Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Générateur de dictionnaires au format Android pour les langues peu dotées

Rémy Bonnet[1] Guillaume Segerer[1]
(1) LLACAN, 7 rue Guy Môquet - BP 8, 94801 Villejuif, France
remy.bonnet@cnrs.fr, guillaume.segerer@cnrs.fr

Résumé

Nous présentons un outil informatique en source libre permettant, à partir d'un dictionnaire sous format électronique de générer une application Android. L'objectif est de tirer profit de l'augmentation spectaculaire des utilisateurs de mobiles multifonctions dans de nombreuses régions du monde, notamment en Afrique, pour permettre aux locuteurs de langues moins dotées de bénéficier de dictionnaires de bonne qualité, faciles à utiliser et de s'impliquer dans la collecte de données. Cet outil est conçu pour être compatible avec les principaux formats lexicographiques.

Abstract

Dictionary App Generator for Less Resourced Languages.
This article presents an open source dictionary Android app generator project for less resourced langages. If these dictionaries can be useful to linguists our main goal is to take advantage of the dramatic increase in the number of smart-phone owners in many parts of the world and especially Africa in order to allow native speakers to benefit from an easily accessible and good quality version of a dictionary of their language and interact with them linguists who work on it. Our application is meant to adapt to as many formats as possible.

Mots-clés : Dictionnaire, Application Android, Langues peu dotées.
Keywords: Dictionary, Android Application, Less Resourced Languages

1 Introduction

Ces dernières années les linguistes de terrain ont pu, grâce aux avancées des technologies numériques, archiver numériquement leur données et ajouter une composante multi-media au traditionnel triptyque "grammaires, dictionnaires, et recueils de textes" (Musgrave & Thieberger, 2014).

Parallèlement l'usage des téléphones mobiles s'est considérablement répandu et démocratisé dans des endroits du monde où l'usage des ordinateurs reste assez marginal. Ainsi le nombre de souscripteurs à des forfaits internet mobiles a quadruplé lors de la dernière décennie (Poushter). L'immense majorité des téléphones mobiles vendus aujourd'hui sont des mobiles multifonctions.

Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

Le système d'exploitation mobile Android, lancé en 2005 par Google équipe aujourd'hui plus de 80 % des mobiles multifonction (Auffray 2016).

Une application Android semble donc être, pour de nombreuses communautés linguistiques minoritaires, un moyen d'accéder à un dictionnaire qui, outre son intérêt linguistique, est souvent perçu comme un résultat tangible du travail des linguistes et peut constituer une incitation à s'impliquer dans la collecte des données d'autant que les mobiles multifonctions, en tant qu'outils de communication, permettent aux utilisateurs de contacter directement les auteurs.

L'outil présenté ici, dont le développement a été financé par le projet DReaM (site internet) et fait suite à une version embryonnaire développée pour le LabEx EFL sous la supervision de Guillaume Jacques et Alexis Michaud, a pour objectif de permettre aux linguistes ayant un dictionnaire numérique de générer une application Android.

2 Application Android

Avant d'évoquer le générateur d'applications dictionnaire lui-même, il convient de présenter le type d'application qu'il génère. Le code étant ouvert les créateurs d'applications ont tout loisir de modifier le rendu final selon leurs besoins.

L'application est constituée de différentes pages :

- Page de présentation : une image qui apparaît 5s. au lancement de l'application.
- Page principale (*fig.* 1) : cette page permet d'accéder directement à une entrée en commençant à la taper et en choisissant parmi les propositions automatiques. Il est aussi possible de lancer une recherche à partir d'une définition dans une des langues de traduction ou d'accéder à la page de recherches avancées ou simplement de faire afficher tous les entrées commençant par une lettre.
- Page de recherche avancée : permet de faire une recherche plus fine avec divers filtres et conditions.
- Page liste de mots : affiche la liste des résultats d'une recherche faite à partir d'une traduction, d'un thésaurus etc.
- Page de description d'une entrée (*fig.* 2) : Cette page présente une entrée particulière, on y trouve les informations classiques (partie du discours, définitions dans les langues choisies par l'auteur du dictionnaire…) des informations complémentaires (thésaurus, notes anthropologiques…) des mises en contextes (aussi bien des exemples classiques que des expressions figées, des proverbes...) des liens vers d'autres entrées (synonymes, hyperonymes...) ainsi que des fichiers son ou image.
- Pages techniques : informations sur le dictionnaire, sur l'application, préférences…

