Language in Africa 1(4), 2020, 27–42. doi: 10.37892/2686-8946-2020-1-4-27-42

ESSAIS D'ANALYSE SYNTAXIQUE AUTOMATIQUE DANS CORBAMA, LE CORPUS DE RÉFÉRENCE BAMBARA

Jean-Jacques Méric
CNRS-LLACAN
jjmeric@gmail.com

Résumé : En s'appuyant sur une boîte à outil bien conçue qui aide à entrer manuellement les éléments d'analyse syntactique d'un corpus textuel en bambara, l'auteur a fait une essai de programmation classique à base de règles. Ce n'est pas une approche conforme à l'état de l'art, mais elle s'est avéré faisable et permet d'avancer en attendant, et de faire face à l'énorme quantité de travail requis.

Mots-clés : bambara, corpus textuel, analyse syntactique

AN ATTEMPT AT AUTOMATIC SYNCTACTIC ANALYSIS ON CORBAMA, THE BAMANA REFERENCE CORPUS

Jean-Jacques Méric
CNRS-LLACAN
jjmeric@gmail.com

Abstract: Building on a well-developed set of tools to help manual editing of syntactic analysis of a Bamana text corpus, the author has made an attempt at rule-based classical programming to automate as much as possible the process. This is not a state-of-the-art approach but it has proved feasible and temporary fills a gap in the huge amount of work at hand.

Key words: Bamana, text corpus, syntactic analysis, parsing

Les travaux de Valentin Vydrine (voir, par exemple, Vydrin 2019a) ont abouti à la constitution d'une collection remarquable de textes écrits en bambara, à leur numérisation, à leur organisation et à leur mise à la disposition du public sur internet :

- une bibliothèque classique accessible dans les limites du respect du droit d'auteur, BambaraBiblio,

- une base de données interrogeable comme une sorte de « moteur de recherche du bambara », le Corpus bambara de référence, Corbama, disponible pour tous sur internet. Celle-ci sert également de base pour la constitution d'un dictionnaire du bambara publié sur le web, et aux études sur l'orthographe et la grammaire du bambara.

Cet article présente d'abord le corpus bambara ; il donne ensuite un aperçu pratique des outils qui permettent de constituer ce corpus de référence et de l'annoter manuellement ; il présente enfin l'un de ces outils d'analyse syntaxique, le désambiguïsateur automatique.

1. Le corpus bambara

1.1. La matière première

En plus des collections personnelles de Valentin Vydrine, les contributeurs ont été nombreux : la bibliothèque de Gérard Dumestre a été mise à contribution, certains élèves et collègues de l'INALCO et de Saint-Pétersbourg ont contribué, ainsi que les collègues d'Afrique de l'Ouest. Valentin Vydrine a systématiquement recueilli des ouvrages de toutes tailles, provenances, graphies : En effet l'écriture du bambara a évolué au cours des siècles : Adjami, écritures alphabétiques latines de l'époque coloniale, et après l'indépendance du Mali, réformes orthographiques de 1967 et de 1982. Les textes actuels du Corpus sont écrits dans les scripts suivants (par ordre chronologique d'arrivée) :

- ancienne orthographe : avec les lettres bambara è, ò, ŋ et ny (environ 2.200 textes),

- nouvelle orthographe 1 : avec les lettres bambara ɛ, ɔ, ŋ et ny (environ 50 textes),

- nouvelle orthographe 2 : avec les lettres bambara ɛ, ɔ, ŋ et ɲ (environ 15.300 textes),

- orthographe tonalisée, utilisant les tons hauts, bas, ascendante et descendants : ́ ̀ ̂ ̌ (environ 20 textes seulement).

Au final, le corpus est adaptable en ce qui concerne les scripts utilisés, la boîte à outils DABA contient tous les convertisseurs nécessaires, comme cela a été prouvé pour l'écriture N'ko (important pour le corpus maninka). Et même un premier texte bambara écrit en adjami a été ajouté au Corpus.

Le corpus bambara, le corpus maninka, et d'autres, ainsi que les outils associés, peuvent être consultés à partir du portail des langues mandé, Corpora Mandeica.[1]

1.1.1. La numérisation

La numérisation des sources écrites en bambara numérisation (« scan ») a été un premier travail d'importance.

