Manuel de saisie

RefLex

Version 1 : mai 2011

Sommaire

Rappel

Le projet

Le projet RefLex a pour but de tester un ensemble d'hypothèses fondamentales sur la structure et l'évolution des langues d'Afrique qui sont souvent répétées dans la littérature, mais dont la validité n'a jamais pu être démontrée de façon concrète. Parmi ces hypothèses figurent, entre autres, l'existence présumée d'africanismes phonologiques, morpho-syntaxiques et lexicaux (Heine & Zelealem (2008), Meeussen (1975)), l'hypothèse selon laquelle la morphosyntaxe des langues Niger-Congo est fortement influencée par les contraintes prosodiques sur les thèmes verbaux et nominaux (Hyman 2004) et les hypothèses sur la classification génétique des langues africaines (notamment Greenberg 1963). Ces hypothèses ont en commun qu'elles ne peuvent être testées que par une approche quantitative, qui quant à elle suppose l'existence d'une documentation assez complète. Or, une minorité des langues africaines a fait l'objet d'une étude descriptive approfondie à l'heure actuelle. Le projet RefLex est né du constat qu'il existe des données lexicales pour environ les deux tiers des langues africaines et que cette richesse de données, aujourd'hui dispersées et souvent difficiles d'accès, est largement sous-exploitée.

L'objectif est de créer un corpus de données lexicales exhaustif sur les langues d'Afrique, ainsi qu'un ensemble d'outils pour les exploiter. La création du corpus sera un véritable travail collectif, dans lequel chaque collaborateur apporte les données lexicales des langues de sa spécialité. En retour, chaque collaborateur aura à sa disposition l'ensemble des données lexicales, standardisées, fiables, manipulables et exploitables pour des buts scientifiques concrets. Tous les spécialistes des langues africaines seront invités à fournir des outils et des données lexicales à RefLex et, évidemment, à exploiter cette ressource dans leurs propres recherches.

Grâce à son approche novatrice, le projet RefLex résout bon nombre des problèmes méthodologiques auxquels se heurtent d'autres projets comparables. D'une part, **toutes les sources bibliographiques qui constituent la base de données RefLex seront accessibles aux utilisateurs sous forme numérique** (par exemple au format PDF), de façon à ce que tout le monde puisse vérifier la fiabilité des données saisies, signaler d'éventuelles erreurs, mais surtout reproduire des mesures expérimentales à partir de données fiables. Le corpus lexical est donc conçu comme un véritable lexique de référence (d'où le nom *RefLex* de *Reference Lexicon*). D'autre part, nous procèderons à une standardisation des données. Il est donc prévu d'adopter des règles de transcription rigoureuses, qui auront pour effet de lisser les variations dues à l'hétérogénéité des sources utilisées, et qui permettront des comparaisons directes de documents de natures très diverses. Enfin, la manipulation et l'exploitation des données seront optimisées grâce au développement et à la mise à disposition de nombreux outils. La mise en commun des spécifications techniques permettra à chacun des participants de développer ses outils spécifiques, dont la communauté entière pourra ensuite tirer profit. Ainsi, outre le corpus proprement dit, le site web de RefLex proposera une véritable bibliothèque d'outils généraux ou spécifiques.

Le corpus RefLex se distinguera par sa taille sans précédent. Par principe, il n'est pas prévu de limiter le nombre des documents qui ont vocation à intégrer le corpus.

Le corpus: une base de données structurée

Une base de données est constituée de plusieurs tables qui peuvent être visualisées séparément sous forme de tableau, voir l'aperçu en Annexe 1. Ces tables sont reliées entre elles par certaines données qu'elles partagent. L'information lexicale proprement dite est rangée dans une table appelée "Lexiques". Elle 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, la classe nominale ou le genre, la classe verbale, 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. Pour sa part, le sens donnera lui aussi lieu à des traitements variés, qui nécessiteront plusieurs champs.

Logiciels

Il n'y a pas besoin d'installer un logiciel spécifique à RefLex pour effectuer la saisie. Nous recommandons de travailler dans Excel ou Toolbox, pour lesquels des protocoles de traitement préalables à l'intégration dans RefLex existent déjà. Ces outils permettent une visualisation et une correction des données plus pratique et plus intuitive.

Excel

Excel gère des tableaux présentés sur des "feuilles" qui font partie d'un classeur. Un fichier Excel (".xls") est un classeur. RefLex met à disposition une feuille où les noms de champ (dans Excel, les titres des colonnes) sont préremplis dans le fichier : **Template.xls** (téléchargeable sur le site[1]). Ce fichier est un classeur dont les feuilles correspondent aux différentes bases du projet RefLex. Celle qui nous intéresse ici est la feuille "lexique".

Toolbox

La saisie peut être faite dans Toolbox, il suffit de créer une fiche modèle en utilisant les champs RefLex (ex "\frm", "\fun", "\fsi", etc...). Il faut veiller à ce que toutes les fiches aient le même nombre de champs pour faciliter l'exportation, même si cela implique un certain nombre de champs vides dans chaque fiche. Il suffira pour cela de commencer par créer un modèle de fiche. Un fichier toolbox vide avec les paramètres nécessaires sera prochainement disponible.

On pourra se reporter au manuel disponible sur le site du LLACAN[2] pour une présentation plus détaillée.

[1] http://sumale.vjf.cnrs.fr/Lexiques/reflex/Template.xls

[2] http://llacan.vjf.cnrs.fr/fichiers/manuels/Shoebox/ToolboxCadre.htm

Clavier & police

La diffusion et l'hétérogénéité des données de RefLex nécessitent l'utilisation du standard Unicode pour l'encodage des caractères. Il existe différents moyens pour faciliter la saisie en respectant ce format, notamment les claviers générés par le logiciel Microsoft Keyboard Layout Creator (MSKLC). Le LLACAN en propose un exemplaire pour les claviers AZERTY adapté aux conventions des africanistes, nommé "AFU" et librement téléchargeable[3], ainsi qu'un manuel d'utilisation du logiciel MSKLC[4]. Pour les claviers QWERTY, la SIL propose un clavier sur le même modèle, plus complet mais aussi plus complexe[5].

