Le projet RefLex [*Reference Lexicon*]

projet soumis à l'ANR - novembre 2008 - G. Segerer - LLACAN

Objectifs et méthodes

Ce projet a pour objectif la standardisation, la mutualisation et l'exploitation scientifique des données lexicales des langues d'Afrique.

1. Enjeux

Plusieurs équipes de recherche françaises et étrangères ont constitué des corpus lexicaux sur les langues d'Afrique. Certains chercheurs de ces équipes ont déjà commencé soit à mettre en ligne (cf. CBOLD, BLR3, IDS), soit au moins à élaborer des corpus numériques. Le problème est que ces ressources sont difficilement exploitables en l'état dans le cadre d'études comparatives ou typologiques, du moins autrement que par leurs auteurs, car

- elles sont non compatibles donc non mutualisables
- elles ne respectent pas toujours les sources originales
- elles utilisent des systèmes de transcription variés.

Les données lexicales sont très éparpillées, hétérogènes et difficiles d'accès, y compris pour la communauté scientifique. De ce fait, les travaux qui ont pour base des corpus complexes (c'est-à-dire construits à partir de plusieurs langues) ne peuvent pas faire l'objet de vérifications, sauf à dépenser un temps considérable à la recherche des sources originales. La validité scientifique des travaux comparatistes en est donc fragilisée.
Par ailleurs l'émergence de la linguistique quantitative, rendue possible par les progrès de l'informatique, nécessite des données nombreuses et fiables.

La solution est de mettre à la disposition de la communauté un **CORPUS LEXICAL DE REFERENCE** pour les langues africaines. Cette démarche nécessite un consensus de plusieurs équipes, en France, mais aussi au niveau international. La constitution d'un noyau de quelques équipes phares sera le point de départ d'une dynamique plus vaste.

Le LLACAN a élaboré un prototype de base de données comportant 30 lexiques pour 25 langues Niger-Congo soit environ 30.000 entrées. Cette base de données est prête pour la mise en ligne et peut constituer une maquette pour le projet.

2. Objectifs

Les objectifs scientifiques sont multiples :

- établir une méthodologie de partage et d'unification des données
- permettre des avancées significatives dans les études comparatives des langues africaines

- faire émerger de nouvelles problématiques liées à la profusion d'informations lexicales standardisées

Ce CORPUS DE REFERENCE intéressera en premier lieu les comparatistes, mais permettra également de mettre au point des méthodes de recherche novatrices dans des domaines aussi variés que la statistique phonologique ou la typologie sémantique.

A terme, notre but est d'impliquer un grand nombre d'équipes de recherches européennes travaillant sur les langues africaines. Il s'agit de parvenir à une mutualisation maximale des ressources disponibles. Dans un premier temps, on procèdera à l'inventaire de ces ressources. Celles-ci seront ensuite collectées, formellement harmonisées et enfin mises à la disposition de la communauté.
Dans les 2 premières années, l'objectif est de mettre en ligne au moins 500000 fiches lexicales indexées.

