Conception et Développement d'un Système Automatique d'Ecriture Amazighe: Etat d'Avancement et Perspectives

Y. Es Saady, B. Bakkass, A. Rachidi, M. El Yassa, D. Mammass
Laboratoire IRF-SIC, Université Ibn Zohr
B.P. 8106, Hay Dakhla Agadir, Maroc
essaady2110@yahoo.fr, b_brahim11@yahoo.fr, rachidi.ali@menara.ma,
melyass@gmail.com , mammass@univ-ibnzohr.ac.ma

Résumé : Aujourd'hui, le développement des ordinateurs personnels et celui des réseaux font de l'informatique un instrument pour écrire et communiquer au même titre que le papier l'est avant. La forte augmentation de texte Amazighe disponible en format papier a fait ressortir la nécessité de concevoir et de développer des outils de traitement automatique de texte amazighe performants dans le but de produire des documents en format numériques. Dans ce cadre et parmi nos axes de recherche, nous essayons de concevoir et de réaliser un système automatique d'écriture Amazighe. Ce système permettra d'effectuer des opérations de base d'un éditeur de texte amazighe.

Mots clefs

Editeur de texte, Ecriture Amazighes, Unicode, système d'écriture.

1. Introduction

Aujourd'hui, le développement des ordinateurs personnels et celui des réseaux font de l'informatique un moyen pour écrire et communiquer au même titre que l'est le papier depuis Cai Lun et l'imprimerie depuis Gutenberg. Mais les langues ne sont pas égales devant le processus d'informatisation et les populations parlant des langues mal dotées ont un accès limité à ces nouveaux moyens, limitation pouvant aller d'une simple gêne à une incapacité totale. L'Amazighe fait parti de ces langues peu dotées informatiquement. Par conséquent, des recherches scientifiques et linguistiques sont lancées pour remédier à cette situation [1]. L'un des volets prioritaire de cette recherche, est de concevoir et réaliser des applications capables de traiter de façon automatique des données linguistiques (données exprimées dans la langue naturelle Amazighe). Parmi les outils logiciels et ressources pour l'Amazighe à développer :

- En informatique multilingue
  - Au niveau des systèmes d'exploitation
    - Encodage des caractères

        - Méthodes de saisie
        - Affichage
    - Au niveau des interfaces de programmation
        - Éditeurs de texte
        - Tri lexicographique
- En traitement automatique des langues naturelles
    - Au niveau applicatif
        - Traduction automatisée
        - Reconnaissance optique des caractères
        - Gestion de dictionnaires
    - Au niveau des ressources
        - Dictionnaires d'usage et dictionnaires bilingues

Dans ce contexte, nous proposons des méthodes et des stratégies pour produire un outil de traitement de texte bien adapté à l'écriture Amazighe offrant des fonctionnalités spécifiques de l'écriture amazighe.

Ce papier est composé de cinq parties. Dans la première partie, nous présentons les éléments de base de système d'écriture informatique Amazighe. La deuxième partie est consacrée à la présentation des fonctionnalités de base d'un éditeur de texte amazighe. Nous présentons dans la troisième partie la réalisation effectuée dans ce projet en présentant les outils de développements et l'état d'avancement de l'application.

2. Base des systèmes d'écriture informatiques

Les systèmes d'exploitation actuels des micro-ordinateurs intègrent la capacité Unicode dans le sens qu'ils présentent une interface de programmation compatible avec Unicode. Ils sont donc nativement multilingues, pour autant que le système d'écriture considéré soit dans Unicode et qu'une police de caractères existe et fonctionne pour ce système d'écriture. L'élément de base permettant de créer du texte est la fenêtre d'édition (fenêtre dans laquelle on peut saisir du texte). Des fenêtres d'édition évoluées permettant l'édition dans plusieurs systèmes d'écriture contenus dans Unicode, voire dans tous, sont incluses dans les environnements de développement sous forme d'objets ou d'interfaces de programmation (API). L'utilisation de ces fenêtres d'édition permet un gain de temps considérable, ces objets étant devenus très complexes avec la prise en compte d'Unicode [2]. Ils réalisent en effet les fonctions suivantes :

- gestion des actions clavier et souris,
- affichage du texte,
- coupures de fin de ligne,
- justification du texte,
- gestion du mouvement du curseur,

- sélection du texte (vidéo inversée),
- copie collage.