D'autre part, un certain nombre de polices, conformes au standard Unicode, couvre la gamme des caractères utilisés dans la majorité des documents. On peut citer : DejaVu[6], Doulos SIL[7], Charis SIL[8], Code2000[9].

[3] http://llacan.vjf.cnrs.fr/fichiers/Chanard/reflex02.zip

[4] http://llacan.vjf.cnrs.fr/fichiers/manuels/Internet/SaisieClavier.pdf

[5] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=UniIPAKeyboard

[6] http://sourceforge.net/projects/dejavu/files/dejavu/2.32/dejavu-fonts-ttf-2.32.zip/download

[7] http://scripts.sil.org/cms/scripts/render_download.php?&format=file&media_id=DoulosSIL4.106b.zip&filename=DoulosSIL4.106.zip

[8] http://scripts.sil.org/cms/scripts/render_download.php?&format=file&media_id=CharisSIL4.106b.zip&filename=CharisSIL4.106.zip

[9] http://www.code2000.net/code2000_page.htm

Vue générale

Dans le projet RefLex, une source est définie comme une association langue-lexique. Ce sont les lexiques/listes de mots/dictionnaires que l'on trouve dans les ouvrages de linguistique, publiés ou non. Ils peuvent être de première main (rassemblés suite à une enquête de terrain) ou non (repris d'un autre ouvrage). On a dans un ouvrage un ou plusieurs lexiques se rapportant à une ou plusieurs langues et chaque lexique extrait de l'ouvrage représente une source.

Par exemple **Moñino, Yves** (1995), *Le proto-gbaya: essai de linguistique comparative historique sur vingt-et-une langues d'Afrique centrale*, contient 21 lexiques issus d'enquêtes de l'auteur et 1 d'une reconstruction, on a donc 22 sources pour cet ouvrage. **Bühnen, Stefan** (1988), *Lexique comparatif des dialectes Baɲun et de Kasanga et Cobiana*, reprend 2 listes de mots publiées dans **Basso Marques, João** (1947), *Aspectos do problema da semelhança da língua dos papéis, manjacos e brames*, en complément des lexiques collectés dans ses propres enquêtes. Ces lexiques peuvent être repris d'une publication antérieure que nous possédons (c'est le cas ici), mais aussi d'un manuscrit ou d'un ouvrage introuvable. Il y a 11 lexiques dans Bühnen 1988 (dont les 2 repris de Basso Marques) soit 11 sources, et on aura par ailleurs 5 sources issues des 5 lexiques de Basso Marques 1947 (dont les 2 repris par Bühnen), soit, au total pour les deux ouvrages, 16 sources distinctes dans RefLex.

Une autre table de la base de données, nommée "Sources", reprend l'ensemble des métadonnées des lexiques saisis ou à saisir. Les informations présentes, dues à l'auteur, reportées dans les différents champs de la feuille de saisie sont les "champs originaux", ceux qui sont ajoutés pendant la saisie sont les "champs ajoutés" (respectivement COR, "Champ d'ORigine" et CHP, "CHamPs renseignés" de la table "Sources"). Cette répartition est évidemment différente en fonction de la source traitée.

La saisie doit être effectuée dans l'objectif de ne pas perdre d'information par rapport à la source.

Champs originaux

La source la plus sommaire doit permettre de remplir au moins 2 champs dans RefLex : FRM (pour FoRMe) & TRA (pour TRaduction de l'Auteur). Ces champs, systématiquement présents, doivent être reproduits à l'identique, tant pour les symboles utilisés que pour la ponctuation. La question de la répartition de l'information entre les champs ainsi que les exceptions aux principes généraux seront abordées dans le détails des champs.

D'autres champs sont réservés aux informations de la source, le champ CGO (pour Catégorie Grammaticale Originale) et le champs COA (COmmentaire de l'Auteur), qui doivent être copiés fidèlement.

Souvent la source fournit la matière à des champs supplémentaires : champs CLS (pour Classe du Singulier) et CLP (Classe du Pluriel), CLV (pour CLasse Verbale), PLU (pour PLUriel), RAC (pour RACine) et ALT (pour ALTernance consonantique). Ces champs peuvent être remplis par l'auteur de la saisie. S'ils n'apparaissent pas dans le champ COR de la table des sources (voir ci-dessus), cela signifiera qu'ils ont été ajoutés lors de la saisie. Ils pourront de toute façon faire l'objet d'un traitement ultérieur.

Champs à renseigner

Au cours de la saisie, certaines informations doivent impérativement être renseignées, qu'elles soient redondantes par rapport à la source ou indicatives. D'autres sont optionnelles et dépendent de la précision de la source.

De façon générale, il est souhaitable que le plus grand nombre possible de champs soit rempli. L'important est de repérer systématiquement, dans la table "Sources", les informations reprises du

document original et celles ajoutées lors de la saisie. Cependant, le format de RefLex permet d'accueillir des documents même incomplets qui pourront être modifiés par la suite.

Informations minimales

1. Le champ REF (pour REFérence) : correspond au numéro de page où la donnée figure dans la source. Obligatoire pour chaque enregistrement.

2. Le champ CGR (CatéGorie Grammaticale) : obligatoire pour chaque enregistrement. Si le champ CGO a pu être recopié de la source, il s'agit simplement de faire correspondre les catégories de la source, parfois propre à l'auteur, aux catégories unifiées de RefLex (la liste provisoire est donnée en Annexe 2). Si la source ne permet pas de choisir l'une des catégories unifiées de RefLex, on peut s'inspirer d'autres indications (traduction, indication de classe, de pluriel...).