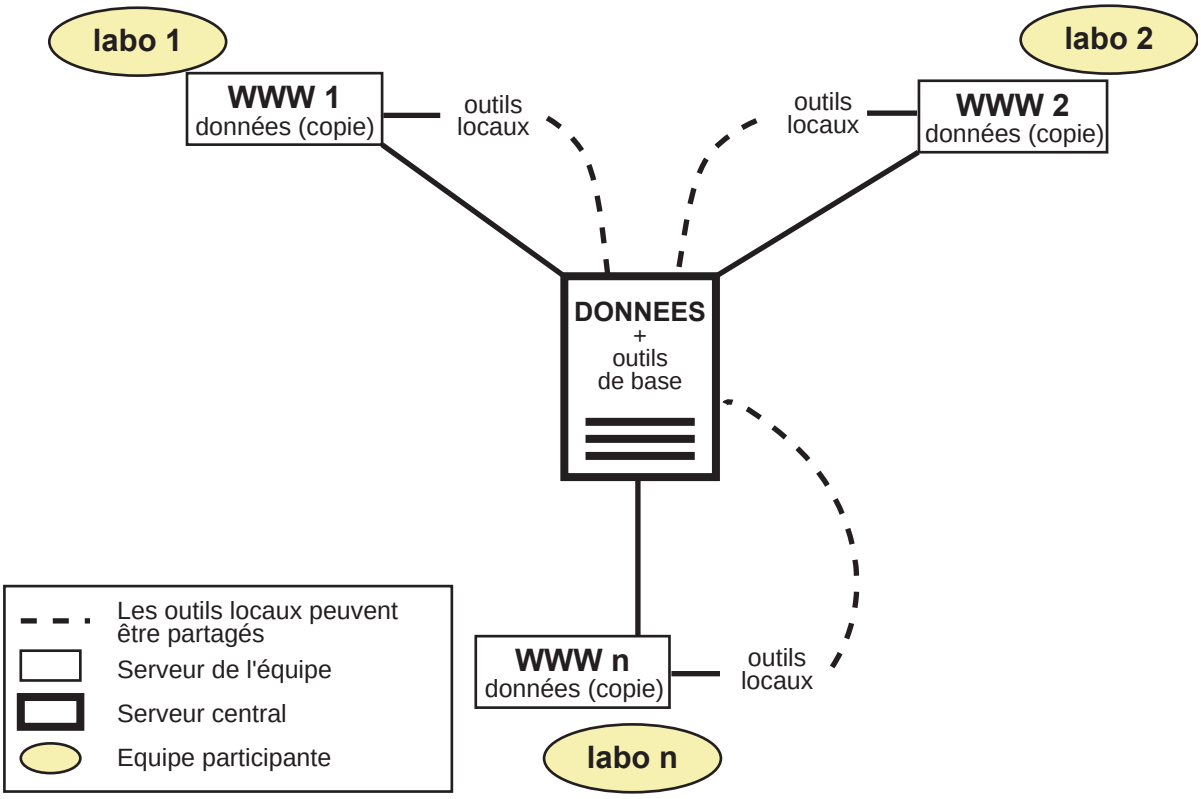
3. Méthodes

La mise en ligne de données lexicales devra obéir à deux impératifs catégoriques :

- présenter sans l'altérer la source originale
- permettre des manipulations (calculs, tris, recherches) sur tous les champs affichés.

En dehors de ces exigences, la plus grande souplesse est souhaitée. Un lexique pourra être mis en ligne sans attendre que tous les champs soient renseignés, le minimum étant la saisie des champs présents dans la source d'origine. La souplesse est ainsi double : le corpus peut être enrichi verticalement (ajout de langues, ou de données pour une langue) et horizontalement (ajout d'informations pour les données déjà présentes : schème tonal, squelettes consonantique et vocalique, structure syllabique, etc.). Ce principe permettra une réelle mutualisation des efforts, puisque chaque membre du projet pourra participer à l'enrichissement du corpus.

La mutualisation des ressources doit également suivre des principes généraux de souplesse et de modularité. D'une part, les données proprement dites (*data layer*) seront séparées des applications développées en vue de leur exploitation (*software layer*). La grande différence entre ces deux couches est la suivante : alors que les données structurées doivent demeurer strictement identiques pour tous (c'est précisément le principe d'un corpus de référence), les applications, au contraire, doivent être spécifiques, ce qui n'empêche pas leur mutualisation le cas échéant. Cette mutualisation sera rendue possible grâce au partage des spécifications techniques de la structuration des données (inventaire des champs et tables, types de données). Le schéma ci-dessous permet de se faire une idée de la façon dont nous envisageons de fonctionner :



4. Outils et ressources

Le LLACAN dispose déjà d'une expertise dans le domaine des bases de données lexicales, notamment par son expérience passée avec le logiciel Mariama. Robert Nicolaï, fondateur du projet Mariama, a d'ailleurs choisi le LLACAN pour être dépositaire du fond Sahelia qu'il a constitué dans ce cadre (400.000 entrées lexicales). Le LLACAN dispose en outre d'environ 150.000 entrées propres. À Lyon est hébergé le projet CBOLD, et nous avons l'accord de ses principaux responsables (F. Pellegrino pour DDL, Larry Hyman, Derek Nurse) pour récupérer ces données (450.000 entrées lexicales environ). En rassemblant ces données déjà saisies, on atteint virtuellement **un million de fiches lexicales**.

Par ailleurs, Guillaume Segerer a commencé à développer des programmes satellites de traitement autour de la maquette de base de données lexicales du LLACAN. Ces modules sont pour l'instant centrés autour de l'analyse phonologique : statistiques, recherche de paires minimales, alignement automatique de formes proches, recherche et gestion de correspondances phonétiques. Il est déjà envisagé de développer d'autres outils, par exemple pour l'analyse morphologique ou la sémantique lexicale.

