Projet DILAF
Dictionnaires Langue Africaine – Français

ORGANISATION INTERNATIONALE DE la francophonie

Méthodologie DILAF

Étape 3 : Choix des noms des éléments XML, structuration

Objectif

Choisir les noms des éléments XML ainsi que la structuration de ces éléments.

Outil

Éditeur de texte Open Office

Motivation

Respecter le standard LMF[1] permettra une utilisation des ressources lexicales par des chercheurs en traitement automatique des langues (TAL) grâce à la normalisation proposée par ce standard.

Pour en savoir plus :
– XML pour les nuls
– Langage XML pour les débutants
– Francopoulo G., Bel N., George M., Calzolari N., Monachini M., Pet M., and Soria C. (2009). Multilingual Resources for NLP in the lexical markup framework (LMF). Language Resources and Evaluation, 43(1), pages 57–70, March.

1 Lexical Markup Framework (en français cadre de balisage lexical).

Méthode

3.1 – Identification et nommage des éléments d'une notice

Les différents éléments composant chaque notice doivent être identifiés et nommés. Ces noms peuvent être exprimés dans n'importe quelle langue mais ils ne peuvent comprendre, notamment, le caractère espace ou l'apostrophe (ceux-ci peuvent être remplacés par un tiret bas "_").

Ce premier ensemble d'éléments, qui repère des informations composant une notice, est nommé "éléments de base".

Exemple :
L'examen des notices du dictionnaire permet d'identifier et de nommer les parties composant les notices. Les noms ont été exprimés en haoussa.
– vedette (exemples : *a, a* ƙ*alla, abarba, babbake*). Nom : kalma
– phonétique (notée entre crochets [et]). Nom : furici
– classe (exemples : *s., s, aik.*). Nom : nau_i
– genre, signalé après "*Jin.:*". Nom : jinsi
– forme au pluriel, signalée après "*Jam.:*". Nom : jam_i
– forme au féminin, signalée après "*Sg.*". Nom : mace
– information morphologique, signalées après "*Morph.:*". Nom : siffolin_kalma
– variante, signalée après "*Yare:* ". Nom : yare
– définition (en gras) . Nom : ma_ana
– exemple d'usage (en italique). Nom : misali
– équivalent français signalé après "*Far.:* ". Nom : makwatanci
Les informations seront regroupées par l'élément de nom : article
Nous remarquons que l'entrée *adadi* présente deux sens.

3.2 – Structuration des éléments composant une notice

Afin de respecter le standard LMF, la structure des notices doit explicitement séparer les informations sémantiques des autres informations. Les informations sémantiques peuvent être regroupées au sein d'un bloc sémantique. Une entrée polysémique présentera donc plusieurs blocs sémantiques.

Cette structuration entraîne la définition d'un "bloc sémantique".

Exemple :
Nous identifions trois informations sémantiques : la définition (ma_ana), l'exemple d'usage (misali) et l'équivalent français (makwatanci). Ces informations peuvent être regroupées au sein d'un bloc sémantique nommé rukunin_ma_ana.
L'entrée "*adadi*" qui a deux sens, présentera, *in fine*, deux blocs sémantiques.

A priori une notice ne définit qu'un seul ensemble d'informations lexicales, morphologiques et phonétiques pour une entrée, puis éventuellement plusieurs blocs sémantiques. Toutefois, il arrive qu'un dictionnaire présente une entrée ayant, par exemple, différentes catégories lexicales. Dans ce cas, l'entrée devra être scindée en plusieurs entrées ayant chacune une seule catégorie lexicale.

3.3 – Identification d'une notice

Dans un dictionnaire certaines entrées homonymes ont plusieurs notices. Il est nécessaire de distinguer ces notices pour bien indiquer d'éventuels liens (de synonymie, d'homonymie, etc.) entre notices. Il est également nécessaire d'identifier les sens d'une notice.

Notices et sens peuvent être identifiés en fabriquant un identifiant unique pour chacun d'eux.

L'identifiant de la notice peut être facilement construit en reprenant l'entrée et en lui adjoignant un numéro. Cet identifiant est inscrit dans l'élément qui encadre l'entrée comme valeur d'un attribut.

De même, l'identifiant du sens peut être construit en ajoutant un numéro à l'identifiant de la notice.

Comme certains caractères (espace, apostrophe, etc.) sont interdits dans les identifiants, il peut être nécessaire de remplacer ces caractères par un tiret bas "_".

Exemple :
Dans notre dictionnaire, deux notices homonymes sont explicitement numérotées 1 et 2 ("*abara*"), d'autres ne le sont pas - mais auraient dû l'être - ("*a*" "*ba*", "*babbake*").
Les identifiants de ces notices seront "*abara1*", "*abara2*".
Les identifiants des notices non homonymes seront simplement construits en ajoutant le chiffre 1 à l'entrée : "*abacada1*", "*abarba1*", etc.
Les identifiants de sens seront construits en ajoutant un tiret bas et un chiffre à l'identifiant de notice : "*abacada1_1*", "*abarba1_1*", etc.
Les entrées utilisant une espace ou une apostrophe sont légèrement transformées pour fabriquer leur identifiant : "*a_kalla1*", "*ba_a1*", "*ba_abzine1*", etc.
Un identifiant (de notice ou de sens) sera noté comme valeur de l'attribut de nom : id.