المعهد الملكي للثقافة الأمازيغية
ⵉⵏⵙⵜⵉⵜⵓⵜ ⴰⵖⵔⵓⵎⵉ ⵏ ⵜⵙⴷⵓⵙⵜ ⵜⴰⵎⴰⵣⵉⵖⵜ
INSTITUT ROYAL DE LA CULTURE AMAZIGHE

ⵙⵉⵏⴰⴳ

Asinag

Dossier

Les Technologies de l'Information et de la Communication (TICs) au service de l'amazighe

Coordonné par Youssef Aït Ouguengay

Revue de l'IRCAM - Numéro 9

ⵓⴰⵎⵓⵍⵓⵅ – Asinag

Revue de l'Institut Royal de la Culture Amazighe
Numéro 9 – 2014

Asinag-Asinag est une revue scientifique et culturelle marocaine dédiée à l'amazighe avec ses composantes linguistique et civilisationnelle. Elle est plurilingue et multidisciplinaire et comprend des dossiers thématiques, des articles, des entretiens, des comptes rendus, des résumés de thèses et des créations littéraires. La revue *Asinag-Asinag* est dotée d'un comité scientifique et ouverte à la communauté scientifique nationale et internationale.

Dépôt légal : 2008 MO 0062
ISSN : 2028-5663
Imprimerie El Maarif Al Jadida – Rabat 2014

Sommaire

Présentation

Démocratiser l'outil informatique : telle est en fait l'une des préoccupations majeures des décideurs en matière de technologies de l'information et de la communication (TICs). En témoigne la tendance actuelle de ces technologies, à savoir l'orientation de l'utilisation des nouvelles productions technologiques (logicielles et matérielles) dans le sens de faciliter au grand public l'accès à l'information et ce, en prenant en considération les spécificités locales de l'utilisateur : langue, culture, niveau intellectuel, etc.

Les communautés et organisations (nationales et internationales) œuvrant dans le domaine de l'information et de la communication ont bien saisi les enjeux de la valorisation de la langue et de la culture dans les TICs. Les langues naturelles peu dotées sur le plan informatique telles que la langue amazighe, ont vu émerger plusieurs plates-formes et projets de recherche nationaux et internationaux qui favorisent le développement de la langue et de la culture sur la base des nouvelles technologies de l'information.

Le présent numéro de la revue *Asinag*-*Asinag*, constitué de douze articles (huit en français, deux en anglais et deux en arabe) exclusivement consacrés au domaine des technologies appliquées à la langue, vise la mise en exergue des avancées réalisées dans ce champ de recherche.

La contribution de Patrick Andries dresse un tableau panoramique à propos de l'introduction de l'amazighe dans les nouvelles technologies à l'échelle internationale. Il retrace les différentes étapes en commençant par le niveau le plus basique qu'est le codage du système de l'alphabet tifinaghe. Ensuite, il aborde l'intégration du clavier amazighe dans les systèmes d'exploitation (Windows, en particulier), les polices de caractères et la prise en charge du tifinaghe dans les langages Web pour arriver au stade le plus avancé. C'est-à-dire le renseignement du registre Unicode CLDR (Common Locale Data Repository) et la reconnaissance du système graphique tifinaghe dans les bibliothèques logicielles.

Dans leur article, Fadoua Ataa Allah et Siham Boulaknadel traitent du rôle des technologies de l'information dans la promotion de l'amazighe au niveau national. La technologie, procédé de sauvegarde du patrimoine culturel, y est considérée comme un vecteur de l'aménagement linguistique de l'amazighe (différentes opérations de codification, gestion des ressources, etc.). Cette technologie offre également des moyens qui facilitent l'enseignement de la langue en permettant la promotion de l'amazighe par l'exploitation de la

technologie éducationnelle, de l'apprentissage à distance et des outils du TAL amazighe. Aussi les auteurs proposent-elles une méthode élaborée pour l'informatisation de l'amazighe à court, moyen et long termes.

Le projet de boîte à outils unifiés de la linguistique computationnelle « SmallCodes », spécialement conçu pour les langues peu dotées, est présenté par Carlo Zoli. Celui-ci explique l'objectif du projet en montrant les démarches des champs disciplinaires mobilisées dans sa recherche et présente une démonstration des performances du système développé. Il propose ensuite une analyse des approches existantes et essaie une projection de son approche sur l'amazighe.

Nora Tigziri et Henri Hudrisier exposent le projet *Humanité DigitMaghreb* qui vise la construction d'une bibliothèque de corpus multilingues (oraux, littéraires et linguistiques), synergiques et normalisés pouvant garantir la communisation inter-langues. Ils expliquent l'utilisation de la norme historique TEI (Text Encoding Initiative) en vue de réaliser leurs objectifs de départ et soulignent les difficultés rencontrées au niveau des Sciences humaines.

Dans le domaine de la construction des corpus annotés, Mohamed Outahajala *et al.* proposent d'améliorer le rendement de leur étiqueteur morphosyntaxique par l'exploitation des processus CACs (Champs Aléatoires Conditionnels). Sur un texte prétraité, ils expérimentent leur système ainsi augmenté et comparent les résultats obtenus par le choix des données d'auto-apprentissage ayant une valeur de confiance élevée à ceux obtenus sur la base de données aléatoires.

L'article de Violetta Cavalli Sfortza traite deux aspects de la pertinence des textes et leur degré de facilité dans le processus de lecture pour la langue arabe. Il s'agit de la lisibilité du texte, indépendamment de l'apprenant ou du programme d'enseignement, et du degré d'appropriation du texte dépendant des connaissances et du vocabulaire de l'apprenant. Après avoir passé en revue la littérature existant sur la question, l'auteur propose un modèle prédictif de cette pertinence et met en œuvre des outils techniques pour une partie de l'analyse (*e.g.* MADA). Elle essaie une projection sur la lisibilité de la langue amazighe, suite à quoi elle présente les limites de l'étude.

