Un vérificateur orthographique pour la langue bambara

Jean-Jacques Méric [1]
(1) INALCO, Étudiant, Département Afrique,
65 rue des Grands Moulins - CS21351 - 75214 PARIS cedex 13
jjmeric@free.fr

Résumé. Un vérificateur orthographique et thesaurus pour le bambara (bamanankan) réalisé à partir d'un dictionnaire électronique de la langue bambara, Bamadaba[1], adapté à un moteur de vérification standard, Hunspell[2], et donc disponible pour Libre Office, Open Office, Neo Office, Mozilla Firefox et Thunderbird, (sur leurs sites web respectifs) ainsi que pour Adobe Indesign et les autres logiciels intégrant Hunspell. L'adaptation a porté essentiellement sur les règles de flexion, dérivation et composition (langue agglutinante), ainsi que sur le dictionnaire des synonymes et variantes. La vérification ne tient pas compte des tons, ce qui poserait des problèmes d'acceptation. Une attention particulière est portée sur les propositions d'orthographe correcte pour les erreurs d'usage les plus fréquentes (forums et blogs). Projets pour les mobiles et pour les logiciels utilisant leurs propres moteurs, et pour l'écriture en N'ko.

Abstract. A spell-checker and thesaurus for the Bambara language (Bamanankan), based on an electronic version of a dictionary of Bambara, Bamadaba, ported to a standard spell-check engine, Hunspell, and made available (through their respective web sites) for Libre Office, Open Office, Neo Office, Mozilla Firefox and Thunderbird, as well as Adobe Indesign and other software using the Hunspell engine. The port concerns essentially rules of inflection, derivation and composition (Bamanankan being an agglutinative language), as well as the compilation of a thesaurus for synonyms and variants. Tones are not checked, mainly because they are currently not indicated in text published in Mali. A special attention is paid to the most frequent errors in current usage (forums and blogs) : appropriate suggestions have been made. Future work is to be done for mobile use and software using specific spell-check engines, and the N'ko script.

Mots-clés : Vérificateur, orthographe, bambara, Afrique
Keywords: Spell-checker, dictionary, Bambara, Africa

1 Le moment opportun

Environ 40 langues bénéficient du support d'un correcteur orthographique, parfois avec un luxe d'options : variétés régionales, support des césures, dictionnaire des synonymes, vérification grammaticale. Seules 3 langues africaines pouvaient jusqu'à présent prétendre intégrer ce groupe : le swahili (Afrique de l'Est), le shona (Zimbabwe), le malagasy (Madagascar). Le bambara (Afrique de l'Ouest) devient la quatrième.

Malgré l'échec de l'apprentissage scolaire du bambara, c'est une des langues mandingues le plus en expansion et dont le statut de "lingua franca" dans la région est le plus affirmé. La connaissance linguistique de cette langue a continué de progresser, notamment sous l'impulsion de la revue Mandenkan, et a atteint ces dernières années un nouveau palier avec la publication de dictionnaires importants (Bailleul, 2007 ; Dumestre, 2011) et de grammaires approfondies (Dumestre, 2003). Ce sont les travaux du professeur Vydrine et de toute une équipe qui ont ouvert une nouvelle étape en mettant à la disposition des chercheurs ces dictionnaires sous forme électronique, dictionnaires qui s'enrichissent à présent grâce à la mise en place de son Corpus bambara de référence et de son alimentation régulière : c'est l'étude de ce Corpus qui permet à présent un nouveau degré dans la compréhension du vocabulaire et de la grammaire bambara.

---

1 http://cormand.tge-adonis.fr/

2 http://en.wikipedia.org/wiki/Hunspell

C'est dans ce contexte de maturation, que l'idée d'un correcteur orthographique pour la langue bambara (Enguehard, Koné, 2010) a pu mûrir et aboutir à un outil utilisable, outil dont l'ambition est de faciliter la production de textes, qui viendront eux-mêmes à leur tour enrichir le Corpus.