Leur transformation en textes utilisables sur ordinateur en a été un autre, de taille. Pour l'essentiel, chaque texte a été saisi par deux dactylos professionnelles. Une tierce personne en a fait la collation, travail qui consiste à isoler les différences entre les deux saisies et à constituer un texte final épuré des erreurs éventuelles de l'une ou de l'autre saisie.

La reconnaissance optique de caractère (OCR) a été mise au point :

1) elle a été adaptée à la langue car elle s'appuie sur un jeu de caractères reconnu et sur des dictionnaires : ce travail a été fait pour le bambara (nouvelle et ancienne orthographe) et fonctionne correctement.

2) malgré les progrès récents, elle ne fonctionne correctement que sur des documents papiers et des numérisations de bonne qualité ; en pratique, c'est rare.

De ce fait la reconnaissance optique ne peut être utilisée que par un opérateur assez spécialisé.

---

[1] http://cormand.huma-num.fr/mandeica

Les originaux sous forme électronique sont rares et ont pâti des acrobaties différentes employées par les unes et les autres dans le passé pour représenter les quatre « caractères bambara » spécifiques (huit avec leurs majuscules). Avec l'emploi progressif d'Unicode, jeu de caractères standard universel qui inclut les caractères bambara, on commence à recevoir des textes employables directement, mais la tendance est encore timide.

1.1.2. Le volume

C'est ainsi qu'environ 17.600 publications ont été numérisées représentant un volume de 11,3 millions de mots. Ce volume est certes modeste comparé aux volumes des corpus des grands langues occidentales, mais il a tout à fait la masse critique nécessaire pour un Corpus utilisable sur le plan linguistique. Pour ce qui concerne l'Afrique, il est le seul, avec le swahili.

Ce chiffre de presque 12 millions de mots semble actuellement un plafond difficilement dépassable : en effet, il y a actuellement peu de publications nouvelles, à part les quelques revues à parution régulière (*Kibaru*, *Jɛkabaara*).

Aujourd'hui pourtant, on pourrait espérer un renouveau dans l'écriture et la création littéraire en langue bambara, tant on dispose enfin d'outils pour les créateurs qui sont tout à fait à la hauteur des outils disponibles dans d'autres langues : sont parus depuis 2007 les dictionnaires de Gérard Dumestre (2011) et de Charles Bailleul (2007), le dictionnaire en-ligne Bamadaba, ouvrages de grammaire des mêmes (Dumestre 2003; Bailleul 2005) et de Valentin Vydrine (2019b), et même depuis quelques années maintenant, un vérificateur orthographique dérivé de Bamadaba, intégré au traitement de texte, comprenant un thésaurus (synonymes du mot, dictionnaire résumé sous un simple clic).

Il y a aussi un potentiel important du côté des archives sonores non encore transcrites : enregistrements de griots et de pièces de théâtre (*Koteba*), cassettes d'émissions radiophoniques ou télévisées, inépuisable

création de chansons de la mondialement célèbre musique malienne. Il existe un cadre pour ce travail : le Corpus sonore, aujourd'hui embryonnaire, mais peu à peu enrichi grâce à la plateforme logicielle Elan-Corp (http://llacan.vjf.cnrs.fr/res_ELAN-CorpA.php).

Et parler seulement de volume du Corpus ne rend pas justice à la variété des genres littéraires (journaux, épopées, contes, proverbes, devinettes, chansons, coran, ancien testament etc.), des thèmes et à la diversité de plus de 2.600 auteurs.

1.2. Le classement des textes

Tous ces textes sont classés par genre littéraire (information, contes, poésie, etc.), thèmes (politique, agriculture, santé, etc.) et par auteur (lien avec un dictionnaire des auteurs). Ce classement est matérialisé par un entête dit « métadonnées » qui vient s'ajouter au texte lui-même. Ce travail de classement est facilité par l'un des outils de la boîte à outils informatique « DABA » développée par Kirill Maslinsky (voir Maslinsky 2019) et librement disponible sur internet.[2] Les recherches dans le Corpus de référence CORBAMA, peuvent être limitées par auteur, thème ou genre littéraire.