La reconnaissance optique des caractères tifinaghes est soulevée par Ali Rachidi. Ce dernier présente une synthèse des travaux réalisés dans ce domaine au niveau national et procède à leur comparaison à la lumière des bases de données des caractères tifinaghes utilisées.

La reconnaissance de la parole est soulevée par Abenaou et *al.* Les auteurs détaillent les méthodes et algorithmes adoptés pour la synthèse des mots en amazighe en proposant des techniques pour optimiser les données et augmenter la vitesse de la reconnaissance. Ils concluent par la présentation de résultats expérimentaux et une évaluation de la performance du système.

La contribution conjointe de Hammou Fadili et Malika Chakiri expose la conception et l'élaboration d'un modèle d'ontologies de domaines ; modèle enrichi par des liens lexico-sémantiques associés aux fonctions lexicales de la Théorie Sens-Texte (Mel'cuk, 1997) et la notion du contexte. Les auteurs fournissent quelques définitions des outils de travail utilisés tels que les ontologies, les langages informatiques utilisés et la relation contexte-ontologie. Ils abordent la constitution du corpus ontologique et notent des difficultés dans le cas de l'amazighe, auxquelles ils proposent des solutions linguistiques et technologiques.

La contribution de F. Ataa Allah, S. Boulaknadel et H. Souifi est consacrée à l'élaboration d'un *Jeu d'étiquettes morphosyntaxiques* de la langue amazighe. Le travail se fonde sur les recommandations EAGLES qui visent la réutilisation des corpus et la comparaison des langues dans le domaine du traitement automatique du langage naturel.

Le domaine des technologies appliquées à l'apprentissage est au centre de l'article de Robert Bibeau. Y sont traitées les problématiques d'accessibilité et de normalisation des ressources d'enseignement et d'apprentissage, notamment dans le cas de plusieurs instances, utilisant différentes langues, sous divers environnements technologiques. Le profil NORMETIC, variante d'application de la norme IEEE 1484.12.1 (LOM) des métadonnées d'objets d'apprentissage, est conçu pour répondre à ces exigences. Eurêka, banque des ressources d'apprentissage, suffisamment décrite dans le présent article, est un exemple de bases de données compatibles avec le profil NORMETIC.

Le volet en langue arabe contient deux contributions. Dans l'une, Mohamed Lguensat aborde la graphie du tifinaghe et propose une approche pour son aménagement. Il appuie sa thèse par des modèles concrets et souligne les défis liés, d'une part, aux déterminants techniques et informatiques et, d'autre part, aux éléments de la communication visuelle. Dans l'autre, Hassan Jaa et Youssef Ait Ouguengay brossent un panorama général des travaux qui ont accompagné, au cours de la dernière décennie, l'introduction de l'amazighe dans les nouvelles technologies de l'information ; notamment les efforts consentis au niveau national et à l'IRCAM.

Le dossier est agrémenté d'un entretien avec Lahbib Zenkouar et Patrick Andries sur des questions relatives à la promotion de l'amazighe dans les TICs ainsi qu'à la recherche scientifique dans ce domaine.

Le présent numéro comprend également un compte rendu élaboré par Fouad Brigui sur l'ouvrage d'Ahmed Boukous intitulé : *Revitalisation de la langue amazighe : Défis, enjeux et stratégies*, publié par l'Institut Royal de la Culture Amazighe en 2012.

La rubrique *Résumés de thèses* est destinée à faire connaître des travaux académiques récents et inédits sur l'amazighe. Y sont livrés deux résumés

portant sur la reconnaissance automatique des caractères tifinaghes imprimés et manuscrits. Le premier, dû à Youssef Es-saady, traite de deux approches disjointes : l'une, syntaxique, utilise les automates finis ; l'autre, neuronale. L'auteur analyse les résultats des deux approches et conclut à leur pertinence. Le second, de Moustapha Amrouche, propose, d'abord, une modélisation markovienne complémentée par une technique de caractérisation de caractères isolés, et, ensuite, la combinaison d'analyses par chemins discriminants (DP-HMM) et des caractéristiques morphologiques de la graphie amazighe. Le test de rendement et les résultats des deux thèses sont effectués sur une base de données nommée AMHCD, élaborée dans le cadre du travail de Youssef Es-saady.

La Direction et le Comité de rédaction de la Revue tiennent à exprimer leurs plus vifs remerciements à toutes les personnes qui ont apporté une quelconque contribution à la réalisation de ce numéro : Patrick Andries, Fadoua Ataa Allah, Aïcha Bouhjar, Belaïd Bouikhalene, Siham Boulaknadel, El Houssine Bouyakhf, Violetta Cavalli-Sfortza, Abdelkrim El Moukhtari, Lahbib Fouad, Abdelfattah Hamdani, El Mehdi Iazzi, Rachid Laabdalaoui, Mohamed Maamouri, Kamal Naït-Zerrad, Patrice Pognan, Mohamed Yeou, Abdellah Yousfi et Lahbib Zenkouar.

