Kassian A., Starostin G., Dybo A. & Chernov V., 2010, The Swadesh wordlist: an attempt at semantic specification, *Journal of Language Relationship* 4, p. 46-89.
Klimov G. A., 1964, *Etimologicheskij slovar' kartvel'skikh jazykov [Etymological dictionary of Kartvelian languages]*, Moscow, Izdatel'stvo Akademii Nauk SSSR.
Kogan L., 2006, Lexical Evidence and the Genealogical Position of Ugaritic (I), *Babel und Bibel 3. Orientalia et Classica* vol. XIV, Winona Lake, Indiana, Eisenbrauns, p. 429-488.
Kurilov G. N., 2001, *Jukagirsko-russkij slovar' [Yukaghir-Russian dictionary]*, Novosibirsk, Nauka Publishers.
List J.-M. & Nelson-Sathi S. & Geisler H. & Martin W., 2014, Networks of lexical borrowing and lateral gene transfer in language and genome evolution, *BioEssays* 36.2, p. 141-150.
Nikolayev S. & Starostin S., 1994, A North Caucasian *Etymological Dictionary*, Moscow, Asterisk Publishers.
Nikolayeva I., 2006, *A Historical Dictionary of Yukaghir*, Berlin – New York, Mouton de Gruyter.
Orel V., 2003, *A Handbook of Germanic Etymology*, Leiden – Boston, Brill.
Pokorny J., 1958, *Indogermanisches etymologisches Wörterbuch*, Bern-München, Francke Verlag.
Rastorguyeva V. S. & Edelman D. I., 2007, *Etimologicheskij slovar' iranskikh jazykov. Tom 3: f-h [Etymological dictionary of Iranian languages. Vol. 3: f-h]*, Moscow, Vostochnaja literatura.
Starostin G., 2010, Preliminary lexicostatistics as a basis for language classification: A new approach, *Journal of Language Relationship* 3, p. 79-116.
Starostin G., 2013a, Lexicostatistics as a basis for language classification: increasing the pros, reducing the cons, in H. Fangerau, H. Geisler, Th. Halling, W. Martin (ed.), *Classification and Evolution in Biology, Linguistics and the History of Science: Concepts – Methods – Visualization*, Stuttgart, Franz Steiner Verlag, p. 125-146.
Starostin G., 2013b, *Jazyki Afriki. Opyt postrojenija leksikostatisticheskoj klassifikacii. Tom I: Metodologija. Kojsanskije jazyki [Languages of Africa: an attempt at a lexicostatistical classification. Volume I: Methodology. Khoisan Languages]*, Moscow, Jazyki slav'anskoj kul'tury.
Starostin S., 1991, *Altajskaja problema i proiskhozhdenije japonskogo jazyka [The Altaic problem and the origins of the Japanese language]*, Moscow, Nauka Publishers.
Starostin S., 1995, Sravnitel'nyj slovar' jenisejskikh jazykov [Comparative dictionary of Yeniseian languages], *Ketskij sbornik. Lingvistika [Ket Volume. Linguistics]*, Moscow, Vostochnaja literatura, p. 176-315.
Turner R. L., 1966, *A Comparative Dictionary of the Indo-Aryan Languages*, Oxford University Press.
Vries J. de, 1962, *Altnordisches Etymologisches Wörterbuch*, Leiden, Brill.

RefLex : la reconstruction sans peine

Guillaume Segerer*

INTRODUCTION

Cet article se propose de décrire le fonctionnement d'un module d'aide à la reconstruction phonologique et lexicale. Ce module est intégré au site web d'exposition et d'exploitation de RefLex (www.reflex.cnrs.fr), projet dont l'auteur de ces lignes est à l'initiative et qui s'articule autour d'une base de données lexicales sur les langues d'Afrique. Par conséquent, les exemples concrets cités ci-dessous sont empruntés aux langues africaines, mais l'ensemble d'outils dont il va être question peut être utilisé pour n'importe quel ensemble de langues.

Je présenterai d'abord le projet RefLex dans son ensemble, puis je décrirai avec davantage de détails le module d'aide à la reconstruction (que l'on a tendance à appeler plus brièvement 'module de reconstruction' bien que la machine ne prenne pratiquement aucune décision). Ce module est, et sera encore probablement longtemps, en cours de perfectionnement. Certaines des caractéristiques décrites ici sont appelées à évoluer rapidement pour offrir encore plus de confort d'utilisation, et l'utilisateur qui voudrait jouer à reconstruire après la lecture de ce texte trouvera très certainement des fonctionnalités nouvelles. Je tenterai ici d'évoquer les nouveautés en cours de développement.