Outre ce classement classique, le Corpus est subdivisé selon la qualité des textes. L'ensemble s'appelle le **« CORBAMA-BRUT »**, celui-ci comprend 17.600 textes soit 11,3 millions de mots :

1) des textes pré-analysés syntaxiquement à 100 % (désambiguïsés par un bon connaisseur du bambara) :

1.1) dont l'original est écrit en bambara tonal : **« CORBAMA-NET-TONAL »** (une minorité) : 20 textes, 35.000 mots ;

1.2) dont l'original est écrit bambara non-tonal, soit la grande majorité des textes existant en bambara : **« CORBAMA-NET-NON-TONAL »** : 1.743 textes pour un peu plus de 1,5 millions de mots.

Parmi ces textes, certains possèdent des traductions en français qui ont été synchronisées avec les phrases originales en bambara, alimentant

[2] https://github.com/maslinych/daba

le « corpus parallèle » : celui-ci permet des interrogations en bambara dans « CORBAMAFARA », ou en français dans « CORFARABAMA », et met les résultats des recherches en vis-à-vis. Cette partie est pour l'instant embryonnaire : 149 textes, 225.000 mots. Un énorme travail fait par Andrij Rovenchak (voir Rovenchak 2003), avec l'aide de Kirill Maslinsky pour la technique ;

2) des textes analysés automatiquement (désambiguïsés à 75%). Cette dernière partie comprend des textes vérifiés et classés (méta-données), mais également depuis peu des textes non encore collationnés ni classés : le « précorpus ». Au total, 15.800 textes pour 9,8 million de mots.

1.3. La structure de la langue dans les textes

Les textes sont constitués en deux parties : La phrase originale, puis sa décomposition mot par mot. Chaque mot, tel qu'il est écrit par l'auteur, « word », est identifié de façon unique par rapport à son entrée dans le dictionnaire : il pourrait être identifié par son numéro unique dans le dictionnaire. Il a été choisi de l'identifier par le triplet qui le définit de façon unique dans le dictionnaire : 1) son orthographe exacte, normalisée, le « lemma » ; 2) son rôle syntaxique « ps » (*part of speech*) : nom, verbe, auxiliaire, adjectif, conjonction, etc. ; 3) son sémantisme, identifié par son sens de base, « gloss ». Le cas échéant, on y ajoute l'identification des suffixes (plus généralement affixes) quand il est dérivé (pluriels, conjugaison etc.), et l'identification complète des mots qui le composent, s'il s'agit d'un mot composé.

On a donc une représentation de ce type, ou « annotation » : *kuma* : *kúma*:n:parole.

Mais décrire cette structure plus en détail n'entre pas dans le cadre de cet article ; en fait les représentations sont diverses selon les outils informatiques qui sont censés les manipuler et les représenter pour les chercheurs : avec des formules à parenthèse, avec des balises de type XML, sous forme de fichier verticaux, etc.

2. Les outils du Corpus bambara

2.1. Le dictionnaire Bamadaba

Le dictionnaire a été construit avec Toolbox, de la SIL, un outil classique pour les linguistes de terrain. C'est un outil extrêmement souple qui permet de construire rapidement un dictionnaire avec les « marqueurs » de base : lemma, partie du discours, glose. Et il est facile d'y ajouter ses propres « marqueurs » ad-hoc.

Construit sur une base très riche, le dictionnaire de Charles Bailleul (2007), Bamadaba s'enrichit de mots nouveaux au fur et à mesure de l'étude du Corpus : ces mots nouveaux sont soit des mots déjà connus mais répertoriés seulement dans le dictionnaire de Gérard Dumestre (2011), ou bien dans des dictionnaires des langues Mandé non publiés (Valentin Vydrine), soit des mots totalement inconnus ; dans ce dernier cas, une liste de discussion, Mandelang, permet d'interroger chercheurs et locuteurs natifs : orthographe exacte, tons, différents sens.

Une version web de ce dictionnaire est publiée périodiquement, en général synchronisée avec la publication d'une nouvelle version du Corpus[3]. Cette publication permet d'entrevoir l'intention à long terme de ce dictionnaire : un dictionnaire piloté par le corpus. Par exemple, on peut dès à présent obtenir des exemples d'emploi directement dans le corpus, à l'aide d'un lien fourni à cet effet ; et pas seulement comme autrefois à l'aide d'exemples soigneusement choisis par l'éditeur du dictionnaire.