Asinag-Asinag

Asinag-Asinag, 9, 2014, p. 11-31

Normalisation et état des lieux de la prise en charge de l'amazighe et des tifinaghes[1]

Patrick Andries
Conseils Hapax, Québec, Canada
Membre du consortium Unicode

In this paper, we look at the progress made since Tifinagh characters were encoded in ISO/IEC 10646 and Unicode in 2005. Eight years ago, it was impossible to send documents written in Tifinagh without reference to a private encoded font. Today, you can create HTML pages, XML documents, email in Tifinagh. Now, There is a standard keyboard approved by the Moroccan standard body to enter Tifinagh text, a standard for sorting Tifinagh strings, and Microsoft provides by default in its newest versions a font that supports the Tifinagh. users can now view HTML pages without having to explicitly install a Tifinagh font on their system. Software libraries like ICU also support Tifinagh and it is possible in theory to have internet domain names in Tifinagh.

Dans cette contribution, nous nous pencherons sur le chemin parcouru depuis le codage des caractères tifinaghes dans l'ISO 10646 et Unicode en 2005. Il y a huit ans, il était impossible d'envoyer des documents en tifinaghes sans se référer à un codage de police privé. Aujourd'hui, on peut créer des pages HTML, des documents XML en tifinaghes, envoyer des courriels. Il existe un clavier normalisé pour saisir des textes tifinaghes, une norme de tri, Microsoft fournit une police qui prend en charge les tifinaghes. L'utilisateur peut désormais voir des pages HTML sans qu'il n'ait à explicitement installer de polices tifinaghes sur son système. Des bibliothèques logicielles comme ICU prennent également en charge les tifinaghes et il est possible, en théorie, d'avoir des noms de domaine Internet en tifinaghes.

[1] Nous tenons à vivement remercier l'IRCAM et plus particulièrement le directeur du CEISIC, Youssef Aït Ouguengay, pour leur accueil chaleureux et l'organisation du colloque international à Rabat au cours duquel une première version de cet article a été présentée. Nous voulons également ici rendre hommage au prédécesseur de M. Aït Ouguengay, le professeur Lahbib Zenkouar, sans lequel la normalisation informatique du tifinaghe n'aurait été ni aussi rapide ni aussi complète.

Introduction

Le 31 mars 2005, Unicode 4.1 était publié. Il comprenait tous les tifinaghes normalisés de l'IRCAM ainsi que les principaux caractères touaregs et kabyles.

Avant cette date, il n'existait aucune façon de coder de manière normalisée des textes tifinaghes. Chaque producteur utilisait un codage, le plus souvent de son cru, lié à une police particulière. Il était donc virtuellement impossible d'échanger des documents produits par des personnes qui utilisaient des polices tifinaghes différentes. C'était vrai pour les courriels, les documents Word, les pages HTML en général et, bien sûr, celles d'une encyclopédie en ligne naissante comme Wikipédia.

Depuis huit ans bien des choses ont changé : les tifinaghes normalisés sont désormais de plus en plus présents sur Internet et dans d'autres produits et normes informatiques. Nous allons brièvement passer en revue ci-dessous ces heureuses améliorations.

Les caractères Unicode

Le lecteur de cette contribution connaît certainement les premiers caractères tifinaghes qui ont été codés dans Unicode 4.1 en mars 2005, car ils ont fait l'objet de plusieurs communications de l'IRCAM et de ses chercheurs[2]. Il ignore peut-être certaines précisions apportées depuis 2005 qui se retrouvent dans les dernières versions d'Unicode.

Les diacritiques

Plusieurs variantes tifinaghes modernes utilisent des diacritiques pour compléter les lettres du bloc tifinaghe. C'est ainsi que la notation Hawad utilise des diacritiques du bloc [U+0300-U+036F] commun à de nombreuses transcriptions latines. Ces signes s'utilisent pour représenter des voyelles ou des consonnes étrangères. Dans cette notation, <U+2D35, U+0307> □ représente un « a » court, <U+2D49, U+0304> □ un « i » long /i:/ et <U+2D31, U+0302> □ permet d'écrire un « p ». On indique certaines voyelles longues à l'aide de deux signes diacritiques, un « é » long /e:/ s'écrit <U+2D49, U+0307, U+0304> □. Ces signes sont affichés côte à côte, et non empilés, au-dessus de la lettre de base dans l'ordre d'apparition dans la chaîne codée.

Quatre caractères ajoutés dans Unicode 6.0 et 6.1

Deux caractères tifinaghes ont été ajoutés dans Unicode 6.0 (et bien sûr dans la version correspondante de l'ISO/CEI 10646) :

- U+2D70 □ SÉPARATEUR TIFINAGHE
  = tazaraste
- U+2D7F ◌□ LIANT DE CONSONNES TIFINAGHE

[2] Voir, par exemple, Lahbib ZENKOUAR (2004 : 173-175).

Le caractère U+2D70 ⵰ est ce signe targui dont Prasse dit[3] « Au Hoggar on nous a donné le séparateur ⊂, à l'intérieur duquel s'écrivait la dernière lettre de chaque mot phonétique. »

Les deux figures ci-dessous illustrent l'utilisation de ce séparateur. On remarquera que le signe est réfléchi quand il s'écrit dans un contexte de droite à gauche.

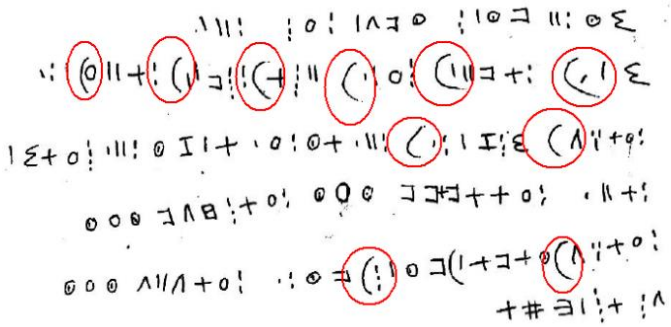


Figure 1 : Exemple d'utilisation du séparateur de mot, Sud de l'Algérie

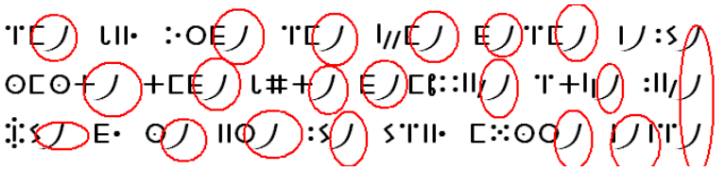


Figure 2 : Utilisation au Niger dans un contexte de gauche à droite

Le *liant de consonnes* U+2D7F ◌⵿ est un caractère de commande qui permet de forcer la formation d'une biconsonne. Nous étudierons son utilisation dans la section 2.3. *Ligatures*.