5. Retombées du projet

Les retombées attendues sont de plusieurs types. D'une part, toutes les équipes concernées constitueront un réseau à l'échelle européenne qui pourra développer d'autres projets coopératifs. A cette occasion, il est permis d'espérer qu'un véritable standard de structuration des données lexicales pourra voir le jour. D'autre part, les possibilités ouvertes par ce vaste corpus permettront des avancées significatives dans les disciplines où la comparaison joue un rôle important : typologie, classification, reconstruction.

Le projet RefLex [*Reference Lexicon*]

projet soumis à l'ANR - novembre 2008 - G. Segerer - LLACAN

Description détaillée

I. L'ELABORATION DE L'OUTIL

I.1. Un catalogue de langues

Un préalable à tout travail d'envergure sur l'ensemble des langues africaines est l'établissement d'un inventaire des langues. Si l'on ne souhaite pas se perdre dans la question indécidable de la délimitation de l'objet, il est nécessaire d'introduire ici une part d'arbitraire. Une solution possible est d'adopter sans complexe le catalogue existant du site www.ethnologue.org, qui est le plus complet puisqu'il recense 2092 langues pour l'Afrique. En fait, la plupart des linguistes de terrain du LLACAN ne sont pas satisfaits de cet inventaire pour les zones qu'ils connaissent bien. C'est pourquoi nous avons jugé utile de modifier sensiblement ce catalogue, notamment pour les zones en question (Afrique de l'ouest, Afrique centrale). Notre version de la liste des langues d'Afrique comporte ainsi 2310 entrées (octobre 2008). Toutefois, un lien est toujours possible entre notre nomenclature et l'inventaire *ethnologue* via la norme ISO/DIS 639-3 (code de 3 lettres). Autre innovation importante : chaque langue est pourvue de coordonnées géographiques sous la forme d'un couple latitude/longitude définissant un point arbitrairement représentatif de la zone où est parlée la langue.

Un des objectifs du projet RefLex est d'**affiner l'inventaire des langues d'Afrique**. Deux voies sont envisagées : d'une part, l'interaction avec les participants au projet aura pour effet de mettre à jour le catalogue pour les zones où le LLACAN ne dispose pas de spécialistes (notamment les zones Bantu, Mande, Gur, Omotiques, Nilotiques...) ; d'autre part, les outils qui seront développés devront permettre d'élaborer des standards méthodologiques qui conduiront à une plus grande cohérence dans l'identification et la délimitation des langues.

I.2. Un catalogue de références

La base de données bibliographique en ligne WebBall (Web Bibliography of African Languages and Linguistics : http://sumale.vjf.cnrs.fr/Biblio/) recense tous les articles, ouvrages et, dans une moindre mesure, travaux universitaires concernant les langues africaines, postérieurs à 1920 (avec quelques exceptions). Elle contient près de 25000 références, et permet de se faire une bonne idée de la documentation disponible. Une recherche sur les termes *dictionnaire*, *dictionary*, *lexique*, *lexicon*, *wordlist*, *glossaire*, *vocabulaire* et *vocabulary* renvoie pas moins de 1881 références (au 20/10/2008) concernant quelques 812 langues. A cela il faut ajouter une partie des quelques centaines de descriptions et grammaires qui contiennent un lexique substantiel. On peut donc estimer que l'information lexicale disponible (de nature très inégale, bien sûr) pour l'Afrique concerne un bon millier de langues, soit près de la moitié.

La nature et la qualité des données sont très variables. On distingue grossièrement trois niveaux de précision : le *glossaire* (*wordlist*), qui ne comporte pas d'autre information que les correspondances mot à mot ; le *lexique* (*lexicon*), qui propose une petite information supplémentaire : catégorie grammaticale, classes nominale, forme de pluriel, d'infinitif, base lexicale, découpage morphologique... ; enfin, le *dictionnaire* (*dictionary*) comporte en général une définition plus détaillée, des informations grammaticales plus complètes et