2.2. Le parseur

Terme du jargon informatique, « parser » en anglais ; il s'agit d'un programme qui analyse un texte à son entrée dans le corpus. Il extrait de chaque phrase tout mot et ponctuation. Il tente de nettoyer chaque mot de ses dérivations possibles (pluriels, conjugaisons, etc.), puis le confronte au dictionnaire. Le mot est alors identifié, soit de façon

[3] http://cormand.huma-num.fr/Bamadaba/lexicon/index.htm

unique : il est alors désambiguïsé, soit de façon multiple : il reste alors à désambiguïser :

kumaw : le nom *kúma* ‘parole’ suivi de la marque du pluriel *-w* : 1 possibilité ;

bɛ : l’auxiliaire marque de l’imperfectif affirmatif (IPFV.AFF), ou bien la copule marque d’existence (ÊTRE) : 2 possibilités.

Le réglage du parseur est particulièrement délicat ; mal paramétré, il peut créer plus d’ambiguïté qu’il n’est nécessaire, en particulier avec une langue qui utilise beaucoup les mots composés. C’est ainsi que le mot très familier aux bambara, *nakun* : ‘la raison’ (*kùn*) de ‘la venue’ (*nà*) pourrait être renvoyé par le parseur comme pouvant signifier également ‘la tête’ (*kùn*) de ‘la sauce’ (*ná*), ‘la mère’ (*ná*) qui ‘s’adapte’ (*kù*) à ‘moi’ (*ń*)... soit une bonne douzaine d’explications parfois farfelues.

Il faut donc éviter que le parseur considère toutes les possibilités avant de retourner le résultat de ses recherches. Il doit travailler par étapes, considérer d’abord les dérivations les plus fréquentes, et s’arrêter là si un résultat est plausible, avant de considérer les dérivations moyennement fréquentes, puis les plus rares ; dans le temps suivant il va s’attacher aux compositions de mots les plus typiques (nom+nom) (nom+verbe) (redoublement), avant de considérer, seulement si nécessaire, les compositions plus complexes (nom+verbe+nom), etc. Heureusement, la grammaire bambara est très bien balisée avec les travaux de Valentin Vydrine (2019b), Gérard Dumestre (2003) et Charles Bailleul (2005), et il n’est pas trop difficile de la formaliser pour le parseur.

Au terme de cette analyse, le parseur a identifié de façon unique environ 25 % du texte, le reste étant composé soit de mots «ambigus» (plusieurs possibilités) soit de mots inconnus : pas de correspondance trouvée dans le dictionnaire, pour des raisons variées qui vont de la faute de frappe dans l’original en passant par l’utilisation d’un nom de lieu ou d’un nom propre inconnu (la cause la plus fréquente, en particulier des noms étrangers), jusqu’à, plus rarement, l’apparition d’un nouveau mot à considérer pour le dictionnaire.

2.3. La désambiguïsation

C'est là que doit intervenir un humain, un « esclave de la désambiguïsation ». Car c'est un travail délicat mais fastidieux que de clarifier l'analyse syntaxique d'un texte de plusieurs centaines de phrases. À l'aide du logiciel GDISAMB, il faut confirmer (et plus rarement, corriger) le travail du parseur, choisir parmi les possibilités listées celle qui convient le mieux. Ce travail demande une bonne compréhension du contexte sémantique, assez fréquemment, il s'agit de décortiquer un mot composé dont la structure aura échappé au parseur, et éventuellement, de noter à part les questions sur lesquelles on bute pour les soumettre à Valentin Vydrine et au groupe de discussion.

Peu nombreux sont les étudiants qui se portent candidats pour ce genre de travail. C'est dommage car son aspect répétitif présente un avantage de taille : vous êtes assurés de voir en peu de temps se graver dans votre cerveau les structures de la phrase bambara. L'effort d'une dizaine de désambiguïsations est largement récompensé par une accélération d'apprentissage extraordinaire.

Au terme de ce travail, on aboutit à un texte complètement analysé, ce qui est une ressource essentielle pour qui compterait s'attaquer à la traduction, dans n'importe quelle langue, du texte original en bambara. Et on constitue un formidable outil de recherche linguistique ; toutefois, il faut là aussi atteindre une masse critique de textes ; on est probablement près du but avec la perspective d'atteindre 2 millions de mots dans quelques mois.

L'outil de travail lui-même, GDISAMB, fait partie de la famille d'outils « DABA » préparés par Kirill Maslinsky, comme le parseur. Il est très pratique, confortable, il est facile de retrouver un mot ou une partie de phrase, de revenir sur une désambiguïsation afin de la corriger.