Unicode 6.1 a par la suite ajouté deux caractères :

- U+2D66 ⵦ LETTRE TIFINAGHE YÉ
- U+2D67 ⵧ LETTRE TIFINAGHE YO

YÉ et YO sont simplement deux voyelles préconisées par l'APT (Association pour la promotion des tifinaghs) au Niger pour transcrire les voyelles « é » et « o ».

Ligatures

Dès le codage des tifinaghes dans Unicode 4.1, il était possible de préciser que l'on préférerait obtenir une biconsonne (ou une triconsonne) en insérant un U+200D ‍ *liant sans chasse* entre les consonnes qui constituent la ligature souhaitée. On peut

[3] Karl-G. PRASSE (1972 : 152).

également utiliser U+200C *antiliant sans chasse* entre deux consonnes pour décourager la formation d'une ligature à partir de ces deux lettres.

C'est ainsi que l'on peut demander la formation de la biconsonne « rt » en écrivant <U+2D54, U+U200D, U+2D5C> (□□□). Les polices pourvues d'une telle ligature devraient en présence de ces caractères afficher le glyphe de celle-ci, par exemple □, □, □, ou □ selon la police. Il existe également quelques cas de triconsonnes : parmi celles-ci notons le groupe consonantique « nft » □□□ dont la ligature est parfois □ et « nkn » □□□ représenté dans certaines régions par □. Pour bien fixer les idées, on demandera la formation optionnelle de la ligature « nkn » à l'aide de la suite suivante de caractères : <U+2D4F, U+200D, U+2D3E, U+200D, U+2D4F>.

Certaines polices pourront être dépourvues de ligatures, d'autres n'en inclure que pour certaines variantes géographiques. Si une police venait à ne pas avoir de ligature correspondant à la suite de caractères liée par un liant sans chasse, la police devrait simplement afficher les deux consonnes de base, à savoir □□ dans notre exemple ci-dessus.

Toutefois, aux yeux d'aucuns, il est apparu que le liant sans chasse qui n'indique que la formation facultative d'une biconsonne ne suffisait pas. Il fallait pouvoir préciser qu'une biconsonne devait impérativement être formée car la présence de cette ligature, dans une graphie non voyellée, indique l'absence d'une voyelle implicite entre les consonnes qui forment la ligature. C'est à cet effet qu'a été introduit dans Unicode 6.0 U+2D7F ◌□ *liant de consonnes tifinaghe*. Ce caractère de commande impose la formation de la ligature. À ce titre, il joue un rôle similaire au U+0652 ◌ْ *soukoûn*, le signe de quiescence arabe. Le tableau ci-dessous illustre l'utilité d'une telle convention, les exemples sont tirés du dictionnaire de Foucauld (Ch. de Foucauld, 1951).

Lettres de base	Graphie touarègue	Translittération	Glose en français
□□□	□□	ənkər	se lever (inhabituel)
	□□□	nâkːər	se lever (habituel)
□□□□	□□□	teɣert	marmite en terre
	□□□□	teɣərit	cri strident, très perçant
□□□	□□[4]	istəɣ	chasser, pousser devant soi
	□□□	ɔsataɣ	chasse, poursuite
□□□□	□□□	təfert	mot, proposition, vers
	□□□□	tefərit	petite aiguille rocheuse
□□□	□□	əndər	excéder les forces
	□□□	ənadar	fait d'être en chaleur, en rut

Dans les exemples ci-dessus, comme les biconsonnes sont considérées comme obligatoires, on n'utilisera pas de U+200D *liant sans chasse*, mais bien un

[4] On aurait aussi pu mettre □ à la place de □.

U+2D7F *liant de consonnes tifinagh*. Le verbe *chasser* □□ se codera donc <U+2D59, U+2D7F, U+2D5C, U+2D57>.

Unicode prescrit que, si une biconsonne obligatoire ne peut être affichée à l'aide d'une ligature correcte, le système de rendu devra faire apparaître *le liant de consonnes tifinaghe (LCT)* de manière visible afin de bien indiquer l'absence de voyelle entre les deux consonnes. Cela pourra prendre la forme d'un losange ◆, d'un ○□ ou de tout autre glyphe. Unicode ne prescrit pas ce glyphe de repli. La forme de repli de <S, LCT, T> sera donc <gS, ○□, gT> où gS est un glyphe de S, gT un glyphe de T et ○□ est un glyphe qui indique la présence du *liant de consonnes tifinaghe*.

Rappelons que, si un système de rendu ne peut afficher une biconsonne optionnelle, c'est-à-dire dont les consonnes de base sont séparées par un U+200D *liant sans chasse (LSC)*, la forme de repli consiste simplement dans l'affichage des consonnes constitutives de cette ligature. La forme de repli de <S, LSC, T> est donc <gS, gT>.