3. Le champ TUF (Traduction Unifiée en Français) peut-être aussi renseigné à partir de la traduction de l'auteur (TRA) si elle est elle-même en français. A défaut, tout dépend des compétences de celui qui effectue la saisie dans la langue de traduction (TRA) ou en français ; des outils seront mis en place pour calculer la TUF à partir d'une TRA en anglais.

4. Le champ FUN (pour Forme UNifiée) : automatiquement généré à partir de la FRM (forme originale) par remplacement systématique des symboles utilisés par l'auteur, afin d'utiliser les symboles standard de RefLex. NB : il est parfois difficile d'établir la correspondance systématique.

Si possible

Les champs CLS (pour Classe du Singulier) et CLP (Classe du Pluriel), CLV (pour CLasse Verbale), PLU (pour PLUriel), RAC (pour RACine) et ALT (pour ALTernance consonantique) doivent être remplis si la source le permet.

Les champs d'emprunts peuvent être remplis lors de la saisie, en fonction des indications données dans la source. Celles-ci apparaîtront en l'état dans le champ COA (COmmentaire Auteur), mais les champs EML (pour EMprunt, Langue), EMF (pour EMprunt, Forme) et EMT (pour EMprunt, Traduction) sont ajoutés lors de la saisie.

Documents électroniques

Si l'on dispose d'un lexique déjà saisi, le plus important est d'établir des correspondances entre les champs et les catégories du document original et les champs et les catégories RefLex. Cette opération n'est pas différente de la saisie d'une source papier, mais elle peut parfois se faire au moyen de procédures automatisées. Comme pour les informations présentes dans une source papier, celles qui se trouvent dans le document électronique seront les "champs originaux".

De manière générale, il faut systématiquement fournir avec ce genre de source un tableau des correspondances entre les champs.

Détails des champs

NB: on trouvera un rappel dans le document Template.xls, feuille "HELP".

Champs automatiques

n (numéro)

C'est un champ qui ne sera pas exporté vers la base de données générale (reflex), il permet de conserver l'ordre des entrées dans le fichier de saisie.

id (identifiant)

Identifiant numérique unique (clé primaire) attribué automatiquement au moment de l'exportation vers la base de données RefLex à chaque entrée du lexique.

Champs de liaison entre tables

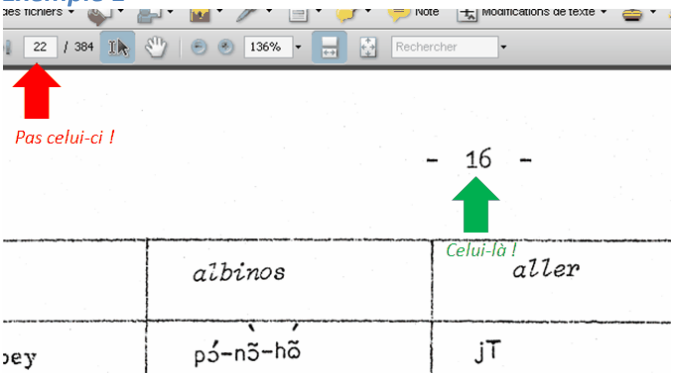
SOU (SOUrce)

Identifiant numérique de la source, il est accessible sur le site internet de RefLex et doit figurer devant chaque entrée du lexique saisi (c'est une clé étrangère).

REF (page)

Le numéro de la page où figure l'entrée dans le document original, quelque soit sa forme (chiffres arabes, romains, lettres...). Ces chiffres ont pu être écrits à la main sur le document au moment de la numérisation, mais doivent être lisibles & doivent être reproduits à l'identique. Il ne s'agit donc pas du numéro de page qu'affiche le logiciel de visualisation des documents (Acrobat Reader le plus souvent) :

Exemple 1



ORP (Ordre dans la Page)

Position de la fiche dans la page. Cette information permettra de retrouver l'ordre des fiches dans le document original.

ORF (Ordre dans la Fiche)

Par défaut, égal à 1. Lorsqu'une fiche de l'original a été scindée en plusieurs fiches pour RefLex (par exemple lorsqu'il y a plusieurs sens pour une forme), cette information permet de restituer l'organisation interne de la source originale.

Champs lexicaux

FRM (FoRMe originale)

Comme précisé plus haut, il s'agit de la forme donnée de l'item dans la source, recopiée à l'identique.

Cela peut donner lieu à des hésitations malgré tout, en particulier dans le cas de la synonymie, quand 2 léxèmes figurent dans une entrée lexicale. Soit les 2 mots sont très différents et on en fera 2 items distincts (en recopiant simplement les différents champs de l'un à l'autre), soit ils sont proches "à vue d'oeil" et on décidera d'en faire des variantes. Dans le dernier cas, on entrera les 2 mots dans le même item en utilisant le même séparateur que dans la source (virgule, esperluette,...). Pour affirmer que 2 mots sont proches, il faut généralement qu'ils aient au minimum une syllabe (attention aux préfixes) en commun, mais ils peuvent avoir aussi toutes les voyelles (ou consonnes) en commun et des consonnes (ou voyelles) proches (un trait de différence). Le pluriel, s'il est donné, est à entrer dans le même item.