2 Utiliser des outils standards et adaptés

HUNSPELL, créé à l'origine par Lazlo Nemeth, est un moteur de correcteur orthographique "libre" :

- standard, en particulier utilisant le standard Unicode/UTF-8. Il était également important de prendre en compte le fait que le bambara est une langue tonale : ces tons sont marqués par des diacritiques, ce qui nécessite aussi Unicode. Le projet initial n'est pas de diffuser un vérificateur d'orthographe tonal, ce qui poserait de gros problèmes d'acceptation, l'usage étant de ne pas les noter, mais la possibilité devait être préservée. Ainsi que la possibilité d'utiliser un autre alphabet que l'alphabet bambara latin : l'alphabet N'ko. Malgré les inévitables difficultés spécifiques, nous avons pu réaliser également des prototypes tout à fait fonctionnels de vérificateur tonal et de vérificateur en N'ko. Nous utilisons également en interne un vérificateur adapté à l'ancien alphabet bambara (avant 1983) ce qui nous aide à alimenter le Corpus en textes imprimés à l'époque.

- adapté à un aspect important du bambara : comme le hongrois (origine de Hunspell), le bambara fait la part belle à la composition, procédé très productif de création de noms et de verbes.

Hunspell a l'avantage d'être intégré à de nombreux outils :

- de traitement de texte : Open Office, Libre Office, Neo Office ;

- de navigation web : Firefox, Chrome,.. ;

- de messagerie : Thunderbird ;

- de mise en page : Adobe Indesign ;

- de ROC : Tesseract (utilisé pour l'alimentation du Corpus bambara de référence!) ;

… et sur les 3 plate-formes principales : Windows, Mac OsX, Linux, mais des adaptations existent également pour les tablettes et mobiles IOS et Androïd, vers lesquels nous portons notre attention.

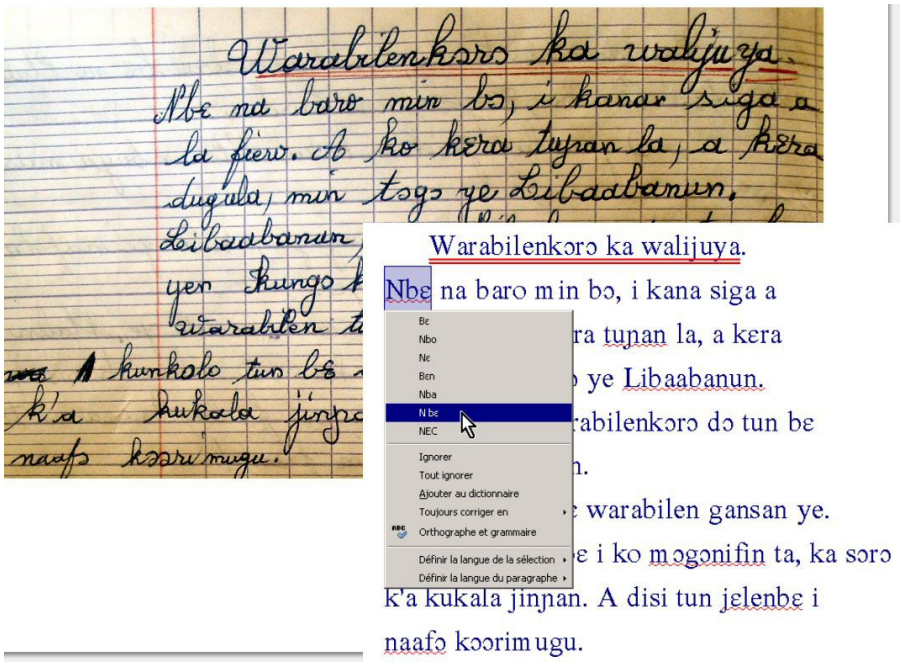

Figure 1: Manuscrit saisi sur ordinateur avec vérificateur orthograhique

J'ai repris un prototype d'Andrij Rovenchak (Université de Lviv, Ukraine) pour en faire un outil utilisable, et aujourd'hui distribué sur les sites internet des logiciels mentionnés plus haut.

3 Description de la solution

Il s'agit pour l'essentiel de fournir au moteur Hunspell deux fichiers :

- un dictionnaire, reformaté pour Hunspell ; ce dictionnaire est à l'origine la version électronique du dictionnaire de Charles Bailleul paru en 2007, sans cesse remanié et enrichi au fur et à mesure de l'enrichissement du Corpus de référence Bambara.

- un fichier des affixes : Dans ce fichier sont formalisées toutes les règles de flexion, dérivation et composition telles que décrites le plus précisément possible dans une grammaire de la langue : il est essentiel d'avoir une grammaire complète et cohérente.