En plus de ces fonctions de base, les fenêtres d'édition courantes (HTML, RTF, Word...) gèrent l'association d'attributs — gras, italique, souligné, police... — à des parties de texte, grâce, généralement, à un balisage du texte. Ces fonctions, déjà assez lourdes à développer pour du texte en caractères latins, deviennent extrêmement complexes avec la prise en compte des contraintes liées à l'ensemble des systèmes d'écriture :

- forme de caractères dépendant de leur voisinage (par exemple arabe, hébreu, thaï et hindi), ce qui n'est pas le cas pour l'amazighe puisque, pour l'instant, on a pas d'écriture cursive.
- bidirectionalité (par exemple un texte qui contient une partie amazighe et une partie arabe ou latin),
- écritures verticales (par exemple chinois [exemple ci-dessous], ouïgour).

Plusieurs classes de fenêtres permettent de gérer ces écritures complexes. Sous Windows, la classe CRichEditCtrl encapsule le contrôle Rich Edit dont la version 3 couvre presque entièrement Unicode 3. Sous Linux/Unix, Windows et MacOS X, QT propose la classe C++ QTextEdit qui intègre plusieurs caractéristiques complexes comme la bidirectionalité (par exemple pour l'arabe et l'hébreu) et la césure des écritures sans séparateur entre mots (par exemple pour le chinois, le japonais, le coréen et le thaï). La bibliothèque Swing de Java offre plusieurs classes dérivées de la classe de base EditorKit qui est la composante « contrôle d'édition » de la classe JTextComponent, permettant en particulier l'édition aux formats HTML (classe HTMLEditorKit) et RTF (classe RTFEditorKit).

Ainsi, des fonctions paraissant aussi basiques que la sélection de texte et même la gestion de la position du curseur deviennent de véritables casse-tête, en particulier avec des textes incluant à la fois les systèmes d'écriture amazighe et latin.

De nombreuses applications compatibles avec Unicode ont été développées pour d'autres langues, en particulier des suites bureautiques et des navigateurs Internet. Certaines proposent des services linguistiques: détection automatique de la langue, formatage automatique de la date, coupure des mots en fin de ligne, segmentation (pour les écritures sans séparateur entre mots), correcteurs d'orthographe, de grammaire et de style, tri lexicographique, dictionnaire de synonymes, résumé automatique, etc.

Par exemple, Office XP, l'une des suites bureautiques les plus répandues, inclut des outils linguistiques pour quarante-huit langues [3] [4]. Certaines de ces

applications sont elles-mêmes des objets pouvant être utilisés comme plates-formes pour informatiser la langue Amazighe.

3. Fonctionnalités de base d'un éditeur de texte Amazighe

Après une étude sur les éditeurs existant qui permet de traiter les langues en général et particulièrement la langue amazighe comme (MS-word, wordpad,…), on a décidé de concevoir notre propre éditeur au format de WordPad. Cet éditeur est une plate forme logicielle qui permet de traiter un texte Amazighe sous format Unicode, qui visait les premiers niveaux du service de traitement du texte. Il inclut les fonctionnalités suivantes :

- La saisie de textes Amazighes indépendante de la police utilisée et utilisant un clavier intuitif ;
- La sélection du texte à la souris et au clavier des graphies Tifinaghs et des mots Amazighes ;
- L'ouverture et l'enregistrement des documents.
- La mise en forme des caractères, des paragraphes et la mise en page ;
- La visualisation et l'impression des pages;
- La recherche et le remplacement d'un texte ;
- L'insertion des objets;
- construction d'un lexique à partir de textes (ajouter, modifier, supprimer une entrée dans un lexique local),

L'interface proposée de notre application est illustrée à la figure 1 ci-dessous.