souvent un ou plusieurs exemple(s). Aux deux extrémités, on peut citer par exemple les listes de mots compilées dans un ouvrage comme *West African Language Data Sheets* (Kropp-Dakubu éd. 1977, 1980, une soixantaine de mots par langue, mais pour plus de 80 langues), et le monumental dictionnaire hausa de Bargery (1934, plus de 60000 entrées, avec des exemples et des informations grammaticales et dialectales). Cette triple partition est importante : le projet RefLex ayant typiquement vocation à proposer des informations du type *lexique*, il s'ensuit que si les 'gros' dictionnaires y ont une place, ce sera sous une forme simplifiée. En effet, il n'est pas prévu pour l'instant (faute de moyens) d'intégrer des informations grammaticales détaillées ou des énoncés illustratifs interlinéarisés. A l'autre extrémité, les simples listes de mots feront l'objet de traitements plus aou moins automatisés visant à augmenter l'information disponible : calcul de schème syllabique, de schème tonal, extraction d'information morphologique (voir ci-dessous II.7.c). Enfin, pour certaines langues peu documentées, les seules ressources disponibles peuvent être des descriptions grammaticales, qu'il faut alors 'dépouiller' pour en extraire les informations lexicales.

I.3. L'accès aux sources

La grande originalité de ce projet, et ce qui doit en faire un véritable outil de **référence** est la garantie d'un accès facile aux sources originales. Chaque document dont le contenu a vocation a être mis en ligne sera donc également accessible sous forme d'image numérisée. De cette façon, les données saisies pourront toujours être confrontées à l'original. Concrètement, pour chaque fiche lexicale, un lien permettra d'afficher la page du document original correspondant. Chaque document sera muni d'un identifiant unique auxquel la communauté scientifique pourra faire référence.
Dans un premier temps, pour éviter autant que possible les problèmes liés au droit d'auteur, on adoptera les positions suivantes :
- On proposera des données supposées libres de droits (documents anciens ou données dont l'utilisation est expressément permise par les auteurs et/ou les éditeurs).
- Les documents originaux ne seront pas accessibles en une fois dans leur totalité. Cela signifie qu'un clic sur une forme mènera vers la page du document original où se trouve cette forme, et seulement cette page.
Cependant, il est évident que ces limitations devront être dépassées. Pour ce faire, il conviendra de passer des accords avec les éditeurs pour permettre la diffusion, dans des conditions qui devront être précisées, des documents les plus récents.

I.4. La numérisation des données

La partie la plus ingrate du projet est bien sûr la numérisation des données, c'est-à-dire à la fois la saisie des divers lexiques mais également la numérisation (*scan*) des documents, leur indexation et l'élaboration des métadonnées. De très nombreuses sources ont déjà été saisies ici ou là, et un énorme travail d'inventaire et d'harmonisation de l'existant est à faire. L'implication de nombreux partenaires prêts à partager leurs données aura pour résultat de rendre beaucoup plus supportable cette charge de travail : en effet, si une dizaine de partenaires participent à égalité à la numérisation des données, cela signifie que chaque partenaire aura accès à 10 documents pour chaque document traité.

I.5. Une structure souple et évolutive

La base de données qui sera réalisée dans le cadre du projet RefLex sera organisée classiquement en plusieurs **tables** reliées entre elles. On chercher à maintenir une structure globale assez souple, avec seulement trois tables indispensables :

a. La table des **Langues**. Il s'agit du catalogue des langues d'Afrique, évoqué ci-dessus (§ I.1.).

b. La table des **Sources**. Si chaque corpus lexical est bien lié à une langue, ce lien ne peut être direct : en effet, une même référence peut contenir des informations sur plusieurs langues, et une même langue peut se trouver renseignée dans plusieurs références. C'est pourquoi il est nécessaire d'introduire le concept de *source*, que l'on peut définir comme la combinaison d'une langue et d'une référence bibliographique. Ainsi par exemple, si cinq références contiennent du matériel lexical sur la langue *manjaku*, il y aura cinq entrées correspondantes dans la table des sources. Inversement, la référence WebBall 11273 (Kraft 1981, *Chadic Wordlists*), qui contient des lexiques de 65 langues, donnera lieu à 65 entrées dans la table des sources.
Cette notion de *source* correspond à peu près à la notion de *doculect*, introduite par M. Cysouw, J. Good et M. Haspelmath[1], avec cette différence qu'ici il est surtout question de documents publiés, ou dont la diffusion est expressément autorisée par leur auteur.
La table des sources est le véritable pivot de la base de données. Elle contiendra les liens vers les documents eux-mêmes, ainsi que les métadonnées correspondantes (comme par exemple des informations sur les conditions de l'enquête, la localisation du document original ou l'auteur de la saisie...).