Ces deux fichiers travaillent en tandem : chaque mot du dictionnaire fait référence à une ou plusieurs règles décrites dans le fichier des affixes : il s'agit des règles applicables à ce mot, **chaque règle est désignée par une lettre de l'alphabet**. Les mêmes groupes de règles sont applicables en général selon la nature des mots, par exemple : la marque du pluriel est compatible avec les noms ou les adjectifs, les suffixes aspectuels se combinent avec les verbes. Mais elles peuvent être individualisées pour des mots particuliers. Si un mot ne fait référence à aucune règle, c'est qu'il est réputé invariable et non susceptible d'être combiné dans une forme composée.

Illustration pour le mot *baarakɛlaw* 'travailleurs', qui, pour être accepté par le vérificateur orthographique, fait entrer en jeu deux règles de composition pour joindre *baara* 'travail' et *kɛ* 'faire', une règle de dérivation en nom d'agent pour *la* 'celui qui fait l'action', et une règle de pluriel pour la marque du pluriel *w*.

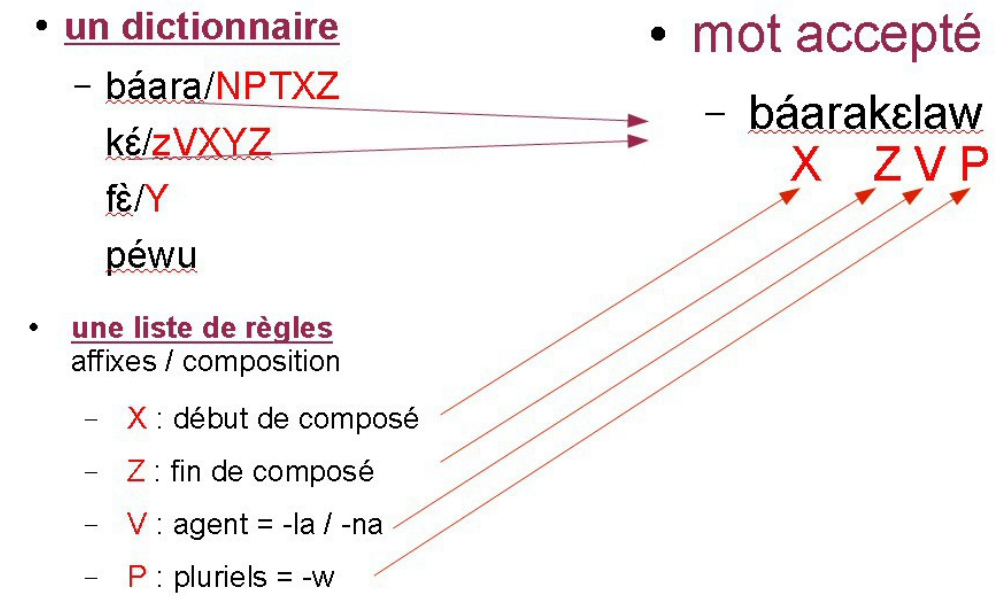

Figure 2 : Assemblage d'un mot composé en bambara

4 Les aspects pratiques.

Outre la technique, deux aspects méritent attention :

4.1 acceptabilité : le vérificateur orthographique ne doit pas trop faire sentir sa présence.

Le bambara n'est pas une langue figée : répandue sur un vaste territoire, elle y côtoie de nombreuses autres langues (17 langues nationales au Mali : peul, maninka, … et français), et elle bouge vite. Les créateurs des grands dictionnaires cités ont eu la sagesse de noter toutes les variantes rencontrées : voyelles longues, nasales et autres modifications. Nous avons conservé toutes ces variantes dans le dictionnaire fourni à Hunspell : leur utilisation est acceptée et n'est pas sanctionnée par un soulignement en rouge intempestif. Seul le dictionnaire des synonymes indique, à qui le consulte en cas d'hésitation, la forme considérée comme "canonique" à l'instant présent. Ce qui ne préjuge pas de ce qu'une étude future des pratiques, à l'aide du Corpus, définira comme celle la plus fréquente ou la plus justifiée. C'est déjà arrivé !

Les composés : ceux-ci obéissent à une série de modèles très limités qui contraignent la formation par ailleurs très libre de noms et de verbes composés. Notre inquiétude initiale était que celles-ci n'étaient qu'imparfaitement couvertes par les règles de composition offertes par Hunspell. Un certain "laxisme" existe donc dans la manière dont le vérificateur soumet les composés à ses inspections ; en pratique toutefois, il est important que le vérificateur ne soit pas trop intrusif : le gain en acceptabilité compense largement les efforts de maintenance et de corrections de bugs inhérents à des contrôles trop exhaustifs.

