Projet DILAF
Dictionnaires Langue Africaine – Français

ORGANISATION INTERNATIONALE DE la francophonie

Méthodologie DILAF

Étape 2 : Conversion d'un fichier vers Unicode

Objectif

Obtenir un dictionnaire respectant le standard de codage Unicode.

Outil

Éditeur de texte Open Office

Motivation

L'écriture de nombreuses langues utilise des caractères spéciaux, c'est-à-dire des caractères absents des alphabets des langues européennes. C'est le cas, par exemple, du o ouvert (ɔ, Ɔ) du bambara et du dioula, du k crossé (ƙ, Ƙ) du haoussa, ou encore du schwa (ə, Ǝ) du haoussa et du tamajaq.

Depuis 1993, le standard de codage Unicode a défini un codage unique pour chacun des caractères, ce qui permet une bonne utilisation des logiciels usuels (comme les éditeurs de textes ou les tableurs).

De nombreuses ressources linguistiques, y compris des dictionnaires, ont été ou sont produites en utilisant des polices incompatibles avec ce standard : les caractères sont mal codés, il faut les modifier afin que le dictionnaire respecte le standard Unicode.

Exemple :
Le dictionnaire haoussa utilise la police "Niger3 SILCharis". Dans cette police, le k

crossé ƙ est dessiné à la place du caractère ø, le d crossé ɗ à la place de ñ.
Si le lecteur dispose de cette police, le texte est lisible (par exemple : il lira la phrase "A ƙalla mutum ɗari sun taru ƙofar gidan sarki.",
sinon il verra "øalla mutum ñari sun taru øofar gidan sarki."
En revanche, pour l'ordinateur, le texte reste "øalla mutum ñari sun taru øofar gidan sarki.", par conséquent l'utilisation de polices redessinées interdit tout traitement automatique des langues.

Pour en savoir plus :
Enguehard, C. Les langues d'Afrique de l'Ouest : de l'imprimante au traitement automatique des langues, Sciences et Techniques du Langage, 6, p.29-50, 2009. (ISSN 0850-3923)

Méthode

2.1 – Déterminer la liste des caractères de la langue

Créer un fichier texte et y faire figurer tous les caractères du système d'écriture de la langue traitée.

L'alphabet (ou le syllabaire) nécessaire à l'écriture d'une langue peut être spécifique à un pays. Il est nécessaire de consulter les décrets en la matière.

Les caractères ainsi que leur code peuvent être trouvés dans le répertoire d'Unicode.

Chaque caractère est présenté dans sa version minuscule et majuscule.

Voir la présentation de l'alphabet haoussa du site web DILAF.

2.2 – Recensement des caractères à modifier

Objectif : Établir la liste des caractères à modifier

Méthode : Les caractères composant le dictionnaire sont observés un à un afin de détecter lesquels sont mal codés.

Astuce :
Le dictionnaire utilise peut-être une police qui respecte le standard Unicode. Penser à le vérifier, si c'est le cas, l'objectif de l'étape 2 est déjà atteint !

2.2.1 – Création d'un fichier de travail

Sauvegarder le dictionnaire dans un fichier de travail (ce dernier sera dégradé lors des traitements).

Pas à pas :
hausa_extrait_v2.odt est sauvegardé sous le nom hausa_travail.odt

2.2.2 – Recensement des caractères

Il faut observer et recenser les différences entre le dictionnaire affiché en utilisant sa police d'origine et le dictionnaire affiché avec une police conforme à Unicode, par exemple "Times New Roman".

Ces traitements sont réalisés sur le fichier de travail.

Pour chacun des caractères du dictionnaire,

— si l'affichage avec "Times New Roman" transforme le caractère, copier ce caractère dans un fichier, y noter également le caractère qui devrait être affiché, ôter ensuite toutes les occurrences de ce caractère affiché avec la même police de caractères et en respectant la casse (minuscule ou majuscule).

— si l'affichage avec "Times New Roman" ne transforme pas le caractère, ôter toutes les occurrences de ce caractère affiché avec la même police de caractères.

Exemple :
Le dictionnaire haoussa utilise la police "Niger3 SILCharis". Il est examiné et traité ligne par ligne.

* Le texte *"a [áa] harafi na biyu a cikin haruffan hausa. Sunan Ado yana farawa da a. Far.: voyelle a"*, devient *"a [áa] harafi na biyu a cikin haruffan hausa. Sunan Ado yana farawa da a. Far.: voyelle a"* lorsqu'il est affiché avec la police "Times New Roman".
Constat : aucun caractère n'est modifié par le changement de police.
Toutes les occurrences de ces lettres (A, a, á, b, c, d, e, F, f, h, i, k, l, n, o, r, S, s, u, v, w, y) ainsi que des signes de ponctuation (. : [ ]) sont ôtés. Les occurrences peuvent être remplacées par une espace par exemple.
Le fichier de travail devient hausa_travail_v2.odt.

* Le texte *"m m ɗ mt t g"*, devient *"m m ñ mt t g"* lorsqu'il est affiché avec la police "Times New Roman".
Constat : ɗ est remplacé par ñ. Ces substitutions sont notés dans le fichier hausa-remplacer_unicode.odt.
Toutes les occurrences de ces lettres (g, m, ñ, t) sont ôtées.
Le fichier de travail devient hausa_travail_v3.odt.

* Le texte *"ƙ à ƙà ƙ ƙ ƙ"*, devient *"ø à øà ø ø ø"* lorsqu'il est affiché avec la police "Times New Roman".
Constat : ƙ est remplacé par ø. Ces substitutions sont notées dans le fichier hausa-remplacer_unicode.odt.
Toutes les occurrences de (à, ƙ) sont ôtées.
Le fichier de travail devient hausa_travail_v4.odt.

* Le traitement se poursuit jusqu'à ce qu'à ce que tous les caractères aient été passés en revue (le fichier de travail apparaît alors vide).
Quatre caractères ont été répertoriés : ɓ, ɗ, ƙ et -
Ils sont notés dans le fichier hausa_remplacer_unicode.odt.

2.3 – Modification des caractères répertoriés

Les modifications répertoriées sont effectuées sur le dictionnaire. Celui-ci peut être entièrement affiché avec une police respectant le standard Unicode, par exemple "Times New Roman".

Exemple :
Le dictionnaire est ouvert. Les caractères précédemment répertoriés sont remplacés en tenant compte de la police de caractères. Ainsi, les caractères ð, ñ et ø affichés avec la police "Niger3 SILCharis" sont respectivement remplacés par les caractères ɓ, ɗ et ƙ.
hausa_extrait_v2.odt est sauvegardé sous le nom fichier hausa_extrait_v3.odt

Les modifications doivent être effectuées en tenant compte de la police de caractères et de la casse (minuscule ou majuscule) afin d'éviter les remplacements inadéquats.

Lors de la transformation d'un dictionnaire tamajaq, nous avions constaté l'emploi d'une police de caractères dans laquelle la lettre 'p' était redessinée en schwa 'ə' pour les parties en tamajaq, alors qu'une autre police était utilisée pour les parties en français. Effectuer des modications sans tenir compte de la police de caracrères aurait remplacé tous les 'p' par des 'ə' y compris dans le texte en français.

Exemple :
Le dictionnaire est ouvert. Les caractères précédemment répertoriés sont remplacés en tenant compte de la police de caractères. Ainsi, les caractères ð, ñ et ø affichés avec la police "Niger3 SILCharis" sont respectivement remplacés par les caractères ɓ, ɗ et ƙ.