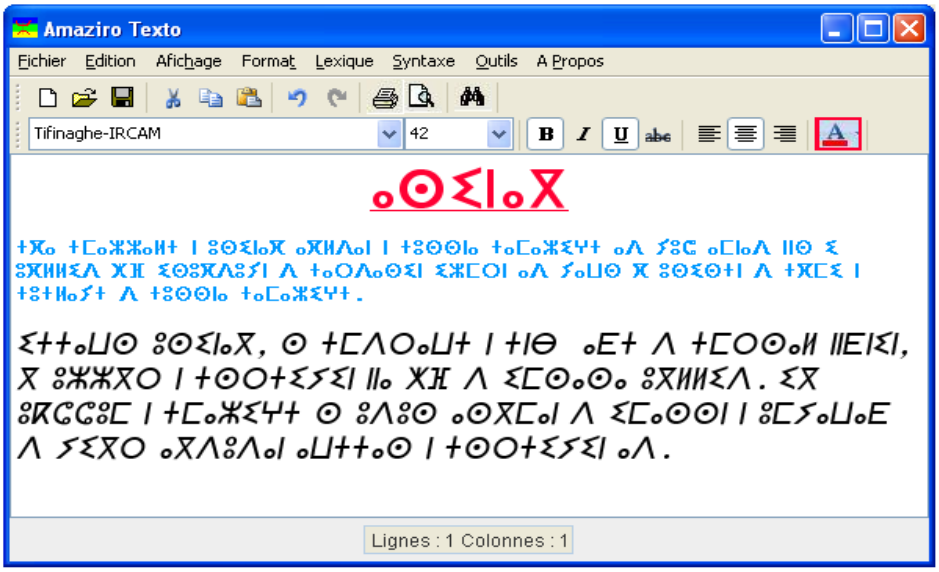

Figure 1 – Application Editeur de texte Amazighe

4. Réalisation d'un traitement de texte pour l'Amazighe

Notre application est développée dans le langage objet java en utilisant Swing qui fait partie de la bibliothèque Java Foundation Classes (JFC). Swing est une API dont le but est similaire à celui de l'API AWT mais dont le mode de fonctionnement et d'utilisation est complètement différent. Nous avons utilisés aussi l'outil de développements Eclipse qu'est un environnement de développement intégré et open source. Il se caractérise par une architecture ouverte à base de plug-ins. C'est l'un des IDE les plus utilisés par les développeurs java. Il bénéficie du support de plusieurs autres projets de taille tels que : JBoss, Jonas, Ant, Tomcat …

La première version de notre application est composée de plusieurs packages dont les principaux sont :

- Package éditeur : regroupe les classes de base de l'éditeur de texte.
- Package Segmentation: regroupe les classes qui permettent de faire la segmentation du texte amazighe.
- Package Dictionnaire: englobe les classes qui font la gestion de notre dictionnaire et calcule de statistique.
- Package Morphologie: rassemble les classes de l'analyseur morphologique (en cours de développement).

Nous avons utilisé les Polices tifinaghes et Claviers UNICODE développés par le centre informatique de l'institut Royal de la Culture Amazighe (IRCAM) qui sont disponibles dans le site de l'IRCAM [5].

Pour rendre l'application cent pour cent amazighiène et dans une deuxième version de notre application, nous avons traduit les textes de l'interface en Amazighe. Nous avons utilisé le lexique d'informatique Français - Anglais – Berbère de Samiya Saad-Buzefran [6] pour traduire les termes des barres de l'interfaces en Amazighe. La figure 2 ci-dessous présente la fenêtre de l'application avec les menus en tifinaghe.