Le langage courant : la quasi absence d'éducation à l'orthographe du bambara à l'école a laissé le champ libre à des façons étranges d'écrire les mots les plus courants. Nous observons régulièrement les blogs et autres forums d'expression et vérifions quelles suggestions de correction est capable de faire le vérificateur orthographique : sont-elles pertinentes ? Si ce n'est pas le cas, nous enrichissons le vérificateur - actuellement quelques centaines de suggestions, par exemple : pour *mousso* : *muso*, pour *dokotoroso* : *dɔkɔtɔrɔso*.

L'environnement d'écriture du bambara doit de préférence être en bambara lui-même. S'il n'entre pas dans le cadre d'un projet de vérificateur de faire la "localization" des menus d'un traitement de texte (un autre de nos projets), nous en avons tenus comte dans le Dictionnaire des synonymes, où les mots de "synonymes", "variantes", et les termes de description grammaticale des mots (parties du discours), sont en bambara ; des exemples d'utilisation sont donnés.

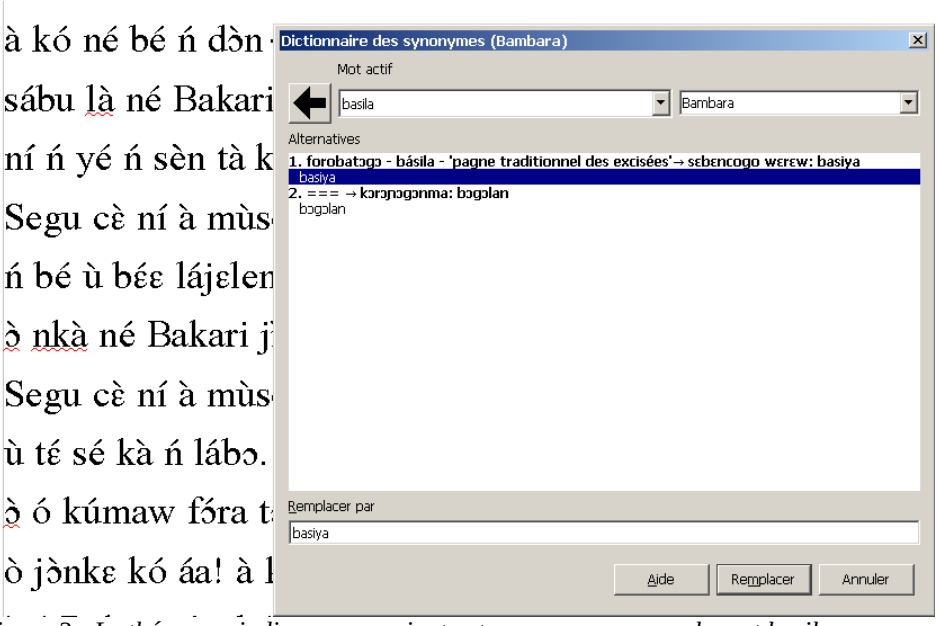

Figure 3 : Le thésaurus indique une variante et un synonyme pour le mot basila

En cela, préserver cette acceptabilité revient à mettre en pratique, au niveau de la production de texte, l'idéologie qui préside au Corpus de référence bambara : "fixer plutôt que normaliser", permettre "de représenter dans le Corpus la pratique langagière bambara telle qu'elle est en réalité." (Vydrine 2014)

4.2 Routine : la mise à jour du vérificateur orthographique doit être facile

Si nous nous plaçons à présent du côté de ceux qui fournissent le vérificateur, les efforts consentis initialement pour le prototype : laborieuses extractions, longues annotations et compilations manuelles... ne sont pas acceptables à long terme : le dictionnaire Bamadaba évolue chaque mois, et parfois plusieurs itérations, les versions du vérificateur doivent pouvoir suivre, sinon ce rythme, au moins chaque trimestre. Il est important d'automatiser : actuellement, un seul programme permet de générer en quelques minutes le dictionnaire nécessaire à toute nouvelle version de Bamadaba.