c. La table des **Lexiques**. C'est là qu'est rangée l'information lexicale proprement dite. Cette table doit comporter un nombre important de champs, à la fois pour tenir compte de la diversité des données et des langues, mais également pour permettre de travailler sur tous les aspects du lexique. Ainsi, les informations concernant la **forme** dépassent de beaucoup la forme elle-même : il est prévu des champs pour la base lexicale, la racine, le genre, la classe nominale (au sg et au pl), le découpage morphologique, les schèmes syllabique, tonal, consonantique, vocalique, le degré d'alternance consonantique... La forme fera aussi l'objet de simplifications successives destinées à augmenter la souplesse des recherches et des tris (voir ci-dessous § I.6). Pour sa part, le sens donnera lui aussi lieu à des traitements variés, qui nécessiteront plusieurs champs (voir ci-dessous § I.7).

D'autres tables peuvent être ajoutées en fonction des besoins exprimés par les collaborateurs du projet. Parmi les possibilités envisagées, on peut citer une table des caractères utilisés avec leurs valeurs phonétiques, qui permettrait par exemple de dresser automatiquement les tableaux phonologiques des langues d'une manière homogène et rigoureuse ; on envisage également d'extraire des sous-ensemble de données que l'on placerait dans des tables séparées pour procéder à des opérations particulières, par exemple la reconstruction lexicale au sein d'un groupe de langue.

I.6. De la transcription unifiée au code phonétique

Lorsque l'on s'emploie à comparer des données linguistiques de provenances diverses, on est très rapidement confronté à des problèmes de transcriptions. Celles-ci varient selon les époques, les auteurs, les traditions académiques, les zones géographiques, et même les modes. Certaines langues disposent d'une orthographe, d'autres non. Bien entendu, c'est le rôle d'un outil de référence que de proposer les données d'une manière exactement conforme à l'original. Mais la variation est telle, dans la transcription, qu'aucune comparaison ne serait possible si l'on devait se limiter à la forme d'origine. Par conséquent, il est nécessaire de prévoir une **transcription unifiée**, qui ne se substituera pas à la transcription d'origine, mais viendra en complément. Celle-ci devra faire l'objet

[1] Voir l'article de Glottopedia : http://urts120.uni-trier.de/index.php/Doculect

d'un consensus, et son élaboration devra être l'un des premiers objectifs du projet. On peut dès à présent avancer les quelques propositions de bon sens suivantes :

- les tons sont notés par des diacritiques suscrits
- tous les tons sont marqués
- les digraphes doivent être évités s'il existe un signe unique équivalent (par ex. **ny** > **ɲ**, **tʃ** > **c**, **sh** > **ʃ**, etc...)

Cette première unification fournira un point de départ pour d'autres opérations de simplification : extraction de la base et de la racine (la différence entre les deux correspondant à la différence entre morphologie et étymologie, respectivement ; la base et la racine peuvent être identiques), et calcul du 'code phonétique'. Cette expression désigne une version simplifiée de la base lexicale dans laquelle chaque phonème est remplacé par une lettre majuscule désignant la classe phonétique à laquelle il appartient : les consonnes labiales sonores (**b**, **v**, **w**, **m**, etc.) sont remplacées par **B** ; les voyelles postérieures fermées sont remplacées par **U**, etc. Ce 'code phonétique' permet de rechercher très simplement toutes les formes correspondant à une certaine trame (*pattern*), mais cette tâche pourrait être effectuée à l'aide d'expressions régulières[2]. Son véritable intérêt est de servir de base de tri lors d'une recherche sur un autre champ : par exemple, si l'on lance une recherche sur un sens, le tri des résultats suivant le code phonétique permet d'afficher côte à côte des formes aussi différentes que **lò**, **ro**, **nō**, **ɖɔ**, etc., qui partagent le code **DO**.