Pour qui s'engage dans ce travail, il n'est pas inutile de bien s'équiper : dictionnaires sous la main, dictionnaire en ligne ou Toolbox sous le clic, et enfin un carnet électronique afin d'y noter les structures de mots complexes, afin de pouvoir faire des copiés-collés (en effet, sur un texte spécialisé, médical par exemple, certains mots reviennent souvent).

3. La désambiguïsation automatique

L'idée est venue progressivement, à force de faire des copiés-collés entre le carnet de notes et GDISAMB : il est certainement possible d'automatiser ce travail répétitif, et l'étendre, des mots isolés à des fragments de phrases.

3.1. Intelligence artificielle

Comme pour la traduction, si l'on veut automatiser ce travail d'analyse syntaxique, c'est vers l'intelligence artificielle (AI) que l'on se tourne aujourd'hui et c'est très probablement l'approche adaptée à ce genre de travail. Damien Nouvel (ERTIM-INALCO) nous a guidé sur cette voie avec le projet MANTAL (Nouvel 2019). Les premiers résultats ont été fort encourageants, avec un taux de désambiguïsation proche de 80 % (d'après ses propres mesures). Toutefois l'AI doit faire son apprentissage sur une base de texte suffisante, et si elle doit comparer les résultats de son intelligence artificielle à quelque-chose, autant qu'elle les compare à une désambiguïsation humaine fiable, avant de continuer à progresser. Notre première expérience dans ce domaine s'est d'abord heurtée à ce problème. Outre la masse critique un peu juste, nous avons constaté de nombreuses erreurs dans les textes désambiguïsés ; il nous a fallu passer un temps considérable pour les détecter systématiquement et les corriger. Cela n'aurait pas été possible sans la puissance des outils d'interrogation du Corpus et sans l'assistance de programmes pour corriger des centaines de textes.

Cette expérience et celle de la désambiguïsation manuelle ont entr'ouvert l'idée d'une solution provisoire, une désambiguïsation automatisée, un bricolage qui n'est pas véritablement de l'intelligence artificielle et n'est pas basée sur un des « moteurs » d'AI connus.

3.2. REPL

L'idée saugrenue de la liste de mots à copier-coller est une base minimale, très insuffisante en elle-même : Tout au plus un dictionnaire

complémentaire, adaptée à des textes spécialisés, ou pour combler les limitations, parfois frustrantes, du parseur.

Plusieurs idées plus précises ont permis une vraie avancée : Certes, un grand nombre de mots sont ambigus, la moindre ambiguïté étant : est-ce un nom, ou bien est-ce un verbe (*kuma* : 'la parole ; parler') ? La première idée est que, aussi ambigus que puissent être un grand nombre mots bambara, leur contexte permet immédiatement de les distinguer.

Qu'entend-on par contexte ? On entend d'abord par là leur proximité dans la phrase : *kuma* entre un auxiliaire verbal et une fin de phrase sera forcément le verbe 'parler' : *a bɛ kuma* {3SG AUX V} 'il parle'. On entend par là aussi, mais cela nécessite que plusieurs mots soient bien identifiés, leur rôle dans la phrase, leur place syntaxique. Il y a bien un autre type de contexte qui entre en jeu. C'est le contexte du sens (sémantique) : le texte parle-t-il des élections ? de la culture du mil ? Cette idée n'est pas hors de portée, notamment parce que tous les textes du Corpus ont une identification claire des métadonnées ; mais elle est complexe à manier, et les textes à propos d'élections au Mali ne manquent pas d'évoquer la culture du mil ou du coton. Nous écartons pour l'instant cette possibilité d'utiliser un marquage de domaine sémantique.

On va donc s'intéresser non seulement aux mots isolés, mais aux mots souvent associés ensemble. De nombreux couples de mots permettent de lever l'ambiguïté sur l'un des deux, et même parfois sur les deux à la fois : *fini foro* sera certainement 'le champ de fonio' alors que *fini* seul est ambigu ('tissu ; fonio') et *foro* seul également ('champ ; atteindre'). On peut certes imaginer une phrase où il s'agit d''atteindre' le 'tissu' mais le sens semble improbable, il est vraisemblable qu'un locuteur natif tournera la phrase autrement pour éviter la confusion possible ; mais surtout, il n'y en a aucune occurrence dans le Corpus.