Indicatif de langue [zgh] de l'ISO 639

En novembre 2012, le comité de l'ISO 639-2 a approuvé l'ajout d'un indicatif [zgh] dans la liste officielle des noms de langues[5] qu'elle tient à jour. Il désigne l'« amazighe standard marocain » soit la norme convergente préconisée par l'IRCAM qui intègre les trois parlers amazighes du Maroc.

Avant cette addition, il n'existait aucun moyen de préciser qu'un document numérique (écrit ou parlé) était en « amazighe standard marocain ». Il fallait soit utiliser un indicatif comme [ber] qui représente toutes les langues de la famille berbère (y compris notamment le kabyle et le touareg) soit choisir de désigner le texte comme une des variantes « dialectales » de l'amazighe marocain.

On peut désormais indiquer qu'un document est en amazighe standard marocain avec l'attribut `lang` égale à « zgh ». Dans une page HTML (voir l'extrait ci-dessous) dont la langue principale est l'amazighe standard marocain, on l'indiquera en ajoutant un attribut `lang="zgh"` sur la balise correspondant au corps de la page (`<body>`) :

```
<html>
    <body lang="zgh">
        <h1>ⴰⵣⵍⴰⵏ ⴰⵎⴰⵣⵉⵖ</h1>
          <p>ⵙ ⵍⴰⵍⴰⵎ ⵣⵓⵖ ⵉⴷⵙ, ⵙⵙⵏⵖ ⵢⴰⵜ ⵜⵖⴰⵍⵙ ⵜⵉⵙⵙ
   ⵙⵏⴰⵜ ⵉⵅⴰⵜⵓⵏ: ⵉⵜⵓⵉ ⵙⵅ ⴷⴷ ⵉⴷⴰ ⵛⵓ ⵉⵎⵎⵛⵓ, ⵉⵎⴰ ⵖⴰⵙ
   ⴰⵏⵜ ⵏ ⵢⴰⵜ ⵜⴰⴷⴰⵓⵜ!</p>
          <p>ⴰⵢⴰ ⵛⵅⵖ ⵜ. ⵙⵙⵏⵖ ⵉⵙ ⵙⵇⵇⴰ ⵏ ...
```

[5] http://www.loc.gov/standards/iso639-2/php/code_list.php

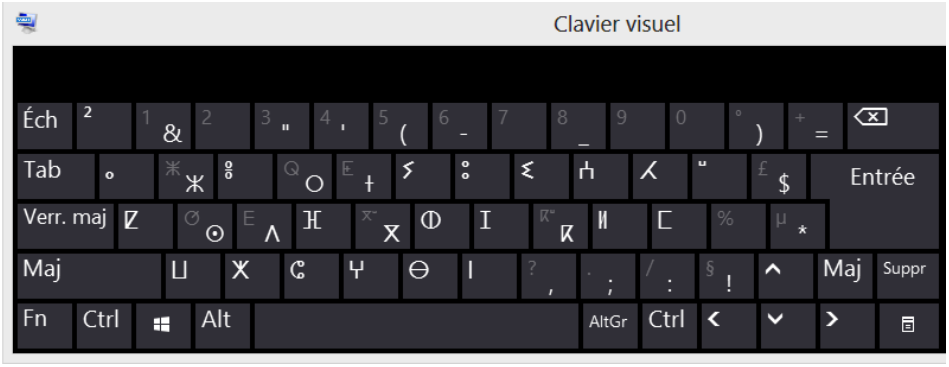
Claviers

Clavier national normalisé

Depuis la normalisation des tifinaghes dans l'ISO 10646 et Unicode, un clavier normalisé marocain pour la saisie de ces caractères a été homologué par le SNIMA [6] . Ce clavier respecte la norme internationale en la matière, l'ISO/CEI 9995. Il a été repris par plusieurs fabricants de Linux et inclus par Microsoft dans Windows 8. Normalisé il y a plusieurs années, ce clavier ne comprend pas les caractères tifinaghes introduits par Unicode 6.0 et 6.1.

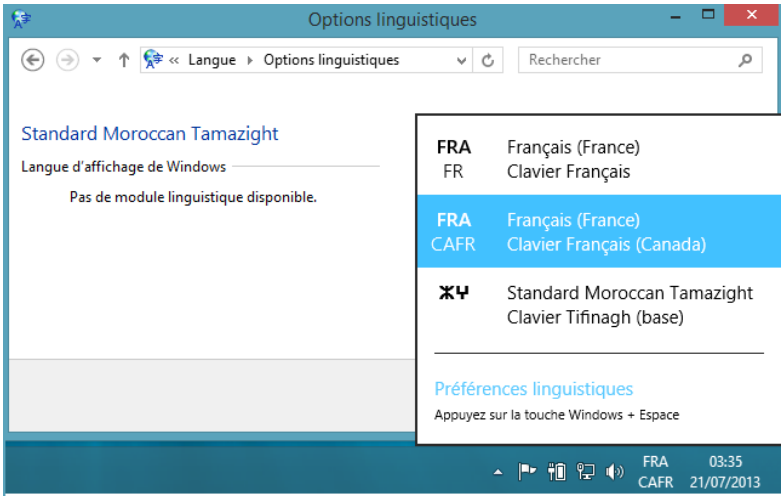
Clavier tifinagh sur Windows 8

Microsoft a adopté le clavier SNIMA dans Window 8 :



Pour le sélectionner, l'utilisateur doit ajouter un profil linguistique à partir du Panneau de configuration de Windows. Le profil linguistique regroupe une langue et une méthode d'entrée correspondante. La langue correspond à un indicatif ISO 639. Dans Windows 8, le seul indicatif « berbère » disponible est [tzm], à savoir le parler amazighe de l'Atlas central. À partir de Windows 8.1, une nouvelle langue et son indicatif sont disponibles : l'amazighe standard marocain [zgh].