Exemple de variantes :

Exemple 2

SOU	REF (pag	FRM
1067	3	mo̍re & mo̍re
1067	3	mbɑ̃̄ni & mbɑ̃̄n
1067	3	ntseɑ̃̄ru & dṣeɑ̃̄ru

cas plus discutable, traité comme des variantes :

Exemple 3

SOU	REF (pag	FRM
1067	17	n·ūán·ūā̃tu & móān·go moā̃tu

(la forme unifiée est éclairante : **ŋuuáŋuuáátu ~ móaango moáátu**)

Exemple de synonymie, dans la source on a :

ń̍họ̀k & mo̍e̍

et dans le fichier de saisie :

Exemple 4

n	id	SOU	REF (pag	FRM
5211		1079	3	ń·họk
5212		1079	3	mo̍e̍

FUN (Forme UNifiée)

C'est le champ qui permet en premier la comparaison, il est ajouté à partir de la FRM en remplaçant systématiquement les symboles utilisés par l'auteur pour la transcription par les symboles utilisés dans RefLex, s'ils sont différents.

Exemple de remplacements systématiques :

Exemple 5

FRM	FUN
dṣîdṣi	**cííci**
ḛ́ba	**ɛ́ɛ́ba**

mais on peut avoir aussi :

Exemple 6

FRM	FUN
ɓótʌ̀ɓà	**ɓótʌ̀ɓà**
a-ɓòxóx	**a-ɓòxóx**

Dans ce champ les variantes sont séparées par un tilde entre blancs et les deux formes correspondant à deux nombres (ou alternances) différents par une barre oblique entre blancs (au contraire de la FRM où on conserve le caractère de séparation de l'auteur) :

Exemple 7

FRM	FUN
onîwu & ōníbu	**onííwu ~ ooníbu**

et

Exemple 8

FRM	FUN
bőka, pl. áōka	**bóóka / áooka**

mais aussi

Exemple 9

FRM	FUN
liőwu & liőbu, pl. aőwu	**lióówu ~ lióóbu / aóówu**

FSI (Forme SImplifiée)

La Forme SImplifiée doit tendre vers la base morphologique ; on reprend la FUN, sans les affixes signalés comme tels dans la source, ni les tons. On ne gardera qu'une forme en cas d'alternance par exemple, ou de réduplication, mais pas pour les mots composés. On aura :

Exemple 10

FRM	FUN	FSI
a-dʸòtárà	**a-ʄòtárà**	**ʄotara**

et

Exemple 11

FRM	FUN	FSI
-dám / -rám / -ndám	**-dám / -rám / -ndám**	**dam**

ou

Exemple 12

FRM	FUN	FSI
ɗám-ɗám	**ɗám-ɗám**	**ɗam**

mais

Exemple 13

FRM	FUN	FSI
dʸák-sʸapàr	ʃák-ʃapàr	ʃak-ʃapar

ou

Exemple 14

FRM	FUN	FSI
dʸàmb ɛ-ngèlá	ʃàmb ɛ-ngèlá	ʃamb ngela

COD (CODe)

Ce champ peut-être calculé à partir de la FSI.

Chaque phonème est remplacée par un symbole de classe, son CODe, défini d'après les principaux ordres phonologiques : les consonnes sont groupées en labiales (symbole P), dentales/coronales (symbole T), palatales (symbole C) et vélaires/uvulaires/glottales (symbole K). En cas de double articulation, le choix est laissé à l'auteur de la saisie, mais les choix suivants sont néanmoins recommandés : labio-vélaires : symbole P ; glottalisées : symbole correspondant au point d'articulation non glottal (ɓ > P, etc.). Pour les voyelles, on distingue trois classes représentées par les symboles A (voyelles centrales), I (voyelles antérieures) et U (voyelles postérieures).

STO (Schème TOnal)

Le Schème TOnal de chaque item est extrait automatiquement de sa FUN, mais il peut être saisi, notamment si l'auteur préfère noter une forme (FRM) sans ton. Il peut aussi être présent dans la source en tant que tel, et être apparemment différent de ce qui figure sur la forme elle-même. Exemple (dictionnaire Punu de J. Blanchon) :

17.	mwâ:li myâ:li	3 / 4	HH	-ali#	-ál+í	-al-		

Ici le schème indiqué est HH et les accents sur la forme suggère une schème descendant. Bien entendu, ce sont les indications de l'auteur (ici HH) qui prévalent. Encore une fois, le fait que le champ STO apparaisse dans la liste des champs d'origine (COR dans la table des sources, voir ci-dessus p. 5) permet de savoir qu'il s'agit bien d'une indication de l'auteur.

CGR (Catégorie GRammaticale)

Pour renseigner la Catégorie GRammaticale de chaque item, on s'aidera en premier lieu du champ CGO et à défaut, on donnera une indication déduite de la traduction, par exemple :

Exemple 15

FRM	CGR	TRA
koyaaməna	N	le fait de se tenir caché,

(si la traduction avait été "se tenir caché", on aurait indiqué "V" en CGR)

Exemple 16

FRM	CGR	TRA
wóy-wóy-wóy	onom	H	1cri de douleur

(ici, la traduction et la forme suggèrent une onomatopée)

S'il y a une CGO, on peut avoir :

Exemple 17

FRM	CGR	CGO	TRA
ree-	gr	monème de conjugaison	aussi utilisé comme monème de conjugaison à l'aspect inaccompli.

ou

Exemple 18

FRM	CGR	CGO	TRA
baw!	id	idéophone	accompagnant le verbe saas- : rire fort

mais aussi :

Exemple 19

FRM	CGR	CGO	TRA
uwe	pr	i.	u he, him, she, her
kwekwekwe	adv	i.	e heartily, completely
hánu	loc	i.	n here