On s'intéressera donc aux groupes de mots dont la fréquence dans le Corpus est attestée, et de préférence élevée. Et c'est ici un point où l'on se rapproche certainement de la façon un peu opaque dont l'intelligence artificielle constitue son auto-apprentissage.

Notre fichier de règles de désambiguïsation automatique va donc se compléter : mots isolés, couples de mots fréquents, groupes

nominaux, qualificatifs, possessifs, distributifs, circonstanciels (souvent aidés par les ponctuations), etc.

Toutefois, on s'épuiserait vite à ne considérer que ces assemblages de mots particuliers. Par exemple tous ces mots qui, en bambara, peuvent être à la fois des noms et des verbes ne pourraient tous être listés dans leurs différents contextes. Il nous faut des formules plus générales. On ajoute donc de nouveaux mots abstraits, ou « objets », qui représentent des familles de mots : « NV » va représenter tous les mots pour lesquels le parseur a envoyé deux possibilités : N : nom *ou bien* V : verbe. « PONCT » va représenter toute ponctuation. Beaucoup de mots en bambara peuvent être à la fois un nom et un verbe. La proximité d'un auxiliaire verbal dans un contexte simple permet d'identifier que l'on a bien affaire au verbe. On peut donc écrire une règle comme :
bɛ NV PONCT = *bɛ*:pm:IPFV.AFF NVverbe PONCT

Soit, en clair : Si l'on trouve dans une phrase le mot *bɛ*, suivi d'un mot qui peut être un nom ou un verbe, suivi d'une ponctuation, alors le remplacer par l'auxiliaire verbal *bɛ*, celui de l'imperfectif affirmatif, suivi du verbe proposé, suivi de la ponctuation en question.

Avec une dizaine de règles de ce type on règle 95 % des ambiguïtés Nom/Verbe.

Plus on avance dans la liste des règles, plus de mots sont identifiés, et plus on peut faire usage de formules qui s'appuient sur les quelques structures fixes et bien identifiées de la phrase bambara, comme celle-ci :
sujet
— auxiliaire
— complément-d'objet
— verbe
— complément-circonstanciel
— postposition
— adverbe.

Supposons que l'on utilise les objets suivants : NOM représente tous les noms (non ambigus), VERBE tous les verbes, POSTP toutes

les postpositions, on peut identifier beaucoup de *yé*, mot très ambigu en bambara (et qui reste une des principales difficultés !) avec une formule comme :
NOM *ye* NOM VERBE NOM POSTP = NOM *ye*:pm:PFV.TR NOM VERBE NOM POSTP

Soit, en clair : si l'on trouve *ye* dans le contexte décrit, alors ce *ye* est l'auxiliaire du perfectif (PFV.TR).

Ce exemple est bien entendu d'un usage restreint car beaucoup de noms sont accompagnés : adjectifs, déterminants, groupes possessifs, etc. ce qui risque d'être très long si l'on veut tenir compte de toutes les combinaisons possibles ! Donc en pratique on a fini par introduire un objet plus puissant qui est le groupe nominal, GN. La formule devient plus générale :
GN *ye* GN VERBE GN POSTP = GN *ye*:pm:PFVTR GN VERBE GN POSTP

La liste des règles de remplacement (REPL) compte environ 30.000 règles : les 2/3 sont des mots et combinaisons de mots, le 1/3 restant sont des formules avec des objets comme NOM, VERBE, NV, POSTP, GN.

Ces règles sont écrites dans un langage facile à comprendre pour un désambiguïsateur. Mais le vrai défi lorsqu'on travaille avec un fichier de 30.000 lignes, c'est l'organisation du fichier : où et dans quel ordre placer telle ou telle nouvelle règle ? Les améliorations ne sont donc pas seulement dans le nombre de règles, elles reposent énormément sur cette organisation, toujours perfectible.

Sur le plan informatique, ces règles sont traduites en « expressions régulières ». Le traitement du fichier prend de une à plusieurs minutes selon la taille du texte. Le traitement de l'ensemble du corpus prend de 1 à 2 jours.

3.3. Les résultats

Les résultats ont tout de suite été très encourageants, certes moins excitants ceux que l'AI mais consistants : travailler avec des textes où

60 % du travail de désambiguïsation a déjà été fait automatiquement s'est tout de suite avéré plus confortable et plus rapide. Toutefois il a fallu beaucoup de travail, et l'introduction de nombreux nouveaux objets, pour atteindre 70 %. Au point actuel, on en est à 75 % de désambiguïsation automatique (soit 25 % fait par le parseur et 50 % fait par REPL).