1. REFLEX : PRÉSENTATION GÉNÉRALE

1.1. Les principes

RefLex est un projet combinant une base de données lexicales sur les langues d'Afrique, des outils de traitement et un site internet pour l'accès aux données et aux outils[1]. Le projet a réellement démarré en 2010, grâce à un financement de

* LLACAN-Sorbonne Paris Cité-INALCO – CNRS. Courriel : guillaume.segerer@cnrs.fr. Ce travail est lié au programme "Investissements d'Avenir" géré par l'Agence Nationale de la Recherche ANR-10-LABX-0083 (Labex EFL).
Avertissement : Cet article décrit une application en ligne, et par conséquent contient inévitablement des images ou 'captures d'écran' dont les couleurs peuvent être difficiles à discerner dans une version en niveaux de gris. Ces images ne peuvent donc fournir qu'une illustration approximative du contenu de l'écran. Le lecteur est donc encouragé à consulter le site original.
[1] Le développement a été assuré par Sébastien Flavier, ingénieur d'études au laboratoire DDL (Lyon), partenaire du projet

l'ANR. Le financement a pris fin en 2015 mais le projet reste un des axes forts du programme de recherche du laboratoire LLACAN.

Ce projet est unique car il associe plusieurs principes qui font généralement défaut dans les bases de données lexicales en ligne :

- il s'agit véritablement d'un lexique de référence, dans la mesure où les données consultables sont liées à des images numériques des sources originales dont elles proviennent. C'est cette caractéristique unique qui a donné son nom au projet (RefLex : Reference Lexicon). Ainsi par exemple, pour la première fois, les mesures lexicostatistiques (un des exercices favoris de la linguistique comparative des langues africaines) deviennent intégralement reproductibles, et les choix concernant les ressemblances lexicales peuvent être discutés.
- Beaucoup des manipulations que l'on peut souhaiter faire sur les données sont réalisables en ligne : recherches sur critères complexes, comptages divers, statistiques combinatoires (voir K. Pozdniakov, ce volume), représentation cartographique des données, et surtout assistance à la reconstruction, module qui fait l'objet de cet article.

1.2. Les données : langues

Pour associer des données lexicales à une langue, il est nécessaire de disposer d'un inventaire des langues d'Afrique. Plusieurs inventaires existent actuellement, les deux principaux étant le catalogue correspondant à la norme iso 639-3, disponible sur le site www.ethnologue.com, et celui de l'institut Max Planck de Leipzig, récemment publié (http://glottolog.org/). Ces deux catalogues visent non seulement à inventorier toutes les langues du monde, mais en proposent en outre une classification généalogique. Les deux classifications sont assez différentes, celle de Glottolog étant beaucoup plus prudente (autrement dit, elle comporte davantage de familles). Notre catalogue (disponible à http://reflex.cnrs.fr/Lexiques/webcal/index.html) est différent des deux catalogues cités ci-dessus. Du fait de la présence au LLACAN de spécialistes reconnus de plusieurs familles de langues africaines, il est plus détaillé pour certaines zones géographiques, notamment l'Afrique centrale, la région Sénégal-Guinées, certaines parties du Nigeria. En septembre 2015, il recense 2441 langues.

1.3. Les données : sources

Comme RefLex fournit des données lexicales provenant de documents publiés (pour la grande majorité, mais voir ci-dessous), avec accès aux documents d'origine, chaque lexique, en plus d'être associé à une langue, est associé à un document. Cette combinaison d'une référence linguistique et d'une référence bibliographique est appelée source dans RefLex, ce qui correspond en partie au concept de doculect[2] introduit par Jeff Good en 2006. Par exemple, le document dont l'identifiant est 18737 est un document de Keith Snider publié au Ghana en 1989 et contenant des listes lexicales pour 5 langues du groupe Kwa, auxquelles

[2] Cf. http://www.glottopedia.org/index.php/Doculect

correspondent donc autant de sources dans RefLex. Inversement, la langue joola karon (Atlantique) est représentée dans RefLex par 4 documents : Barry 1987, Carlton & Rand 1994, Sambou 2007 et Wilson 2007, ce qui fait donc 4 sources.

Une source de RefLex n'est pas la combinaison unique d'une langue et d'une référence bibliographique. En effet, lorsqu'un document contient des informations à propos de variantes dialectales, il n'est pas toujours facile de décider si ces variantes doivent avoir le statut de langue. Par exemple, Fresco 1970 (réf. 6910) fournit des listes lexicales pour le yoruba commun ainsi que pour 7 variétés dialectales de yoruba. Ces 8 sources sont toutes la combinaison de la référence n° 6910 (Fresco 1970) et de la langue n° 835 (yoruba).

Actuellement (juin 2016), le nombre de sources dans RefLex est de 1164 et correspond à 705 langues.

1.4. Les données : lexiques

Toutes les données lexicales des langues d'Afrique ont par nature vocation à être intégrées à RefLex. Pour des raisons pratiques cependant, certaines critères peuvent influer sur le choix des données à importer en priorité :

- Les données publiées sont préférées aux données brutes, pour des questions de stabilité. Cependant, dans certains cas, notamment lorsqu'il s'agit de données de première main concernant des langues peu ou pas documentées, des données non publiées peuvent être incluses dans RefLex.
- Les données déjà numérisées (en partie ou en totalité), ou celles déjà saisies sont évidemment plus faciles à intégrer.
- Les données qui font l'objet d'un travail de recherche particulier sont également prioritaires. Le projet ANR RefLex comportait des tâches précises et impliquait une vingtaine de chercheurs. Nous avons donc mis l'accent sur les données concernées par ces tâches.