Pour l'utilisateur assidu, il doit être facile de passer à la nouvelle version. C'est le cas : supprimer l'ancienne version, ajouter la nouvelle, redémarrer l'application.

5 Utilisation en pratique et retours sur expérience

La diffusion publique du vérificateur orthographique, non plus à travers notre propre site spécialisé mais à travers les sites des éditeurs de logiciels (Apache Open Office, Libre Office, Mozilla...) a été accompagné par une série d'articles sur le blog malien fasokan.org, blog primé en 2012 (Best Of Blogs award). On ne peut pas encore parler de diffusion massive, mais il s'agit quand même de quelques centaines de téléchargements.

Cela a cependant permis un premier contact tout à fait excitant : Le projet international dokotoro.org a sollicité notre aide. Une équipe de plusieurs rédacteurs maliens travaille à la publication d'un manuel de médecine de campagne en bambara, et le problème qui se posait était d'assurer une homogénéité dans la qualité de bambara écrit ; un vérificateur leur a paru d'une aide précieuse. Il s'ensuit des échanges enrichissants sur le vocabulaire médical.

Nous mettons à profit également le vérificateur "en interne" dans le processus d'alimentation du Corpus : de nombreux textes, imprimés anciens ou manuscrits (illustration de la première page) sont saisis par des dactylos, ou scannés par reconnaissance optique de caractères (ROC ou, en anglais, OCR), le vérificateur permet d'améliorer le contrôle qualité de ces textes sous forme électronique, avant leur analyse et entrée dans le Corpus de référence Bambara ; c'est en fait une boucle qui se met en place : les textes analysés qui alimentent le Corpus permettent d'enrichir le dictionnaire, ce qui permet d'améliorer l'OCR et le vérificateur, etc.

Produire plus de textes en bambara : Ce projet de vérificateur s'ajoute à d'autres projets qui poussent dans le même sens, comme par exemple les claviers permettant la saisie des caractères de l'alphabet bambara[3].

Si l'intention était d'aider les rédacteurs potentiels à produire plus de textes en bambara, nous n'en sommes certes pas encore là, d'autant que d'autres obstacles existent : sur le plan culturel, la façon dont la langue est perçue ; sur le plan de l'éducation, la façon dont elle est enseignée ; enfin sur le plan des outils informatique, la faible diffusion des ordinateurs de l'utilisation d'internet, la préférence d'utilisation de Word, voire, dans les blogs et forums, l'utilisation de mobiles...

Nous restons toutefois portés par l'enthousiasme des quelques contacts que nous avons au Ministère de l'éducation et des langues nationales, qui en perçoivent au moins l'intérêt pédagogique[4], ce qui nous pousse à développer d'autres projets ludiques s'appuyant sur le Corpus, en particulier : un Scrabble en bambara[3], des mots croisés générés automatiquement. Et nous commençons à en percevoir les effets sur la qualité du bambara écrit sur le blog mentionné plus haut, qui est une des sources "bambara contemporain" du Corpus bambara de référence.

Enfin, de même que nous pensons appliquer à d'autres langues (malinké, Nk'o) l'expérience acquise avec le vérificateur bambara, nous sommes prêts à partager celle-ci avec quiconque voudrait se lancer dans cette aventure.

---

[3] http://www.mali-pense.net/Ressources-pour-la-pratique-du.html (Ressources pour la pratique du bambara écrit)

[4] intérêt documenté dans quelques rares études (apprenants 2ème langue, avec encadrement pédagogique 1ère langue)

Références

BAILLEUL C. (2007) *Dictionnaire bambara-français*. Bamako: Editions Donniya

DUMESTRE G. (2011) *Dictionnaire bambara-français*. Paris: Karthala

DUMESTRE G. (2003) *Grammaire fondamentale du bambara*. Paris: Karthala

DUMESTRE G. (2006) *Bamanankan Maben [grammaire du bambara, en bambara, rédigée avec un groupe de linguistes et pédagogues maliens]*. Bamako: Editions Donniya

ENGUEHARD C., KANÉ S. (2010) *Langues africaines et communication électronique : développement de correcteurs orthographiques*. LABORATOIRE D'INFORMATIQUE DE NANTES-ATLANTIQUE – NANTES- FRANCE/CENTRE NATIONAL DES RESSOURCES DE L'EDUCATION NON FORMELLE – BAMAKO – MALI

VYDRINE V. (2014) *Instructions pour le Corpus bambara*. Paris: non publié