[6] http://hapax.qc.ca/pdf/NM%2017.6.000.pdf



L'illustration ci-dessus représente le dialogue de préférence linguistique dans Windows 8.1, la prochaine version de Windows. Au moment d'écrire ces lignes, il n'était pas encore totalement traduit en français.

Une fois le clavier tifinaghe et la langue amazighe choisis, les lettres □□apparaissent en bas à droite de l'écran pour indiquer que le clavier actif est amazighe. Sur l'illustration ci-dessus, le clavier actif est canadien-français (FRA/CAFR).

Norme de tri

Un ordre précis de tri des caractères tifinaghes a également été normalisé par le SNIMA, il s'agit d'un « delta » de la norme internationale en la matière, l'ISO/CEI 14651. Le même ordre est également mis en œuvre par les tris de caractères Unicode (voir ci-dessus 8. *Les bibliothèques logicielles*).

Les documents XML, HTML

Les fichiers XML peuvent en général contenir sans encombre du contenu tifinaghe. Ces documents – comme l'extrait ci-dessous – sont conformes.

```
<?xml version="1.0" encoding="UTF-8"?>
<texte xml:lang="ber" lang="ber">
  <h1>□□□□□□ □□□□□</h1>
  <p>□ □□□□□ □□□ □□□, □□□□ □□□ □□□□□□ □□□□
  □□□□ □□□□□□: □□□□ □□ □□ □□□□ □□ □□□□□□,
  □□□□ □□□ □□□□ □ □□□ □□□□□□□ !</p>
</texte>
```

En 2008, lors d'une précédente contribution (Andries, 2008) où nous avons abordé ce sujet, il existait une restriction dans les caractères admis dans les noms

d'élément et d'attribut ainsi que dans certaines [7] valeurs d'attribut dans les documents XML 1.0. En effet, à l'époque, XML 1.0 ne permettait pour ces noms que des caractères appartenant au répertoire d'Unicode 2.0. Cela signifiait que ces noms ne pouvaient contenir des caractères tifinaghes, ou d'autres provenant d'écritures comme l'éthiopien également introduits après Unicode 2.0, pas plus que des caractères importants récemment ajoutés à des écritures qui existaient déjà dans Unicode 2.0 comme le latin, l'arabe, le cyrillique ou le chinois.

Il n'était donc pas permis d'écrire en XHTML :

```
<a name="□□□" id="□□□">attribut touareg</a>
```

Pas plus que ceci n'était permis en XML 1.0 :

```
<□□□□>…</□□□□>
```

Au vu de cette incongruité, le W3C chargé de la normalisation de XML a décidé de publier à la fin 2008 (le 26 novembre très précisément) la cinquième édition de XML 1.0[8]. Outre la mise à jour de quelques références bibliographiques et la correction de quelques errata, le grand changement introduit par cette édition est de permettre la quasi-totalité des caractères Unicode et notamment les tifinaghes dans les noms d'élément et d'attribut ainsi que les valeurs d'attribut.

En théorie, la pratique est identique à la théorie, mais en pratique cela peut bien sûr être différent. Pour qu'un nom d'élément en tifinaghes soit accepté par un logiciel de traitement de documents XML, il faut tout de même que les analyseurs (parseurs) XML soient modifiés pour mettre en œuvre la nouvelle règle de formation des noms. Il faudra donc encore attendre quelques années avant que les analyseurs en place soient mis à jour pour que l'échange de documents XML/XHTML avec des tifinaghes dans les noms d'éléments et d'attributs puisse se faire sans encombre. Par contre, si les documents XML ne sont utilisés qu'en interne où sont contrôlés les analyseurs XML, il se peut que, en mettant à jour ceux-ci pour choisir une version qui prend en charge la cinquième édition de XML 1.0, vous puissiez dès aujourd'hui utiliser des documents XML dont les noms d'élément, d'attribut et les valeurs d'attributs contiennent des tifinaghes.

Les polices

Ebrima sur Windows

Depuis Windows 7, Microsoft inclut en série dans son système d'exploitation une police « panafricaine », la police Ebrima. Elle est fournie en deux graisses : normale et grasse. Ebrima a été conçue pour prendre en charge un grand nombre de langues africaines. Elle

[7] Les attributs CDATA peuvent contenir n'importe quoi, la restriction était sur les attributs déclarés ID, IDREF, NMTOKEN, etc.

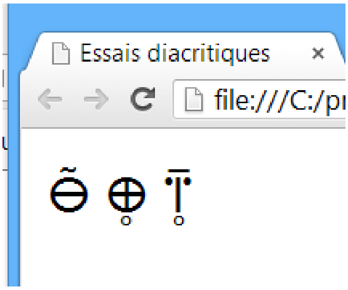
[8] <http://www.w3.org/TR/2008/REC-xml-20081126/>

contient des glyphes pour le n'ko, les tifinaghes, le vaï et l'osmanya. Les glyphes latins de la police sont accompagnés des diacritiques et autres signes utilisés pour représenter les langues africaines.

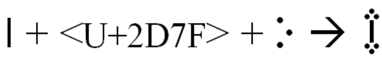
Les œils tifinaghes d'Ebrima sont directement inspirés de ceux utilisés dans les tableaux Unicode (c'est-à-dire de Hapax Berbère).