La taille des lexiques est très variable. Les plus petits ont quelques dizaines d'entrées, les 3 plus importants on plus de 20000 entrées chacun (il s'agit du bambara, du peul du Maasina et du sereer).

Leur structure est également très variable et va de la simple liste de mots au dictionnaire complexe. Certaines informations, comme les exemples d'illustrations, ne sont pas reprises dans RefLex. En revanche, les formes citées dans les sources servent de base pour extraire de l'information nouvelle : schème tonal, schème syllabique, code phonétique. Ces informations vont permettre d'effectuer des recherches et des tris qui peuvent se révéler très utiles lors de la recherche de cognats.

Par ailleurs, les deux champs indispensables que sont la forme et le sens font l'objet de procédures d'harmonisation. Pour la forme, il s'agit surtout d'unifier les transcriptions de façon à pouvoir facilement repérer des formes semblables. En effet, selon les époques, les traditions linguistiques ou orthographiques, la langue de l'auteur ou la langue cible de la traduction, les transcriptions sont extrêmement variables. En ce qui concerne le sens, deux champs sont réservés à des interprétations simplifiées (respectivement en français et en anglais) de la traduction originale. Comme pour les formes, cette simplification sert à ce que lors des recherches et des tris, des sens semblables soient affichés ensemble. Par

exemple, il n'est pas rare que le même mot désigne à la fois 'le bras' et 'la main'. Mais suivant les cas, les traductions proposées peuvent présenter des variations, même mineures, pouvant rendre plus difficile la comparaison : 'bras, main' ; 'main, bras' ; 'le bras, la main', 'hand, arm', 'arm, hand', etc. Une recherche sur le terme 'bras' par exemple ne renverra pas les mots dont la traduction originale est en anglais ; les définitions commençant par 'main' seront affichées loin de celles commençant par 'bras', etc. On a donc jugé utile d'ajouter, à côté de la traduction d'origine, deux champs (en français et en anglais) contenant une version relativement standardisée de cette traduction

Bien sûr, la simplification de la traduction est beaucoup plus complexe que la simplification de la transcription, et il subsiste dans RefLex de nombreuses irrégularités. Mais le principe même, malgré un énorme coût en temps de préparation, s'est montré très efficace.

Au mois de juin 2015, le nombre des entrées lexicales dans RefLex a dépassé le million.

1.5. Les outils

Outre le module d'aide à la reconstruction présenté ci-dessous, RefLex propose un module statistique et un module cartographique.

- Le module statistique permet de réaliser toutes sortes de comptages et d'afficher les résultats sous différentes formes, dont des tables de contingences pour les statistiques dites 'combinatoires' et les statistiques dites 'croisées'. Les statistiques combinatoires comptent des combinaisons à l'intérieur d'un champ unique. Typiquement, on les utilise pour explorer les combinaisons de phonèmes et/ou de tons dans les langues. Les statistiques 'croisées' permettent d'évaluer la dépendance entre le contenu de deux champs, par exemple la structure syllabique et les parties du discours. L'obtention de ces statistiques et leur exploitation scientifique sont évoqués dans ce volume (KP, réf ?).

- Le module cartographique affiche le résultat d'une recherche sur une carte d'Afrique, chaque langue étant représentée par un point. Il s'agit d'un véritable module, et non d'une simple fonction, car les cartes ainsi obtenues peuvent être modifiées par l'utilisateur, par exemple pour afficher certaines informations (formes, sens, nom de langue) ou pour modifier certains paramètres (forme et couleur des points, ajout de données, etc.).

Ces deux modules peuvent être utilisés indépendamment de la base de données, ce qui n'est pas le cas du module de reconstruction.

2. Le module d'aide à la reconstruction

La puissance de calcul de l'ordinateur peut être d'une aide précieuse dans les tâches fastidieuses de la reconstruction phonologique et lexicale. Cependant, aucune des diverses tentatives pour remplacer complétement l'esprit humain dans cette entreprise n'ont donné de résultats convaincants. Qu'il s'agisse de rechercher automatiquement des cognats potentiels, ou d'aligner correctement les segments au sein des séries lexicales, l'algorithmique n'est pas pour l'instant

en mesure de fournir des réponses correctes à tous les pièges que posent les langues naturelles. En revanche, les capacités de l'ordinateur à trier, stocker, retrouver, organiser l'information peuvent être utilisées pour soulager le linguiste et lui apporter un maximum de confort de travail. C'est dans cet esprit qu'a été conçu le module d'aide à la reconstruction de RefLex.