(l'auteur explique dans l'introduction que le "i." signifie "invariable", qui n'est pas une catégorie de RefLex ; on a donc attribué aux items d'autres catégories dans CGR en tenant compte de l'information fournie par la CGO). Pour certains types de mots n'entrant pas dans les grandes catégories du nom ou du verbe, on a un certain choix : les adjectifs par exemple peuvent être codés comme **Adj** (si on a de bonnes raisons de penser qu'il s'agit bien d'adjectifs) ou comme qlt (pour 'qualitatif', pour signaler qu'il s'agit d'un mot permettant d'affecter une qualité, sans plus de précision) ; d'autres catégories sont basés sur des traits sémantiques, pour ne pas faire d'hypothèse en l'absence de critères morphosyntaxiques explicites : **qnt** ('quantitatif', notamment pour les mots dont la valeur est 'tou', 'aucun', 'rien', 'beaucoup', 'plusieurs', etc.), **loc** ('locatif'), **temp** ('temporel'). Il n'est pas rare de trouver dans les sources des entrées constituées non pas de mots, mais d'expressions complètes : impératifs, salutations, formules plus ou moins figées, termes pour 'oui', 'non', 'merci', etc. Pour ces cas, on utilisera l'abréviation **exp** ('expression').

CGO (Catégorie Grammaticale Originale)

Comme précisé plus haut, la Catégorie Grammaticale Originale doit être recopiée à l'identique. Cependant, elle doit parfois être extraite de la traduction donnée dans la source, comme dans l'exemple 17 ci-dessus, ou encore :

Exemple 20

FRM	CGR	CGO	TRA
-ao	gr	affixe de dérivation verbale	affixe de dérivation verbale, ayant un sens passif

SCS (SChème Syllabique)

Le SChème Syllabique peut être calculé à partir de la FUN ou de la FSI.

CLS & CLP (Classe du Singulier & Classe du Pluriel)

La CLasse nominale du Singulier et la CLasse nominale du Pluriel sont données par la source (généralement sous forme numérique pour les langues Bantu, elle peuvent aussi être identifiées par un marqueur de classe). Il arrive qu'on ne sache pas si on a affaire à une classe du singulier ou du pluriel, on utilisera le champ CLS par défaut, de même pour les langues où les genres sont masculin (masc.), féminin (fem.) et neutre (neutre). On aura :

Exemple 21

FRM	CGR	CGO	CLS	CLP	TRA
omuhoozi	N	n.	1		ho avenger
abahoozi	N	n.		2	ho avengers
eshûngwe	N	n.	9	10	sh awn(s) on grass

ou

Exemple 22

FRM	CGR	CGO	CLS	CLP	TRA
ɛ-gɔ́bʌ́r	N		7	8	2 tas de pierre

CLV (CLasse Verbale)

On procèdera de la même façon pour les CLasses Verbales, en reproduisant les indications de la source.

PLU (PLUriel)

Le nombre (sg, pl, du...) pourra être précisé s'il n'est pas donné par la classe.

SCM (SChème Morphologique)

Le SChème Morphologique peut être calculé à partir de la FUN ou de la FSI.

SCC (SChème Consonantique)

Le SChème Consonantique peut être calculé à partir de la FUN ou de la FSI. Il s'agit d'une copie du champ de départ (FUN ou FSI) où toutes les voyelles sont remplacées par le signe '_'.

SCV (SChème Vocalique)

Le SChème Vocalique peut être calculé à partir de la FUN ou de la FSI. Il s'agit d'une copie du champ de départ (FUN ou FSI) où toutes les consonnes sont remplacées par le signe '_'.

RAC (RACine)

La RACine est aussi reprise telle que donnée par la source. Par exemple :

Exemple 23

FRM	RAC	TRA	COA
pakasɛ	kas	tronc d'arbre posé sur un cours d'eau pour traverser.	pont kas- traverser une rivière sur un arbre

ALT (ALTernance consonantique)

Le degré d'ALTernance consonantique : 1 (fricatives, continues / lenis), 2 (occlusives / fortis), 3 (prénasalisées) est toujours repris de la source, mais il peut n'être donné que par les différentes FRM, comme dans 0les exemples 24 et 25 :

Exemple 24

FRM	FUN	FSI	ALT	TRA
-nɔngʷ / -lɔngʷ	**-nɔngʷ / -lɔngʷ**	**nɔngʷ**	N 3 / 1	aîné, devant

ou

Exemple 25

FRM	FUN	FSI	ALT	TRA
-gum / -ɣum / -ngum	**-gum / -ɣum / -ngum**	**gum**	gpr 2 / 1 / 3	ça ; ce ; cette; ces

TRA (TRAduction originale)

La TRAduction originale est celle donnée par l'auteur, dans la ou les langues de la source. Elle doit être reprise à l'identique, y compris si l'auteur donne la traduction en plusieurs langues. La seule liberté que l'on peut s'autoriser consiste dans la répartition des informations dans les différents champs quand elles ne sont pas clairement délimitées dans l'original. Par exemple, dans la source on a :

Exemple 26

birin (kobirin): 1. depuis (préposition)
Birin paaki jeen-te-men en.
Depuis hier, je ne t'ai pas vu.
2. depuis que, dès que (conjonction de subordination, employée avec **wen**.)

et dans le lexique :

Exemple 27

FRM	CGR	CGO	TRA	COA	
birin / kobirin	temp	préposition	depuis	depuis	
birin / kobirin	conj	conjonction de subordination	depuis que, dès que	dès qu	employé avec wen

De même pour les définitions encyclopédiques :

Exemple 28

de: morphème grammatical employé après les verbes dans plusieurs types d'énoncés, soit à l'aspect accompli, à l'aspect inaccompli ou dans des énoncés de situation. Il est employé pour focaliser l'action ou le procès exprimé par le verbe d'un énoncé.

et dans le lexique saisi :

Exemple 29

FRM	TRA	COA	
de	morphè morphème grammatical [...] Il est employé pour focaliser l'action ou le procès exprimé par le verbe de lénoncé	FOCAL	... employé après les verbes dans plusieurs types d'énoncés, soit à l'asp

ou :