Figure 1: Page principale

– Page d'interaction avec l'utilisateur : si il est possible d'envoyer un mail à partir de la page principale, il existe une page spécifique accessible depuis la page de description d'une entrée. Elle permet d'envoyer un commentaire au sujet d'une entrée spécifique. Il est possible de saisir le commentaire qui apparaîtra dans le corps du mail mais aussi d'enregistrer un fichier son qui sera ajouté en pièce jointe.

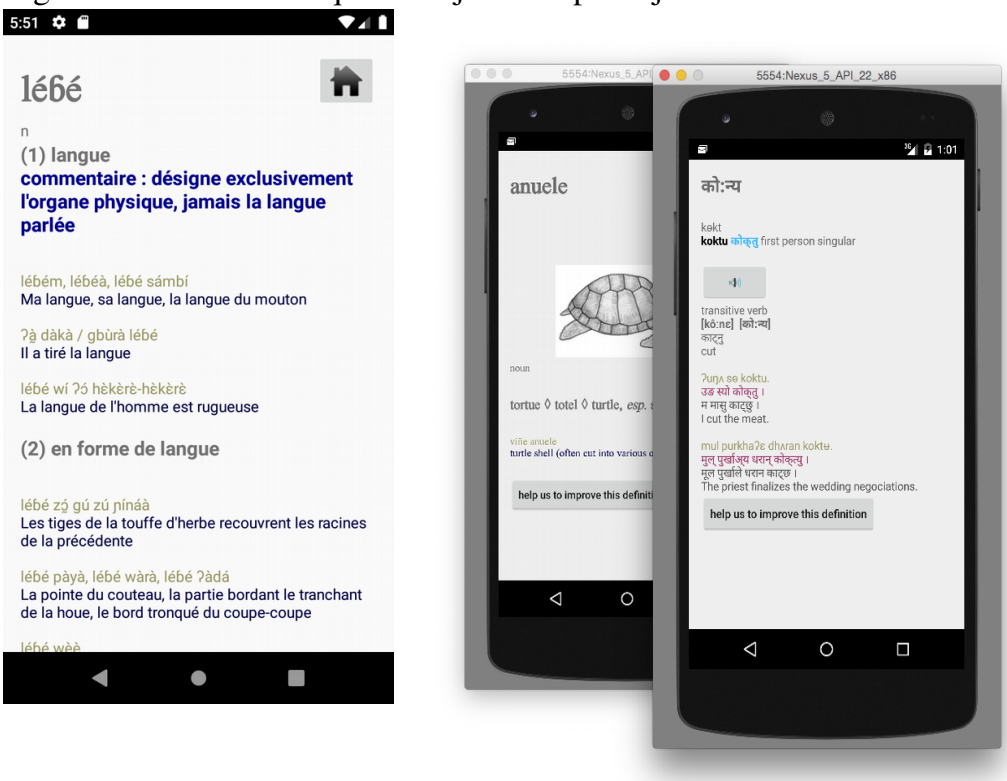

FIGURE 2: Exemples de page description

3 Générateur d'application

Le générateur d'applications peut être utilisé avec une ligne de commande ou à l'aide d'une petite interface (*fig*. 3). Conformément à la philosophie du logiciel libre, notre outil produit un code source. Ce code est écrit en Java (licence publique générale GNU). Java a été choisi car il est l'un des deux langages de programmation officiellement pris en charge par Android et qu'il reste un des langages informatiques les plus utilisés. Le code s'accompagne d'une base de données (bdd) en SQLite dans laquelle sont stockées les données du dictionnaire. Le format de la bdd est inspiré de LMF (Francopoulo, 2013) et compatible avec les formats lexicaux les plus utilisés (notamment MDF et LIFT qui doivent leur succès aux applications Toolbox et Flex développées par la SIL). Pour mettre à jour une application suite à une modification du dictionnaire il suffit de mettre à jour la bdd. Pour générer la base de données il faut fournir à l'application, outre le dictionnaire lui-même, un fichier décrivant les balises spécifiques à ce dictionnaire, l'ordre alphabétique de la langue (il est possible de spécifier les caractères spéciaux ou les lettres ne pouvant pas être à l'initiale d'un mot) et l'emplacement des éventuels fichiers son et image. Le générateur a donc un mode création d'application et un mode mise à jour qui se contente de modifier la bdd.