D'une manière générale, l'utilisateur de RefLex manipule des *objets*. Ces objets peuvent avoir une existence préalable (les sources, les fiches lexicales) ou être créés par l'utilisateur. Le premier objet qu'il est nécessaire de créer pour utiliser le module de reconstruction est le *dataset*. Il s'agit de l'ensemble de sources au sein duquel se fera la recherche de cognats (cet anglicisme désigne une série lexicale dont les éléments sont supposés descendre d'un étymon commun)[3]. L'intérêt de travailler sur un nombre réduit de sources découle de la nature même de RefLex, qui a pour vocation à accueillir toutes les données disponibles, indépendamment de leur qualité. Dans le cadre d'un travail de reconstruction, il est important de ne sélectionner que les sources les plus dignes de confiance, lorsque le choix existe.

A partir du dataset, l'objectif est de créer des cognats, ou *cognate-sets* (voir ci-dessus). La méthode la plus simple est de lancer une recherche à l'aide de l'interface appropriée :

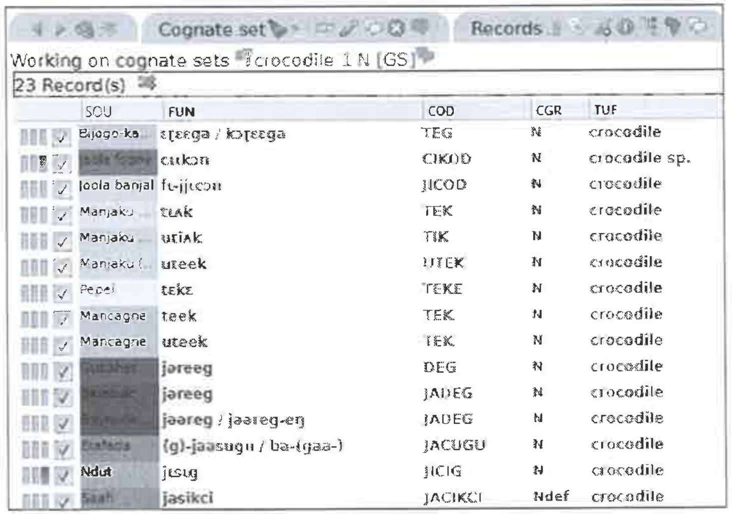
Par défaut, la recherche porte sur tous les champs dédiés au sens : la traduction d'origine, mais aussi les champs de traduction simplifiée en français et en anglais. Il est bien sûr possible de lancer une recherche sur un autre champ, comme il est possible de sélectionner certaines sources seulement, ou de combiner plusieurs critères de recherche. D'autres fonctionnalités (expressions régulières, exclusion) rendent l'outil de recherche extrêmement puissant.

L'affichage des résultats se fait sur la zone centrale de l'espace, sans que ne disparaisse l'interface de recherche. Par défaut, seuls quelques champs sont

[3] Il importe dès à présent de signaler que beaucoup des objets de RefLex peuvent être partagés entre plusieurs utilisateurs. Le partage d'un objet implique le partage des objets qui en dépendent, ce qui permet un véritable travail collaboratif.

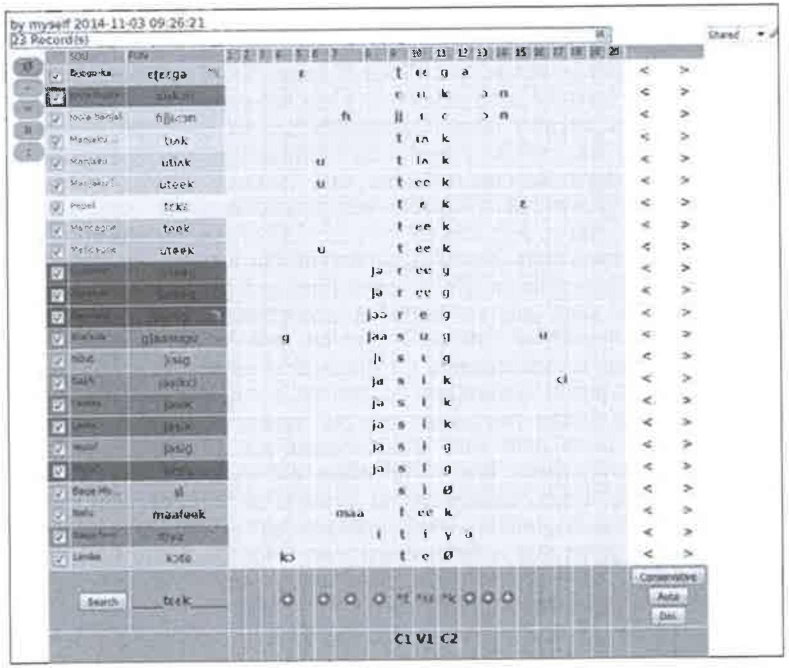
affichés et l'ordre des fiches suit l'ordre des sources. Ce dernier peut sembler arbitraire puisqu'il dépend directement de l'ordre dans lequel les sources ont été créées dans la base de données. En outre, chaque source est repérée par une couleur spécifique, ce qui s'avère extrêmement pratique lorsque l'on est appelé à travailler un certain temps sur un dataset. Ces deux caractéristiques, l'ordre des sources et leur couleur, sont personnalisables. On peut ainsi afficher par défaut les sources dans un certain ordre et avec certaines couleurs, par exemple en fonction des regroupements généalogiques présumés au sein du dataset. Cette possibilité peut sembler anodine, mais elle fait gagner un temps précieux. C'est même ce type de fonctionnalité 'de confort' qui est au centre de la philosophie de RefLex.