I.7. Vers un thesaurus global

De la même manière qu'il a été jugé nécessaire d'unifier les transcriptions (cf. § précédent), il convient, sinon d'unifier, au moins d'harmoniser les traductions. Il y a à cela au moins deux bonnes raisons en amont :

- Les sources peuvent être dans des langues différentes : la documentation des langues africaines est le plus souvent en français ou en anglais, mais on rencontre aussi des documents en allemand, italien, portugais, néerlandais, pour ne citer que les principales langues européennes utilisées. Pour permettre des recherches par sens sur l'ensemble du corpus, il est nécessaire de disposer d'au moins un champ dont la langue est fixe. En fait, on prévoit pour l'instant deux champs : français et anglais.
- L'exigence de base, qui est de pouvoir disposer d'une reproduction exacte de l'original, conduit inévitablement à multiplier les formulations pour des notions qui peuvent être unifiées. Par exemple, beaucoup de langues africaines disposent d'un mot dont le sens est à la fois 'main' et 'bras'. Dans certaines sources, la traduction est notée 'main, bras'. Dans d'autres, elle est notée 'bras, main', 'bras ; main', 'bras ou main', etc.. Autre cas : pour les espèces végétales non identifiées, on rencontre des formulations aussi diverses que : 'espèce de plante', 'sorte de plante', 'plante sp.', 'plante', 'variété de plante', 'genre de plante'. Si l'on affiche une liste de mots triée suivant le sens, on voit bien que toutes ces expressions seront complètement dispersées, ce qui peut gêner considérablement certains types de recherche.

L'harmonisation des traductions est beaucoup plus délicate que l'unification des transcriptions. Il est par conséquent peu probable que l'on parvienne ici à un consensus. Il y aura inévitablement des conflits entre d'une part la nécessité de rester fidèle au sens (présumé) de départ, et d'autre part la tentation de regrouper des valeurs proches sous des notions générales. Il est possible, mais ceci doit être discuté avec les partenaires du projet, d'affecter plusieurs champs aux différents niveaux d'harmonisation des définitions : dans un premier champ, l'harmonisation pourra n'être que formelle (ponctuation, ordre des sens

[2] Une 'expression régulière' est un motif qui décrit un ensemble de séquences de caractères possibles. Par exemple, l'expression *[bvmw][aou]+* désigne toute séquence commençant par un symbole pris dans l'ensemble (bvmw) suivi d'un ou plusieurs symbole(s) pris dans l'ensemble (aou).

mulptiples). Dans un second, on procèdera à une simplification sémantique (par ex. 'couper' pour 'couper en morceaux', 'découper', 'trancher', etc.). On peut imaginer un troisième champ où figurera la 'grande notion' correspondante. Chacun de ces champs doit être doublé par son équivalent en anglais. Quoi qu'il en soit, il faut répéter que ces champs ne se substituent pas à la traduction originale, il s'y ajoutent. On ne perd aucune information, au contraire.

II. LES APPLICATIONS SCIENTIFIQUES

L'existence d'un corpus maximalement complet pour les lexiques de langues africaines ouvrira des potentialités sans précédent dans les disciplines suivantes : phonologie, comparatisme et reconstruction, sémantique lexicale, et évidemment typologie, grâce à des outils développés spécifiquement pour ce corpus, notamment dans le domaine de la statistique.

II.1. Phonologie

L'accès instantané aux inventaires phonologiques (consonnes, voyelles et tons) de centaines de langues offrira une perspective inédite sur la typologie des systèmes phonologiques attestés en Afrique.

II.2. Comparatisme et reconstruction

L'accès immédiat aux données lexicales de – potentiellement – toutes les langues d'Afrique permettra de renouveler la pratique de la comparaison de masse, initiée par Greenberg il y a 50 ans avec des données limitées. Mais surtout, le fait de pouvoir disposer, pour un groupe de langues donné, de matériel lexical standardisé et structuré permettra enfin de procéder à l'application systématique de la méthode comparative aux langues africaines, avec cet avantage considérable que représente la mise à disposition des données, grâce à laquelle les hypothèses pourront être testées et vérifiées par la communauté scientifique.

II.3. Sémantique lexicale

La structuration du lexique fait l'objet d'études renouvelées par l'outil informatique (cf. PROX). Il est évident que la masse de données accessible via RefLex constituera un champ d'investigation infini pour les recherches lexicales.

II.4. Autres champs d'investigation possibles : typologie des systèmes de classification nominale, de la dérivation verbale, études des emprunts et de la diffusion lexicale, étude de la stratification lexicale.