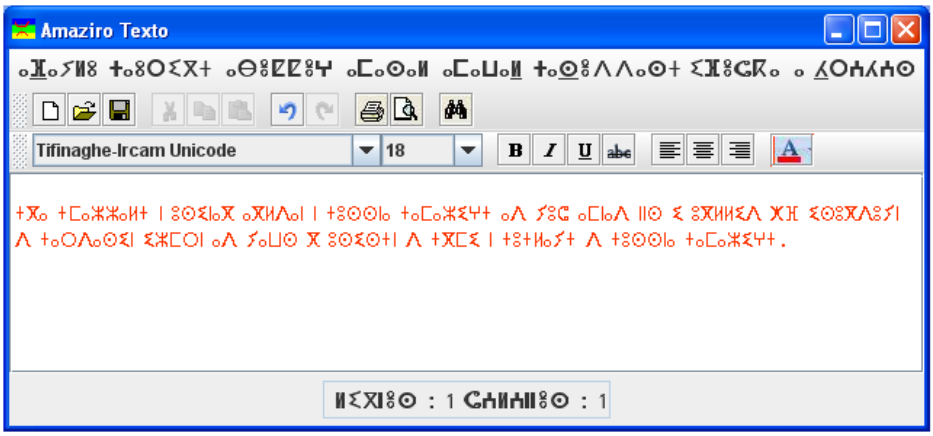

Figure 2 – Application Editeur de texte Amazighe avec l'interface en Tifinaghe

5. Conclusion et Perspectives

L'informatisation de langue amazighe est devenue primordiale pour la promotion de la culture Amazighe. L'existence du standard Unicode a récemment permis la réalisation de systèmes d'exploitation et de logiciels couvrant de nombreux systèmes d'écriture tout en évitant la multiplication des incompatibilités entre plates-formes. L'Amazighe bénéficie ainsi d'outils d'édition performants. Le codage et les principes de base étant communs aux différents systèmes d'écriture. La dynamique Unicode a ainsi conduit à la réalisation de logiciels performants et génériques couvrant en grande partie le premier niveau des services de traitements du texte. En perspective, nous essayons d'intégrer des fonctionnalités avancées à notre éditeur telles que l'analyseur lexical, et l'analyseur morphologique. Pour arriver prochainement à ajouté l'analyse syntaxique dont le but major est de faire les corrections orthographiques et grammaticales de la langue amazighe.

Références

[1] A. Rachidi, D. Mammass: Informatisation de La Langue Amazighe: Méthodes et Mises En Œuvre, SETIT 2005 3rd International Conference: Sciences of Electronic Technologies of Information and Telecommunications March 27-31, 2005 – TUNISIA.

[2] Vincent Berment méthodes pour informatiser des langues et des groupes de langues « peu dotées », thèse de Doctorat de l'université Joseph Fourier, Grenoble 1, UFR d'informatique et mathématiques appliquées, 18 mai 2004.

[3] l'atelier sur les langues minoritaires des conférences LREC (tous les deux ans depuis 1998):

- http://www.lrec-conf.org/fr/index.html,
- http://www.lrec-conf.org/lrec98/ceres.ugr.es/_rubio/elra/minority.html,
- http://www.lrec-conf.org/lrec2000/www.cstr.ed.ac.uk/SALTMIL/lrec00.html,
- http://www.lrec-conf.org/lrec2002/lrec/wksh/WP15agendaF.html

[4] l'atelier associé à TALN 2003 « Traitement automatique des langues minoritaires et des petites langues » : http://www.sciences.univ-nantes.fr/irin/taln2003/page/acte_sommaire.html#atelier.

[5] Institut royal de la culture Amazighe, centre des études informatiques et des systèmes d'information et de communication, Polices et Claviers UNICODE http://www.ircam.ma/fr/index.php?soc=telec&rd=3

[6] Samiya Saad-Buzefran, Lexique d'informatique Français - Anglais – Berbère. ISBN : 2-7384-4650-7 • 1996.