La police Ebrima ne comprend pas les deux voyelles nigériennes (yé et yo) introduites par Unicode 6.1, elle comprend cependant depuis Windows 8 plusieurs innovations par rapport aux versions précédentes :

1. On peut ajouter des diacritiques à toute lettre tifinaghe[9] :



2. De nombreuses biconsonnes ont été ajoutées à la police, elles sont formées à l'aide du U+2D7F ◌⵿ LIANT DE CONSONNES TIFINAGHE, notamment



3. Quand une biconsonne ne peut être formée parce qu'elle n'est pas prise en charge par la police, Ebrima affiche une série de points souscrits sous la paire de consonnes :

Les polices incorporées dans les pages HTML

Avec une écriture aussi récente – en termes de normalisation informatique – que le tifinaghe, il n'est pas rare qu'un utilisateur qui désire afficher une page Internet qui contient des caractères tifinaghes n'ait pas de police tifinaghe à sa disposition.

Malgré la diffusion récente d'Ebrima, le problème demeure entier aujourd'hui, car il existe encore de nombreux utilisateurs qui n'utilisent pas Windows et *a fortiori* Windows 7.

En outre, que faire quand on veut être sûr qu'une page s'affiche dans un style

[9] Ceci fonctionne bien avec les trois grands navigateurs (Chrome, Firefox et Internet Explorer) et dans Windows 8.1 avec MS Word 2013.

tifinaghe particulier ? On risque donc dans ces cas-là d'être confronté à des pages remplies de petits rectangles blancs comme dans l'illustration ci-dessous, chaque rectangle y représente un caractère (tifinaghe ici) qui ne peut être représenté par manque de police adéquate.

Il existe plusieurs remèdes possibles à ce désagrément. Une solution consiste à prévenir les lecteurs des pages en tifinaghe qu'ils doivent installer une ou plusieurs polices en fournissant un lien permettant de télécharger ces polices. Il existe cependant une autre solution : les polices dynamiquement téléchargeables. Cette solution consiste à envoyer les glyphes nécessaires à l'affichage d'une page HTML avec la page en question. On parle alors de polices embarquées ou incorporées.

Extrait du dictionnaire touareg-français de Charles de Foucauld (page 339)

□□ éferi □□□ sm. φ (pl.iferân □□□), daṛ ăferi (ĕferi), daṛ ferân || aiguille rocheuse (quelconque) || diffère d'ăḍaouḍa « aiguille rocheuse très mince (ressemblant à un doigt) ». Tout ăḍaouḍa est un éferi, mais non inversement.

— téferit □□□□ sf. φ (pl. tiferâtîn □□□□□), daṛ tăferit (tĕferit), daṛ tferâtîn || dim. du pr.

□□ tăfirt □□□ sf. φ (pl. tifir □□□), daṛ tfir || mot (une syllable ou plusieurs syllabes réunies exprimant une idée) || p. ext. « quelques paroles (paroles en petit nombre) » || p. ext. « vers (assemblage de mots rythmés d'après des règles déterminées en poésie) » Le sing. tăfirt signifie un vers, le pl. tifir signifie 2 ou plusieurs vers.

□□ aferra || v. □□□ afri .

Figure 3 : Police manquante dans une page HTML

Il y a près de deux ans, en 2008, nous avions déjà abordé les techniques alors disponibles (Andries, *ibid*.). Rappelons que les polices incorporables dans les pages HTML ne sont pas une idée neuve. Dès 1998, CSS 2 permettait de préciser un lien vers une police que les fureteurs internet pouvaient télécharger. Microsoft et Netscape prirent en charge cette syntaxe et permettaient de télécharger des polices. Malheureusement, aucun des deux fabricants ne prenait en charge directement le format de police le plus populaire : TrueType. Microsoft choisit le format de police EOT (un format propriétaire) dans Internet Explorer 4.0 alors que Netscape 4.0 jeta son dévolu sur un format rival TrueDoc qu'il abandonna deux versions plus tard, car Mozilla ne pouvait rendre public le code source de TrueDoc, propriété de Bitstream.

Depuis 2008, cette question a connu un vif regain d'intérêt. À partir de 2010, on a assisté à un véritable engouement pour les polices internet. Subitement, coup sur coup, on a assisté à l'élaboration d'un format de police conçue pour le téléchargement sur Internet (WOFF), la prise en charge de ce standard naissant par

les quatre grands concepteurs de moteurs de composition HTML[10] et enfin la mise en place de nombreux sites de partage ou de vente de polices incorporables dans des pages HTML[11].

- ***WOFF***

WOFF est un format de police embarquée qui a été proposé au W3C par Microsoft, Mozilla et Opera. Il a été conçu pendant l'année 2009. Le 27 juillet 2010, le W3C publiait un « projet de travail ». Le 13 décembre 2012, WOFF devenait une « recommandation » W3C, un standard Internet.

Fondamentalement, WOFF est une enveloppe qui recouvre les polices de type *sfnt* (à savoir TrueType, OpenType ou Open Font Format) qui ont été comprimées à l'aide d'un outil qui leur permet d'être incorporées dans des pages HTML. Ce format utilise une compression *zlib* qui assure habituellement une réduction de plus de 40 % par rapport à la police TTF équivalente.

WOFF est pris en charge par Firefox depuis sa version 3.6 et par Chrome de Google depuis la version 6.0. Microsoft, pour sa part, la prend en charge depuis la version 9 de son navigateur. Enfin Opera, le permet depuis sa version 11.10.

WOFF en tant que tel ne contient aucun mécanisme de sécurité qui empêcherait la copie de la police. Certains navigateurs, toutefois, ne téléchargent que les polices qui sont hébergées dans le même domaine que la page qui y fait référence. De même, l'optimisation qui consiste à ne sélectionner et envoyer sur le réseau que le sous-ensemble de glyphes correspondant aux caractères utilisés dans la page HTML demandée (ou un ensemble de pages) est laissée à des outils tiers qui ne font pas partie de la norme.

