultats globalement positifs.

7. Perspectives

Le travail de développement du Corpus Bambara de Référence est en cours. Dans la perspective la plus proche, il est prévu d'y inclure une sélection des numéros des périodiques en bambara, *Kibaru* et *Jɛkabaara,* couvrant toutes les périodes de leur existence (au moins un numéro par an), mais aussi des périodiques plus éphémères (*Kɔlɔnkisɛ, Saheli, Kalamɛnɛ, Ɲɛtaa, Jama*) ; la traduction bambara du Quran. Très prochainement, le volume du CBR doit dépasser 2 millions de mots. Au même temps, on travaille constamment sur le nettoyage du CBR et le perfectionnement de son outillage.

Au moment donné, nous avons une sélection suffisante pour démarrer un corpus parallèle bambara-français. Des recherches dans cette direction ont été effectuées par Andrij Rovenchak et Solomija Buk (2013). Ce corpus parallèle peut être lancé avec relativement peu d'efforts, une fois que nous trouvons du financement.

[2] Je n'ai pas d'information précise concernant son utilisation dans d'autres universités européennes, américaines et africaines où le bambara est enseigné.

Un corpus oral bambara est un autre objectif important. Pour le moment, les forces humaines et surtout le financement nous manquent pour nous lancer dans ce projet, tandis que les méthodes de sa création, les logiciels nécessaires et les enregistrements audio sont disponibles.

Le Corpus Maninka est à son stade initial, et beaucoup d'efforts doivent être appliqués pour l'amener au niveau d'élaboration comparable avec celui du CBR. Cependant, nos partenaires guinéens de l'Académie N'ko se disent prêts à participer activement dans le travail de désambiguïsation et développement du dictionnaire électronique, ce qui nous donne beaucoup d'espoir.

Il y a une bonne perspective pour un corpus dioula de Côte d'Ivoire. Actuellement, deux collègues ivoiriens, Kalilou Téra et Aby Sangaré, travaillent activement sur un dictionnaire dioula-français (avec un support logistique de notre équipe), et ce dictionnaire (sous format Toolbox) est déjà à un stade avancé. On peut passer assez facilement au lancement d'un corpus, en utilisant l'outillage de CBR. Un grand obstacle représente le petit volume de textes disponibles en dioula de Côte d'Ivoire.

On pourrait développer un corpus du dioula du Burkina Faso ; pour cela, nous comptons à une coopération avec les linguistes burkinabé.

Références

BAILLEUL, CH. (2007) *Dictionnaire Bambara-Français.* 3e édition corrigée. Bamako : Donniya.

DUMESTRE G. (2011). Dictionnaire bambara-français suivi d'un index abrégé français-bambara. Paris : Karthala

ROVENCHAK A., BUK S. (2013). Masadennin (The Little Prince in Bamana): Experimental online concordance with parallel French and English texts. *Mandenkan* 50, 117-130.

VYDRINE V. (1999). Manding-English Dictionary (Maninka, Bamana). Vol. 1. St. Petersburg: Dimitry Bulanin Publishing House.