Une fois les résultats affichés, l'utilisateur va souhaiter en sélectionner certains, qu'il estime constituer un ensemble de réflexes d'un étymon supposé. La sélection se fait de manière très intuitive, et les fiches sélectionnées sont affichées dans la partie supérieure de l'écran. Cet ensemble de fiche peut alors être sauvegardé : il s'agit d'un cognate-set. Voici à quoi ressemble le cognate-set 'crocodile' créé par moi-même sur le dataset 'atlantique_base':



SOU	FUN	COD	CGR	TUF
Bijogo-ka	ɛɺɛɛga / kɔɺɛɛga	TEG	N	crocodile
	cɩkɔn	CIKOD	N	crocodile sp.
Joola banjal	fʊ-jjɩcɔn	JICOD	N	crocodile
Manjaku	tʌk	TEK	N	crocodile
Manjaku	utiʌk	TIK	N	crocodile
Manjaku (...	uteek	UTEK	N	crocodile
Pepel	tekɛ	TEKE	N	crocodile
Mancagne	teek	TEK	N	crocodile
Mancagne	uteek	TEK	N	crocodile
	jəreeg	DEG	N	crocodile
	jəreeg	JADEG	N	crocodile
	jəəreg / jəəreg-eŋ	JADEG	N	crocodile
	(g)-jaasugu / ba-(gaa-)	JACUGU	N	crocodile
Ndut	jɩsɩg	JICIG	N	crocodile
	jasikci	JACIKCI	Ndef	crocodile

A ce stade, il s'agit d'une simple hypothèse sur le fait que tous ces termes ont une origine étymologique commune. Parmi les arguments susceptibles de renforcer cette hypothèse, il y a le fait que les correspondances phonétiques au sein de cet ensemble appartiennent à des séries relativement régulières. Il convient donc de procéder à l'*alignement* des formes pour que les segments en correspondance soient affichés les uns sous les autres. Les séries ainsi formées vont alors apparaître en colonnes, et chacune de ces colonnes pourra être nommée. Voici par exemple l'alignement que j'ai créé pour le cognate-set 'crocodile' :



On voit que je n'ai pas souhaité nommer toutes les colonnes. En effet, les langues du groupe Atlantique, dont il s'agit ici, sont toutes des langues à classes nominales présentant une morphologie souvent complexe. J'ai donc surtout cherché à aligner les segments appartenant à la racine lexicale, et la position des autres éléments dans le tableau est plus ou moins arbitraire. On voit également que certaines cases du tableau contiennent plus d'un segment, d'autres contiennent l'élément Ø, et certaines cases vides sont insérées à l'intérieur d'un mot. Ces propriétés du sous-module d'alignement permettent de gérer facilement les digraphes, les fusions et les délétions. Toutefois, ce type de tableau ne permet pas pour l'instant de gérer efficacement les métathèses.

Les boutons situés en bas à droite du tableau permettent de nommer les correspondances automatiquement (selon certains critères, qu'il n'est pas nécessaire de détailler ici).

Les deux dernières lignes du tableau permettent de donner respectivement un nom et une étiquette à chaque colonne. Seules les colonnes dans lesquelles figure plus d'un élément peuvent être nommées. Le nom est laissé au libre choix de l'utilisateur. L'étiquette est contrainte, et sert à identifier la position au sein du mot, ou d'autres caractères utiles facilement représentables par des symboles[4]. Dans l'exemple ci-dessus, les symboles C1, V1 et C2 ont été utilisés pour

[4] La liste peut en fait être modifiée par l'utilisateur.

indiquer respectivement la première consonne, la première voyelle et la deuxième consonne de la base lexicale.

A gauche de la partie basse du tableau figurent deux cellules : la première contient un bouton 'search', qui permet de chercher parmi les correspondances déjà nommées celles qui sont le plus compatibles avec les correspondances du tableau. La seconde contient une pseudo-reconstruction, élaborée en concaténant les noms des correspondances précédés de '*'. Les correspondances non nommées sont remplacées par '_', celles non précédées de '*' sont remplacées par '@'. Cette information sera reprise lors de l'affichage en tableau de la liste de toutes les correspondances. Le choix d'un symbole 'neutre' (le signe @) pour représenter une correspondance dont le nom n'est pas précédé de * découle du principe suivant : le nom qui est donné à une correspondance est considéré comme une simple hypothèse tant qu'il n'est pas précédé de *, et ne peut donc être utilisé pour une reconstruction. Ce parti pris est destiné à favoriser les 'bonnes pratiques', parmi lesquelles on trouve aussi le fait qu'il n'est pas souhaitable de faire figurer plus d'un mot par source dans un alignement, ou d'utiliser la même fiche dans plus d'un cognate-set. Ces règles ne sont pas imposées par le programme, mais l'interface est conçue pour faciliter leur respect. Ainsi, lors d'une recherche au sein d'un dataset, les fiches qui appartiennent déjà à un cognate-set vont apparaître sur un fond gris, avec à droite le nom du cognate-set auquel la fiche appartient. Ce nom est surligné en rose si la fiche est présente dans plus d'un cognate-set (y compris si ces cognate-sets appartiennent à des utilisateurs différents) :