II.5. Cartographie

Chaque langue du catalogue étant dotée de coordonnées géographiques, tous les résultats des recherches effectuées sur la base RefLex pourront être visualisés sur une carte, chaque langue étant représentée par un point. Cette fonctionnalité sera extrêmement utile pour tous les domaines scientifiques abordés.

II.6. Méthodologie et outils

La base de données élaborée dans le cadre du projet RefLex sera bien entendu accessible sur Internet. Les conditions exactes de l'accessibilité (aux contributeurs, à la communauté scientifique, à tous) devront être définies par les participants au projet. Il est possible que l'on soit amené à définir plusieurs niveaux d'accessibilité.
Dans tous les cas, il sera nécessaire de développer des outils pour la recherche et le tri des données. Mais les besoins spécifiques des chercheurs et les exigences particulières de chaque discipline devront également susciter le développement d'outils plus spécifiques.

Parmi les outils déjà en cours de développement, on peut citer un module de statistique phonologique, ou un ensemble d'outils d'aide à la reconstruction lexicale.

III. DEROULEMENT DU PROJET

Ce projet peut tout à fait ne pas être limité dans le temps. Cependant, dans le cadre d'une réponse à l'appel d'offres ANR, sa durée initiale est fixée à **48 mois**. Les phases suivantes sont envisagées :

III.1. Catalogue des langues

L'objectif n'est pas pour l'instant d'élaborer une norme permettant de définir une langue, mais de s'accorder sur un **catalogue** opérationnel des langues d'Afrique. En effet, chaque document appelé à être intégré à la base de données devra être muni d'une référence précise comportant un code pour la langue. Lors de cette phase de mise au point, chaque partenaire sera sollicité en fonction de son domaine de spécialité pour affiner le catalogue existant (voir liste en annexe § IV.2). Il est probable que ce travail aura pour résultat une augmentation sensible du nombre des langues.

A l'issue de cette phase, on procèdera à l'élaboration définitive de la table des langues (structure et données).

III.2. Inventaire des sources et des données

Chaque partenaire sera également sollicité pour participer à l'inventaire des sources, qui seront affectées à un **niveau d'accessibilité**, selon la grille suivante :

niveau 1 : ce qui existe ou a existé
niveau 2 : ce qui est physiquement accessible
niveau 3 : ce qui est potentiellement intégrable (cf. questions de droits § I.3)
niveau 4 : ce qui est saisi
niveau 5 : ce qui est formaté et prêt à intégrer
niveau 6 : ce qui est intégré

A partir du niveau 3, et si possible également pour le niveau 2, il conviendra de recueillir pour chaque source les métadonnées minimales : type de document, langue concernée, date, auteur, etc.

Au cours de cette phase, on procèdera à l'élaboration définitive de la structure de la table des sources.

III.3. Mise au point technique

Sur la base des propositions détaillées ici (I.6 et I.7 ci-dessus, et annexe IV.1 ci-dessous), on s'accordera sur la structure définitive de table des lexiques : liste des des champs, contenu des champs, standards de transcription.

Parallèlement, une maquette de site web sera développée (basée sur les structures de base de données existantes), qui devra permettre l'accès aux données. La version définitive du site ne pourra être conçue que lorsque la structure définitive de la base de données sera arrêtée.

Il est prévu deux réunions au cours de la première année pour les phases initiales de structuration. Ces réunions seront également l'occasion pour les chercheurs d'exprimer leur souhaits au niveau du développement d'outils spécifiques d'exploitation des données.

III.4. Production et exploitation

A partir de la 2ème année les structures seront en place pour accueillir les données. Le travail de préparation des données (saisie, numérisation des documents, formatage), qui pourra avoir commencé dès le début du projet, entrera dans sa phase productive. Les

données seront intégrées à la base au fur et à mesure de leur disponibilité, jusqu'au terme du projet (et si possible après).

La version opérationnelle du site web sera développée.

Les premiers outils seront testés et finalisés.

Les 3ème et 4ème années seront consacrées à l'exploitation des premiers outils, au développement et à l'exploitation de nouveaux outils, et bien sûr à l'enrichissement de la base de données. En particulier, la quatrième année verra la préparation d'une publication réunissant les résultats des travaux menés sur ce corpus.

III.5. Objectifs chiffrés

Etant donné le volume des données déjà disponibles et le volume des données partiellement formatées, il n'est pas déraisonnable d'envisager que la base de données contienne, au bout de deux ans, au moins 500 sources pour environ 500000 entrées lexicales. A la fin du projet, il est prévu d'atteindre **1 million** d'entrées lexicales, pour au moins **1 millier** de sources, représentant entre 500 et 700 langues différentes.