Exemple 30

kacamma: houe, daba
(**samm-** désherber, cultiver)

et dans le lexique :

Exemple 31

FRM	CGR	RAC	TRA	COA	
kacamma	N	samm	houe, daba	houe	samm- désherber, cultiver

ou encore :

Exemple 32

FRM	CGR	TRA	COA
siifa	N	jeu d'enfant [...]	jeu ... qui consiste à donner une tape à son camarade avant de s'en aller en courant

TUF (Traduction UNifiée)

Sur le modèle de la distinction entre FRM et FUN, la Traduction UNifiée paraphrase la TRAduction originale de façon à permettre la comparaison entre les langues. Le but est de faire ressortir des ensembles de même sens grâce à un moteur de recherche ou à un classement alphabétique. Par exemple :

Exemple 33

FRM	CGR	TRA	TUF
ferenteku	N	petite calebasse	calebasse
kaku	N	grande calebasse utilisé pour le lait	calebasse
kalama	N	calebasse-louche	calebasse
keemiran	N	calebasse réservée pour le repas des hommes	calebasse
korun	N	petite calebasse, servant de mesure.	calebasse
masameredin	N	calebasse réservée aux hommes, pour le partage du repas.	calebasse
paɲuɛ	N	petite calebasse.	calebasse
pədandan	N	grande calebasse pleine.	calebasse
pomma(pomm)	N	calebasse.	calebasse
boli	N	calebasse-gourde, en forme de bouteille	calebasse

ou

Exemple 34

FRM	CGR	TRA	TUF
paddan	N	tam-tam.	percussion
tabulee	N	gros tambour.	percussion
taman	N	tambour d'aisselle.	percussion

Toujours dans cette optique, on pourra utiliser l'abréviation "sp.", dans un sens biologique pour désigner une variété par son espèce, notamment pour les TRA qui commencent par "variété de", "espèce de", "genre de", "sorte de", etc... On aura :

Exemple 35

FRM	CGR	TRA	TUF
ciicoor	N	espèce d'oiseau	oiseau sp.
cuurnə	N	calao	oiseau sp.
fedde	N	engoulevent (variété doiseau)	oiseau sp.
kanduudu	N	variété d'oiseau	oiseau sp.

Cependant, on évitera "vertébré sp" ou mammifère sp", trop larges. De façon générale, on adaptera les catégories aux nombres de mots qui en relèvent dans le lexique, en particulier pour les animaux ou plantes très représentés :

Exemple 36

FRM	CGR	TRA	TUF
cisad	N	cobe onctueux	antilope sp.
paɲuron	N	variété d'antilope.	antilope sp.
patoŋoŋə	N	variété de grande antilope.	antilope sp.
patunkə	N	mâle du guib harnaché.	antilope sp.
pawci	N	variété de petite antilope.	antilope sp.
wancafɛ	N	guib harnaché.	antilope sp.

ou

Exemple 37

FRM	CGR	TRA	TUF
basi	N	variété de gros mil	mil sp.
konkosaalo	N	variété de mil tardif.	mil sp.
təponpo	N	variété de mil hâtif.	mil sp.
yeeɲo	N	variété de mil blanc.	mil sp.

à ne pas confondre avec les différents mots qui désignent la même variété, dans des états différents par exemple :

Exemple 38

FRM	CGR	TRA	TUF
kunjinə	N	mil germé cuit et pilé.	mil
təpər	N	mil grillé.	mil
tuuŋi	N	mil qu'on a mis à germer.	mil

D'autre part, les différents sens d'un mot peuvent souvent être regroupés en un seul :

Exemple 39

FRM	CGR	TRA	TUF
kuniinɛ	N	jeu, danse, amusement.	jeu

quand c'est impossible, on séparera les différents sens par un point-virgule et on les classera par ordre alphabétique :

Exemple 40

FRM	CGR	TRA	TUF
safee	N	écrit, amulette.	amulette ; écrit

Les valeurs des différents morphèmes grammaticaux (cf Annexe 3 pour une liste non-exhaustive) seront notés en majuscules :

Exemple 41

FRM	CGR	CGO	TRA	TUF
-a	gr	suffixe	suffixe ajouté au radical verbal qui suit les auxiliaires	SUFFIXE
-a	gr	affixe de dérivation verbale	affixe de dérivation verbale à valeur de voix moyenne	VOIX MOYENNE
a	gr	suffixe de dérivation nominale	suffixe de dérivation nominale à valeur d'agent.	AGENTIF
-aad	gr	affixe de dérivation verbale	affixe de dérivation verbale à valeur d'intensif	INTENSIF
-aan-	gr	affixe de dérivation verbale	affixe de dérivation verbale à valeur de causatif ou de factitif.	CAUSATIF; FACTITIF
-aana	gr	suffixe de dérivation nominale	suffixe de dérivation nominale à valeur d'instrument ou de moyen.	INSTRUMENTAL
-aar-	gr	affixe de dérivation verbale	affixe de dérivation verbale qui confère un sens négatif au radical ver	NEGATIF

Pour les marqueurs de classes, on aura "CL *affixe* (~ *variante(s)*)" :

Exemple 42

FRM	CGR	CGO	TRA	TUF
cen	CL	classificateur nominal	classificateur nominal. Il est surtout utilisé pour les termes d'emprunt.	CL cen
fan	CL	classificateur nominal	classificateur nominal. Il est utilisé poux les noms qui commencent par fa- ou fan-.	CL fa ~ fan
kan	CL	classificateur nominal	Il détermine aussi les infinitifs	CL kan
kan	CL	classificateur nominal	Il détermine un certain nombre de noms qui commencent par les préfixes ka- ou ha-.	CL ka ~ ha
kon	CL	classificateur nominal	classificateur nominal pour les mots qui commencent par ko- ; kon- ; koo-.	CL ko ~ kon ~ koo

Pour les marques personnelles, "*personne nombre (FONCTION)*" :

Exemple 43

FRM	CGR	TRA	TUF
a	pr	tu	2 sg S
an	pr	moi, me, à moi	1 sg O
an	pr	le, la, lui, à elle, à lui	3 sg O
an	pr	son, sa, ses	3 sg P

(cf les abréviations utilisées en Annexe 4)

TUE (Traduction Unifiée - English)

La Traduction Unifiée - English pourra être calculée semi-automatiquement d'après les lexiques complétés précédemment. Cependant, il est souhaitable de la renseigner sur le modèle de la TUF si cela n'allonge pas trop la saisie.