	kuyi-kuβei	II	num	quatre-vingts	quatre-vingts.		
Manjaku	bakər	BAKAD	num	quatre	quatre	quatre 1	
Manjaku	berənas	BODAC	temp	dans quatre jours	dans quatre jours		
Manjaku	bakər	BAKAD	num	quatre	quatre	quatre 1	
Manjaku	buənas	BIDAC	temp	dans quatre jours	dans quatre jours		
Manjaku	kabakər	BAKAD	N	pagne sp.	pagne formé de quatre bande...	quatre 2	
Pepel	ŋuakər	GUAKAD	num	quatre	quatre.	quatre 2	
Pepel	ŋumakər	GUBAKAD	num	quatre	quatre.	quatre 2	
Pepel	uaka	UAKA	num	quatre	quatre.	quatre 2	
Pepel	uakr	UAKD	num	quatre	quatre.	quatre 2	

La recherche de cognats, surtout au début, se fait principalement à partir du sens. On sait que la probabilité de trouver des formes apparentées est plus grande pour les termes du vocabulaire de base. Mais RefLex offre des possibilités pour étendre la recherche dans plusieurs directions. On peut par exemple, une fois qu'un cognate-set est défini, chercher à l'étendre en recherchant des formes proches, à l'aide du champ COD (*phonetic code*). Il s'agit d'une variante très simplifiée de la forme d'origine[5], dans laquelle chaque phonème est remplacé par une lettre majuscule représentant la classe de sons correspondante. Ainsi, la lettre B représente toutes les consonnes labiales sonores, soit *b*, *v*, *m*, *w*, *ɓ*, *β* par exemple. Cette simplification permet non seulement d'effectuer très simplement des recherches qui autrement demanderaient l'écriture de fastidieuses formules

[5] Lorsque la morphologie de la langue est connue, on utilise de préférence la base lexicale au lieu de la forme complète.

(des *expressions régulières*, lesquelles sont également permises bien sûr), mais offre aussi et surtout la possibilité de trier les résultats d'une manière qui fasse apparaître *ensemble* les formes non identiques mais proches. Il est alors beaucoup plus facile de repérer des termes apparentés qui ou bien n'auraient pas été traduits de la même façon dans les différentes sources, ou bien auraient des sens proches mais non similaires. C'est ainsi que le terme *kùmbá* ('hippopotame' dans la langue *baga mboteni* de Guinée) a pu être rapproché d'une racine signifiant 'porc' dans de nombreuses langues de la région (sud Sénégal / Guinées) :

Balante fr...	kumba / k	KUBA	N	porc
Balante k...	kumba / kkumba	KUß	N	porc ; phacochère
Bulfada	(w)-kumba / ge-h.	KUBA	N	porc
Kasanga	kumbaaru	KUBADU	N	porc
Pular	kummbaaru	KUBADU	N	porc
Baga Mb...	kùmbá	KUBA	N	hippopotame

Lorsqu'un cognate-set semble raisonnablement valide, on peut procéder à son alignement pour établir les séries de correspondances (voir ci-dessus). Une fois que l'on dispose d'un certain nombre d'alignements, il est possible de rechercher des régularités. Cette recherche n'est pas automatique mais est facilitée par les nombreuses options d'affichage des correspondances. Il est notamment possible de rechercher les occurrences d'un symbole particulier dans un contexte donné, pour une source donnée. Le tableau des résultats permet de visualiser directement les irrégularités, ce qui peut conduire à modifier les cognate-sets en conséquence. Dans l'exemple suivant, extrait d'un dataset contenant 14 sources sur les langues du groupe Joola (Niger-Congo, Atlantique), on a sélectionné les séries de correspondances nommées {*f}, triés sur la source *joola banjal*. On voit immédiatement que certaines sources ont un *f* où d'autres ont un *h*, qu'il existe quelques exceptions, et que le joola Banjal présente assez régulièrement des réalisations géminées lorsque la voyelle reconstruite est longue :

Correspondences : 46

Cognate set	Assumed reconstruction	Class	Label	joola fogny	joola banjal	joola kasa esundaalu	joola kasa	keersk	joola mlomp	joola keron	joola kwestey	Tags
racine	fufaar__		*f	f	w		h	h	h			C1
grenouille	__ɛfɔɔl		*f	f	ff	f	h		h		h	C1
traîner	__fuul__		*f	f	f	ff	h	h			h	C1
se coucher	___fu(l/nt)_		*f	f	f	ff	h		h	h		C1
souffler	____fuu(l/t)__		*f	f	f	ff		f	Ø			C1
train	_fə(g/j)__		*f	f	f	ff	f	h	h	h	h	h	h	h	h	h	C1
déterrer	___fɔk_		*f	f	f	f		h	h	h						h	C1
enterrer	___fɔk__		*f	f	f	f	f	h		h						h	C1

Le tableau peut être trié sur chacune des colonnes, par un clic sur son en-tête. En cliquant sur le nom d'un cognate-set à gauche, on peut éditer celui-ci, et modifier l'alignement ou le nom des correspondances.