III.6. Tableau récapitulatif du déroulement du projet

<table>
<tr><td></td><td colspan="2">année 1</td><td colspan="2">année 2</td><td colspan="2">année 3</td><td colspan="2">année 4</td></tr>
<tr><td>inventaires</td><td>langues</td><td>sources</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>structuration</td><td></td><td>lexiques</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>alimentation</td><td colspan="2">saisie</td><td colspan="6">saisie, harmonisation& intégration lexiques & sources</td></tr>
<tr><td>valorisation</td><td colspan="2"></td><td colspan="4"></td><td colspan="2">public. résultats</td></tr>
<tr><td rowspan="2">outils</td><td colspan="2" rowspan="2">propositions</td><td colspan="4">développement</td><td></td><td></td></tr>
<tr><td></td><td colspan="5">exploitation</td></tr>
<tr><td>site web</td><td colspan="2">maquette</td><td colspan="2">développement</td><td colspan="4">intégration des outils</td></tr>
<tr><td></td><td colspan="8">recrutement ingénieur 48 mois</td></tr>
</table>

IV. ANNEXES

IV.1. Liste des champs de la table **Lexiques**

	nom	**contenu**
1	id	identifiant unique
2	SOU	ID de la source
3	REF	Référence de la fiche dans la source (page, num, etc.)
4	FRM	Forme d'origine, exactement conforme à la source
5	FUN	Forme d'origine, transcription unifiée
6	FSI	Base lexicale
7	COD	Code phonétique
8	STO	Schème tonal
9	CGR	Catégorie grammaticale (inventaire unifié)
10	CGO	Catégorie grammaticale utilisée par l'auteur
11	SCS	Schème syllabique
12	CLS	Classe nominale singulier
13	CLP	Classe nominale pluriel
14	GEN	Genre (m, f, n, ...)
15	PLU	Forme de pluriel
16	SCM	Schème morphologique
17	SCC	Schème consonantique
18	SCV	Schème vocalique
19	RAC	Racine
20	ALT	Degré d'alternance consonantique

21	TRA	Traduction de l'auteur
22	TUF	Traduction unifiée français
23	TUE	Traduction unifiée anglais
24	COA	Commentaire de l'auteur
25	CSA	Commentaire ajouté lors de la saisie
26	FIA	Indice de fiabilité
27	EMP	Langue source (pour emprunts)
28	FEM	Forme dans la langue source (pour emprunts)
29	TEM	Sens dans la langue source (pour emprunts)
30	TRI	Champ libre pour tris
31	DAT	Date de dernière modification

IV.2. Familles de langues et spécialistes partenaires pressentis

Niger-Congo		
Adamawa	R. Boyd	LLACAN
Atlantique	K. Pozdniakov	LLACAN
	G. Segerer	LLACAN
Benue-Congo	R. Blench	
Dogon	R. Blench	
Bantu	K. Bostoen	Tervuren
	Mark van de Velde	LLACAN
	L. Van Der Veen	DDL
	Sophie Manus	DDL
	G. Philippson	DDL
Gur	K. Winkelmann	Bayreuth
Mande	V. Vydrine	Univ. St Petersburg
	H. Tröbs	Univ. Mayence
	F. Lüpke	SOAS
Oubanguien	P. Nougayrol	LLACAN
	Y. Moñino	LLACAN
	R. Boyd	LLACAN
Afro-Asiatique		
Tchadique	D. Ibriszimow	Univ. Bayreuth
	B. Caron	IFRA Nigeria, LLACAN
Couchitique	M. Tosco	Univ. Turin
	G. Sava	Univ. Naples
	R. Kießling	Univ. Hambourg
	M. Mous	Univ. Leiden
Songhay	R. Nicolaï	IUF
Nilo-Saharien		
Soudanique Central	P. Boyeldieu	LLACAN
SBB	P. Boyeldieu	LLACAN
	P. Nougayrol	LLACAN
Soudanique Oriental	C. Rilly	LLACAN
Nilotique	G. Dimmendaal	Univ. Cologne
Saharien	G. Dimmendaal	Univ. Cologne
Khoe-San	G. Starostin	Univ. Moscou