DV1, DV2 & DV3 (DiVers)

Ces champs DiVers seront laissés libres à l'utilisateur. En fonction des informations présentes dans la sources, ils peuvent permettre de ne pas surcharger le champ COA ('COmmentaire Auteur'), mais ils servent également à conserver des informations qui ne se rangent pas 'naturellement' dans l'un des champs disponibles. Par exemple, dans le lexique Punu de J. Blanchon figurent les mentions des séries comparative de Guthrie pour le bantu commun, ainsi que les formes reconstruites. Ces informations sont typiquement de celles que l'on préfèrent mettre dans des champs séparés plutôt que dans le champ COA. Ces champs 'libres' peuvent également accueillir, par exemple, les dénominations scientifiques des espèces végétales ou animales. Idéalement, la mention du type d'information rangé dans les champs DV1, DV2, DV3 doit figurer dans la fiche 'source' correspondante (champ 'commentaire').

EML (EMprunt, Langue)

Il s'agit de la langue source pour les emprunts identifiés. On complètera le champ EMprunt, Langue, d'après la norme ISO 639-3[10], en fonction des informations trouvées dans la source, mais aussi éventuellement d'une appréciation personnelle pour les cas les plus évidents.

Exemple 44

FRM	CGR	TRA	TUF	EML	COA
wuri	N	bouillie à base de riz.	bouillie	mlq	(mal)
wuro	N	parc à vaches, troupeau.	parc	fuc	(pl)
wusun	N	sorte de panier conique utilisé lors de la pêche collective.	panier	mlq	(mal)

(ici pour les langues peul et malinké).

Mais aussi :

Exemple 45

FRM	CGR	TRA	TUF	EML	COA
kipiŋ-	V	s'occuper de, prendre soin de.	s'occuper	eng	(fr)

[10] code *Ethnologue*, accessible sur le site : http://www.ethnologue.com/web.asp

Si l'auteur mentionne une langue source, cette mention doit figurer dans le champ COA. Cette redondance permet d'identifier les cas où c'est l'auteur de la saisie qui a identifié l'emprunt.

EMF & EMT (EMprunt, Forme & EMprunt, Traduction)

Les champs EMprunt, Forme et EMprunt, Traduction contiennent respectivement la forme qui a été empruntée et sa traduction en français. Ici aussi, ces informations, si elles sont fournies par l'auteur, doivent également figurer dans le champ COA.

COA (COmmentaires Auteur)

Ce champ est apparu plusieurs fois dans les exemples donnés pour le détail à propos des champs précédents (comme dans les exemples 23, 27, 29, 31, 32, 44 et 45). C'est dans les COmmentaires Auteur qu'on reproduira à l'identique les informations qui figurent sur la source et que l'on n'utilise pas ailleurs. Dans l'exemple 23, la racine est donnée par l'auteur avec sa traduction, elle est reproduite telle quelle dans COA ; d'une façon générale, toute information qui permet de définir un champ "formaté" (dont le format est prédéfini et ne suit pas celui de l'auteur), est copiée dans COA. D'autre part, on copiera aussi la fin des définitions encyclopédiques, comme dans l'exemple 29 ainsi que les précisions apportées par l'auteur qui n'entre pas dans le cadre de la traduction de l'item, comme dans l'exemple 44. Pour indiquer que la fin d'une traduction se trouve dans COA, on utilise le symbole "[...]" dans TRA :

Exemple 46

FRM	CGR	TRA	TUF	COA
ɲako	N	impureté rituelle qui nécessite une purification rituelle [...]	impureté	... comme la toilette des nouveaux initiés à la fin de leur séjour en brousse. Ou la toilet

CSA (Commentaire SAisie)

Le champ Commentaire SAisie est libre, il peut être utilisé aussi bien pour un commentaire sur le contenu de la source que pour signaler une hésitation lors de la saisie. On peut par exemple signaler ici des informations que seule la saisie permet de connaître, par exemple que tel mot apparaît à plusieurs endroits avec plusieurs sens, ou que la graphie utiulisée est incohérente avec le reste du document. Un cas particulier intéressant est celui des dictionnaires bilingues qui comportent 2 parties, l'une de la langue africaine vers le français (ou l'anglais, etc.), et l'autre dans le sens inverse. En principe, on saisit et on utilise uniquement la première partie. Toutefois, il arrive que certains mots figurent dans le seconde partie mais pas dans la première. Par exemple, dans le dictionnaire limba-anglais / anglais-limba de Clarke (1922), le mot pour 'trois' ne figure bizarrement que dans la partie anglais-limba (p. 142). Evidemment, en ne saisissant que la première partie, on risque de passer à côté de cette information, mais si le hasard fait que l'on en ait connaissance, alors, typiquement, la mention de cette asymétrie doit figurer dans le champ CSA.

DAT (DATe)

C'est la DATe de dernière modification de l'entrée.