Il serait évidemment très utile d'obtenir automatiquement l'inventaire des correspondances les plus régulières, et cette option est actuellement à l'étude. Mais cette automatisation n'aura pas que des conséquences positives, dans la mesure où les séries les plus triviales seront favorisées, risquant de masquer des correspondances plus exotiques. Dans certain cas en effet, une correspondance unique peut être extrêmement convaincante et influer sur la recherche des cognats et la découverte d'autres correspondances du même type, comme va le montrer l'exemple suivant, tiré de notre travail sur les langues atlantiques :

Dans ces langues, le mot pour 'œil' est extrêmement stable, ce qui signifie que presque toutes les langues documentées présentent une racine lexicale semblable. Voici la plupart de ces formes, dont on a isolé la consonne finale pour faire apparaître la correspondance[6] :

LANGUE	SOURCE	'OEIL'	C#
Bijogo-kagbaaga	Segerer 2002	nɛ / ŋɛ	Ø
Joola fogny	Sapir 1970	cil	l
Joola kasa	Wintz 1909	jikiɬ ~ jikill / ku-	ɬ
Joola banjal	Bassène 2006	ji-cil	l
Joola kwaatay	Payne 1992	dɪkɪn	n
Karon	Sambou 2007	nikin	n
Bayot	Diagne 2009	sio	o
Manjaku bok	Doneux 1975	kəs	s
Manjaku de Bassarel	Buis 1990	pəkəs	s
Pepel	Ndao 2010	kil	l
Mancagne	Trifkovic 1969	kəʃ	ʃ
Balante ganja	Creissels & Biaye 2015	f-gít / g-gít	t
Balante fraasɛ	Wilson 2008	f-kit / k	t
Balante kentohe	Doneux & al 1984	fkit / kkit	t
Gubaher	Cobbinah 2013	sijil	l
Baynunk guñaamolo	Bao Diop 2013	si-gil / i-gil	l
Cobiana	Wilson 2008	si-ggih / ɲi	h
Casanga	Wilson 2008	si-gir / ga / ɲi	r
Basari	Ferry 1991	ɑ-ngɔ̀s	s
Tanda	Wilson 2008	a-ngəz / ba	z
Bedik	Ferry 1991	gí-ngùs	s
Bapen	Ferry 2006	a-ngʌs / ɓa-	s
Konyagi	Santos 1996	i-nkɔ́r / vì-nkɔ́r	r
Biafada	Wilson (p.c.)	(w)-gəra / maa-g.	r
Badiaranke	Meyer 2001	maasɛ	s
Palor	d'Alton 1987	ʔil	l
Saafi	Mbodj 1983	has	s
Noon	Williams & Williams 1993	kʷas	s
Laala	Pichl 1981	kɔʔas	s
Wolof	Fal, Santos & Doneux 1990	bət	t

[6] La ligne horizontale dans le tableau sépare les langues du groupe Bak (au-dessus) des langues du groupe Nord (au-dessous).

Pular	Bah 2009	yiitere	t
Peul Adamawa	Tourneux 1999	yitere / gite	t
Peul Massina	Seydou 2014	yitere / gite	t
Sereer	Crétois 1973/77	ngid a...al / kid a...ak	d
Baga Mboteni	Ferry (p.c.)	cír / si-kyír	r
Nalu	Seidel 2013	cet	t
Baga fore	Golovko 2015 (p.c.)	cét / cí-ncét-él	t

Comme on le voit, cette correspondance n'est pas triviale, puisqu'on y trouve les éléments suivants : Ø, d, h, l, ɬ, n, r, s, ʃ, t, z.

La série 'œil' est la plus complète. Cependant, d'autres séries réduites présentent les mêmes correspondances, ou plutôt des correspondances compatibles :

LANGUE	C#	'MORDRE'	'INTESTIN'	'TROIS'
Bijogo-kagbaaga	O			
Joola fogny	l			
Joola kasa	ɬ			
Joola banjal	l			
Joola kwaatay	n			
Karon	n			
Bayot	o			
Manjaku bok	s		pəs	
Manjaku de Bassarel	s		pəs	
Pepel	l		pʊl	
Mancagne	ʃ		pəʃ	
Balante ganja	t		m-mbùté	
Balante fraasɛ	t		b-mbutɛ	
Balante kentohe	t		kəmbute	
Gubaher	l	bʊŋal		lall
Baynunk guñaamolo	l			lall
Cobiana	h	ŋah		
Casanga	r	ŋar		
Basari	s	a-ɣ̃ás		tàs
Tanda	z			
Bedik	s	u-ŋás		tás
Bapen	s			tas
Konyagi	r	i-ŋɛ̀r		tɛ̀r
Biafada	r	ŋar-	bu-bur	
Badiaranke	s	ŋas		
Palor	l			
Saafi	s			
Noon	s			
Laala	s			
Wolof	t	matt	butit	
Pular	t	ŋat		tat
Peul Adamawa	t	ŋat		tat
Peul Massina	t	ŋat		tat
Sereer	d		fud o...ol	tad
Baga Mboteni	r	ŋérà	nio-búrùk	
Nalu	t	ŋat		paat
Baga fore	t			tet