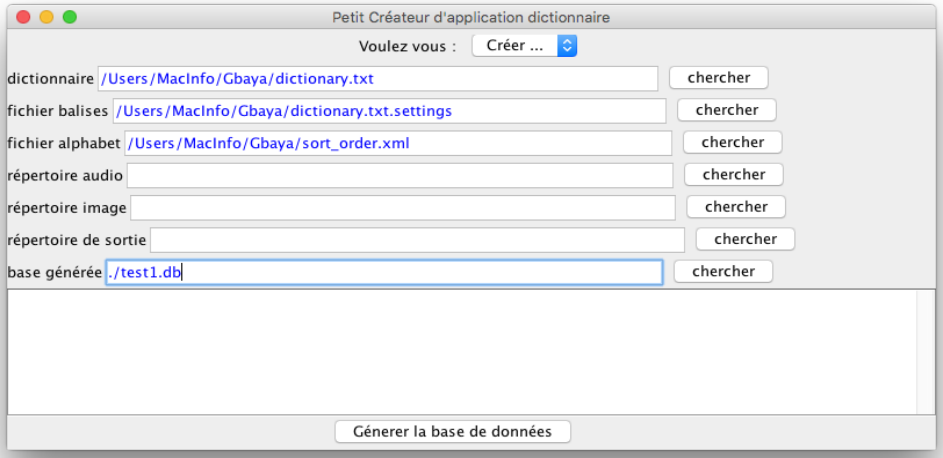

FIGURE 3: Page principale de l'outil de génération

Le code source est peu différent d'une application à une autre, c'est pourquoi sa génération nécessite seulement les informations supplémentaires par rapport à la création d'une bdd : les noms du ou des auteurs, l'adresse électronique à laquelle seront envoyées les commentaires des utilisateurs, l'image à afficher dans la page de présentation ainsi que l'image à utiliser pour l'icône de l'application (qui apparaît sur le menu du terminal mobile).

4 Premiers résultats

Outre trois dictionnaires (khaling, japhug et bena) actuellement accessibles mais issus d'une version antérieure de l'application, la version électronique du dictionnaire gbaya écrit par Paulette Roulon Doko est actuellement en version beta sur le "Play Store" de Google, elle sera accessible au grand public dès la fin de la période de confinement. Des dictionnaires kakabé, bambara, wolof et keeraak seront tester par des utilisateurs volontaires à ce moment là. D'autres dictionnaires devraient suivre rapidement notamment des dictionnaires plus anciens comme le dictionnaire shisumbwa-français écrit en 1901 par le père Capus qui a été numérisé. Le code source du générateur d'applications dictionnaire sera disponible sous licence open-source d'ici la fin de l'été 2020 sur le site du LLACAN (http://llacan.vjf.cnrs.fr/).

Références

AUFFRAY C, (2016) *Chiffres clés : les OS pour smartphones* (zdnet)

FRANCOPOULO G. *(Ed.). (2013). LMF: Lexical Markup Framework. Wiley Online Library.*

MUSGRAVE S. & THIEBERGER N. (2014, November). *Rethinking grammatical description: from Heath to hypertext*. Lecture, Research Unit for Indigenous Language, University of Melbourne.

POUSHTER J. (2016). Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies in https://www.pewresearch.org/global/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/