Annexe 1. Extrait de la table "Lexiques"

id	SOU	REF	FRM	FUII	FSI	COD	STO	CGR	CGO	SCS	CLS	CLP	CLV	PLU	SCM	SCC	SCV	RAC	ALT	TRA	TUF	TUE	DV1	DV2	DV3	EML	EMF	EMT	COA	CSA	DAT
70575	1	280	kaysɛ / ŋaysɛ	kaysɛ / ŋaysɛ	yisɛ	JICE		N			ka	ŋa							0	arbre sp.	arbre sp.										2005-10-03
70576	1	280	kɔɔda / ŋaɔda	kɔɔda / ŋaɔda	ɔda	ODA		N			ko	ŋa							0	arbre sp.	arbre sp.										2005-10-03
70577	1	280	nakpakpɛ / nkpakpɛ	nakpakpɛ / nkpakpɛ	kpakpɛ	PAPE		N			ne	M							0	arbre sp.	arbre sp.										2005-10-03
70578	1	280	nopota / mpota	nopota / mpota	pota	POTA		N			ne	M							0	arbre sp.	arbre sp.										2005-10-03
70579	1	280	ŋuntaage / muntaage	ŋuntaage / muntaage	ntage	TAGE		N			ŋo	mo							0	arbre sp.	arbre sp.										2005-10-03
70580	1	280	uaga / ŋaaga	uaga / ŋaaga	aga	AGA		N			u	ŋa							0	arbre sp.	arbre sp.										2005-10-03
70581	1	280	ukɛnɛm / ŋakɛnɛm	ukɛnɛm / ŋakɛnɛm	kɛnɛm	KEDEB		N			u	ŋa							0	arbre sp.	arbre sp.										2005-10-03
70582	1	280	ɛmudu / kumudu	ɛmudu / kumudu	mudu	BUDU		N			e	ko							0	arbre sp. (Acacia albida)	arbre sp.										2005-10-03
70583	1	280	kamudu / ŋamudu	kamudu / ŋamudu	mudu	BUDU		N			ka	ŋa							0	arbre sp. (Acacia albida)	arbre sp.										2005-10-03
70584	1	280	kajauti / ŋajauti	kajauti / ŋajauti	jauti	JAUTI		N			ka	ŋa							0	arbre sp. (Acacia sp.)	arbre sp.										2005-10-03
70585	1	280	ɛsumbe / kusumbe	ɛsumbe / kusumbe	sumbe	CUBE		N			e	ko							0	arbre sp. (Alchornea cordifolia)	arbre sp.										2005-10-03
70586	1	280	kɛnsum / ŋaensum	kɛnsum / ŋaensum	ɛnsum	ECUB		N			ko	ŋa							0	arbre sp.(Alstonia congensis)	arbre sp.										2005-10-03
70587	1	280	nodokoŋ / ndokoŋ	nodokoŋ / ndokoŋ	dokoŋ	DOKOG		N			ne	M							0	arbre sp. (Anisophilea laurinea)	arbre sp.										2005-10-03
70588	1	280	ɛbaŋa / kabaŋa	ɛbaŋa / kabaŋa	baŋa	BOJO		N			e	ko							0	arbre sp. (Annona	arbre sp.										2005-10-03

Annexe 2. Les catégories grammaticales

(liste non-exhaustive, à discuter)

adj	adjectif
adv	adverbe
cl	marque de classe
conj	conjonction
conn	connectif
cov	coverbe
def	défini
dem	démonstratif
det	déterminant
exp	expression
gr	morphème grammatical
id	idéophone
int	interrogatif
interj	interjection
loc	locatif
N	nom
NP	nom propre
Nc	nom composé
Nd	nom dérivé
Npl	nom pluriel
num	numéral
onom	onomatopée
part	particule
pr	pronom / marque personnelle
prep	préposition
qlt	qualitatif
qnt	quantitatif
rel	relateur
temp	temporel
V	verbe
var	variante
Vaux	verbe auxiliaire
Vc	verbe composé
Vd	verbe dérivé
Vi	verbe intransitif
Vt	verbe transitif

Exemples de valeurs des morphèmes grammaticaux (aspect, cas,...)

(liste non-exhaustive)

ASSOCIATIF	AUGMENTATIF
BÉNÉFACTIF	CAUSAL
CAUSATIF	CL (pour CLassificateur)
CONTREFACTUEL	DISTRIBUTIF
FACTITIF	FORTE PROBABILITÉ
FUTUR PROCHE	INACCOMPLI
INCHOATIF	INSTRUMENTAL
INTENSIF	INVERSIF
ITÉRATIF	LOCATIF
NÉGATIF	OBLIGATIF
OBLIGATION	ORDRE
PASSIF	PÉJORATIF
PONCTUEL	PROXIMITÉ
RELATIF	RÉMANSIF
SIMULTANÉITÉ	SOUHAIT
TOTALITÉ	VOIX MOYENNE

Annexe 3. Les fonctions des marques personnelles

(liste non-exhaustive)

NB : Pour les marques personnelles, les TUF (Traduction UNifiée) sont de la forme : "***personne nombre (FONCTION)***"

<table>
<tr><td>S : Sujet, dont :</td><td colspan="4">S1 : sujet - aspect/mode 1 (accompli, réel, aoriste)
S2 : sujet - aspect/mode 2 (inaccompli, inachevé, prospectif, virtuel, habituel)
S3 : sujet - aspect/mode 3 (nécessaire, impératif, injonctif, exclamatif)
S4 : sujet - aspect/mode 4 (duratif)</td></tr>
<tr><td>O : objet
P : possessif
T : tonique
R : réfléchi</td><td>s : singulier
p : pluriel
d : duel</td><td>f : féminin
m : masculin
n : neutre
i ou ind. : indéfini</td><td>a ou anim : animé
z ou inan : inanimé</td><td>exc : exclusif
inc : inclusif
L ou Log : logophorique</td></tr>
</table>