Ce qu'on sait des correspondances possibles grâce à la série 'œil' va permettre de rechercher des cognats pour les séries incomplètes, et/ou d'éliminer des faux cognats. Ainsi, pour la série 'mordre', j'avais dans un premier temps intégré une racine *ŋar* pour les langues du groupe Joola, avec le sens 'prendre'. Le changement sémantique supposé 'prendre' <> 'mordre' paraissait plausible. Mais la consonne finale ne s'accordait pas avec celle prévue par la série 'œil'. Celle-ci aurait dû être *l*. La série 'œil' m'a donc incité à rechercher une forme *ŋal* dans le groupe Joola. Cette forme existe en joola fogny et en banjal-gusilay avec le sens de 'mâchoire' et en joola kasa avec le sens de 'joue'. La variation sémantique 'mordre' <> 'mâchoire (~ joue)' est assez plausible elle aussi. On a ainsi découvert une racine vraisemblablement proto-atlantique, avec une probable innovation sémantique en joola.

Le module de reconstruction intègre un grand nombre de fonctionnalités, parfois expérimentales, dont certaines sont encore en cours de développement à l'heure où ces lignes sont écrites. En voici quelques-unes :

- Au moment de nommer une correspondance après avoir procédé à l'alignement des cognats, le bouton 'search' (visible sur la figure 3 ci-dessus p. 207) doit permettre de trouver, parmi les alignements déjà enregistrés, ceux qui sont plus ou moins compatibles avec l'alignement en cours.
- Plusieurs alignements différents peuvent être proposés sur un même cognate-set, ce qui permet de tester différentes hypothèses.
- Un cognate-set est par défaut conçu comme une étape vers la reconstruction, via l'alignement. Mais d'autres options sont possibles, qu'il serait trop long de détailler ici.
- Un petit module permet de proposer aux autres utilisateurs d'un dataset partagé de modifier un cognate-set, par l'ajout ou la suppression de fiches. Ces propositions de modification peuvent être acceptées ou non par le destinataire, et/ou donner lieu à des discussions.
- Pour faciliter le travail d'alignement sur de grandes quantités de cognate-sets, leur affichage peut être limité à ceux dont l'alignement n'est pas encore enregistré.

Conclusion

La reconstruction phonologique et lexicale est un domaine ou la linguistique africaine est souvent considérée comme déficiente. Pendant longtemps, il était d'usage d'invoquer le manque de données, ou la difficulté d'accès aux données. Le projet RefLex vise justement à mettre un terme à ces obstacles. Aujourd'hui les données sont très nombreuses : on peut estimer que l'ensemble des dictionnaires et lexiques publiés concernant les langues africaines contient au moins 10 millions de mots, pour au moins 1000 à 1500 langues. RefLex permet d'accéder très facilement à une partie non négligeable de ce corpus, et l'ensemble des données accessibles est appelé à continuer à croître. Même si de nombreux groupes de langues sont encore insuffisamment représentés pour pouvoir envisager une reconstruction sérieuse (notamment dans les groupes

Mande, Gur, Kwa, Kru), les possibilités et le confort d'utilisation proposés par RefLex devraient inciter les linguistes intéressés par la reconstruction à y intégrer davantage de données. Déjà, le travail sur les langues Atlantiques et en bonne voie.

Par ailleurs, l'équipe de DDL à Lyon est en train d'élaborer une base de données sur les langues amérindiennes, qui a la même structure que RefLex et dispose donc des mêmes outils. D'autres initiatives similaires, concernant d'autres régions du monde, sont possibles et bienvenues.

Bibliographie

Barry A., 1987, *The Joola languages: subgrouping and reconstruction*, London, School of Oriental and African Studies (SOAS), University of London (PhD thesis).

Carlton E. M. & Rand S. R., 1994, Compilation de listes de mots Swadesh modifiées recueillis parmi les langues du groupe Diola, sud du Sénégal, *Journal of West African languages* 24-2, p. 85-119.

Fresco E. M. 1970, *Topics in Yoruba dialect phonology*, Los Angeles, African Studies Center & Department of Linguistics, University of California (UCLA).

Sambou P., 2007, *Morphosyntaxe du joola karon*, Dakar, Université Cheikh Anta Diop (Thèse de doctorat de 3e cycle).

Snider K. L., 1989, *North-Guang comparative wordlist: Chumburung, Krachi, Nawuri, Gichode, Gonja*, Legon, Institute of African Studies, University of Ghana.

Wilson W. A. A., 2007, *Guinea Languages of the Atlantic Group*, Frankfurt, Peter Lang.