Comme une grand part du travail récent de désambiguïsation manuelle a été faite sur le journal *Kibaru*, c'est sur ces articles que les résultats sont les plus satisfaisants, montant à parfois plus de 90 %. Désambiguïser ces textes est donc beaucoup plus rapide ; mais ce n'est pas tout : la désambiguïsation manuelle restante permet d'identifier plus facilement les difficultés résiduelles et les progrès qui restent à accomplir, mais aussi parfois les erreurs à corriger, les règles à rendre plus précises, ou plus génériques.

3.4. Les bénéfices

Cette approche est donc d'ores et déjà une aide pour les désambiguïsateurs. Elle permet aussi de rendre le corpus « CORBAMA-BRUT » moins « brut » et permet de faire des recherches linguistiques pertinentes sur un corpus plus étendu que le seul « CORBAMA-NET » : 9 millions de mots élucidés au lieu de moins de 2 millions. On peut même envisager que le dictionnaire renvoie non plus aux exemples tirés directement de « CORBAMA-NET » mais à 4 fois plus d'exemples dans « CORBAMA-BRUT ».

4. Les perspectives

Il est peu vraisemblable que cette approche permette d'atteindre 90 % pour l'ensemble des textes, elle plafonnera vraisemblablement à 80 %. D'ici là, la quantité de nouveaux textes désambiguïsés à 100 % aura vraisemblablement atteint ou dépassé les 2 millions de mots. Cet ensemble formera une base plus consistante et plus saine pour l'apprentissage d'un système à base d'intelligence artificielle, qui est

la véritable perspective : un tel système pourra par exemple contourner les difficultés d'analyse liées à un mot mal orthographié, comme le ferait un lecteur, plus facilement que le programme actuel.

On ne saurait trop souligner ici que tout ceci n'aurait été possible sans le colossal travail d'étude réalisé par les professeurs et chercheurs qui nous ont précédés, les dictionnaires qu'ils ont publiés, les approfondissements de la grammaire dans tous ces aspects, les articles publiés dans les revues scientifiques, les cours dispensés à l'université, sans compter la constitution d'un bibliothèque unique de textes écrits en bambara. Sans ces bases solides, ces outillages : corpus CORBAMA, DABA, REPL,... n'auraient pas même pu commencer à être construits, il eût été inutile de les envisager, quand bien même l'annonce solennelle d'un appui politique et d'un financement aurait été faite.

Références

Bailleul, Charles. 2005. *Cours pratique de bambara.* Bamako: Editions Donniya.

Bailleul, Charles. 2007. *Dictionnaire bambara-français.* Bamako: Editions Donniya.

Dumestre, Gérard. 2003. *Grammaire fondamentale du bambara.* Paris: Karthala.

Dumestre, Gérard. 2011. *Dictionnaire bambara-français.* Paris: Karthala.

Maslinsky, Kirill. 2019. Positional skipgrams for Bambara: a resource for corpus-based studies. *Mandenkan* 62. 165–183.

Nouvel, Damien. 2019. *Traiter et désambiguiser les langues – Reconnaissances & résolutions.* Présentation au Colloque Digital Armenian.

Rovenchak, Andriy. 2013. Masadennin (The Little Prince in Bamana): Experimental online concordance with parallel French and English texts. *Mandenkan* 50. 117–130.

Vydrin, Valentin. 2019a. Vers une lexicographie mandingue sur la base de grands corpus annotés. *Mandenkan* 63. 89–110.

Vydrin, Valentin. 2019b. *Cours de grammaire bambara.* Paris: Presses de l'Inalco.

Ressources électroniques

Bamadaba : http://cormand.huma-num.fr/Bamadaba/lexicon/index.htm
BambaraBiblio : http://cormand.huma-num.fr/biblio/index.jsp
Corpora Mandeica : http://cormand.huma-num.fr/mandeica/
Corbama : http://cormand.huma-num.fr/
Daba : https://github.com/maslinych/daba
Revue Mandenkan : http://mandenkan.revues.org
Toolbox (Field linguist's toolbox) : https://software.sil.org/toolbox/

Received 03.06.2020. Received in revised form 31.07.2020. Accepted 24.11.2020