- ***Autres formats***

Malgré le fait que les nouvelles versions des grands navigateurs prendront toutes en charge le format WOFF, il est important de considérer le parc actuel des navigateurs internet, trois autres types de polices téléchargeables y sont utilisés :

TTF	Fonctionne bien avec la majorité des fureteurs (mais pas IE, ni l'iPhone), volumineux car non comprimé.
EOT	Uniquement sur IE, nécessaire pour IE 5 à 8, comprimé, permet de ne sélectionner qu'un sous-ensemble de glyphes qu'on retrouve dans une

[10] Les quatre grands moteurs de composition sont *Trident* de Microsoft, *Gecko* de Mozilla, *Webkit* utilisé notamment par Google et Safari et *Presto* d'Opera. Chacun de ces moteurs de rendu est utilisé dans plusieurs applications : Presto se retrouve ainsi, non seulement dans le fureteur Opera, mais aussi dans des produits Nintendo, Nokia, Sony, Adobe et Macromedia.

[11] Parmi ces sites, on peut citer : <typekit.com>, <fontsquirrel.com>, <typotheque.com>, <openfontlibrary.org>, <fontshop.com> et <fontfont.com>.

	page HTML donnée.
SVG	Format XML nécessaire pour l'iPhone et l'iPad avant la version 4.2, volumineux.

- ***Hapax Berbère et Hapax Tifinar Carrée***

Les polices Hapax Berbère allégée et Hapax Tifinar Carrée sont disponibles sous quatre formats (TTF, EOT, SVG et WOFF) respectivement aux adresses suivantes <hapax.qc.ca/essai-polices-incor/hapaxber-sousensemble-webfont.zip> et <hapax.qc.ca/extrait-foucault/hapaxtifinarcarree.zip>.

La police Hapax Berbère allégée ne comprend qu'un sous-ensemble restreint de la police Hapax Berbère alors que la police Hapax Tifinar Carrée est complète (et donc assez volumineuse). On peut les copier et les utiliser à loisir. Deux pages de démonstration permettent de voir ces polices Web en action : <hapax.qc.ca/essai-polices-incor/HapaxBerbereRegular-demo-hex.html> et <hapax.qc.ca/extrait-foucault/foucault-p339.html>.

- ***@font-face et CSS***

Revenons à notre extrait du dictionnaire touareg de Charles de Foucauld. Comment s'assurer que la page s'affiche correctement tant sur les anciennes versions de MSIE, les iPhone, les iPad que les nouvelles versions de Firefox ?

La première étape est de créer un ensemble de polices TTF, SVG, EOT et WOFF. Un site comme Font Squirrel[12] permet de fournir sa police en entrée et de créer des versions SVG, EOT, TTF et WOFF de celle-ci ainsi que de ne sélectionner qu'un sous-ensemble des glyphes de la police d'origine afin de réduire encore la taille des fichiers produits.

Dans notre cas, la police Hapax Tifinar Carrée utilisée dans la page du dictionnaire Charles de de Foucauld a déjà été transformée en ces quatre formats. Il suffit de télécharger l'ensemble[13], de les dézipper et de les placer dans un répertoire du serveur. Le code fourni ci-dessous assume que les polices sont présentes dans le même répertoire que les feuilles de style CSS.

Il faut ensuite définir la police Hapax Tifinar Carrée dans CSS à l'aide d'une déclaration `@font-face`. C'est ce que fait le code ci-dessous. Il définit cette même police dans deux déclarations que les principaux navigateurs comprendront. Si les polices ne sont pas stockées dans le même répertoire que le code CSS, on n'oubliera pas de changer les `url()` pour y indiquer l'URL qui correspond à leur position.

[12] Voir <fontsquirrel.com/fontface/generator>. Font Squirrel fournit également les @font-face à inclure dans vos feuilles de style.

[13] Disponible ici : <hapax.qc.ca/extrait-foucault/hapaxtifinarcarree.zip>.

```
 @font-face {
       /* Déclaration destinée à Internet Explorer
*/

       font-family: 'HapaxTifinarCarree';
       src: url('hapaxtifinarcarree.eot') ;
 }

 @font-face {
       /* Déclaration destinée à tous les autres */

       font-family: 'HapaxTifinarCarree';
       src:     url(//:) format('pasde404'),
            url('hapaxtifinarcarree.woff') format ('woff'),
            url('hapaxtifinarcarree.ttf')
format('truetype'),
            url('hapaxtifinarcarree.svg#webfontrYQ4E2jU')
                                        format('svg') ;
       font-weight: normal ;
       font-style: normal ;
  }
```

Extrait du dictionnaire touareg-français de Charles de Foucauld (page 339)

□□ éferi □□□ sm. φ (pl.iferân □□□), daṛ ăferi (ĕferi), daṛ ferân || aiguille rocheuse (quelconque) || diffère d'ăḍaouḍa « aiguille rocheuse très mince (ressemblant à un doigt) ». Tout ăḍaouḍa est un éferi, mais non inversement.

— téferit □□□□ sf. φ (pl. tiferâtîn □□□□□), daṛ tăferit (tĕferit), daṛ tferâtîn || dim. du pr.

□□ tăfirt □□□ sf. φ (pl. tifir □□□), daṛ tfir || mot (une syllable ou plusieurs syllabes réunies exprimant une idée) || p. ext. « quelques paroles (paroles en petit nombre) » || p. ext. « vers (assemblage de mots rythmés d'après des règles déterminées en poésie) » Le sing. tăfirt signifie un vers, le pl. tifir signifie 2 ou plusieurs vers.

□□ aferra || v. □□□ afri .

Nous reviendrons sur ces déclarations par la suite. Pour l'instant, il suffit de savoir que ces déclarations permettent aux navigateurs de savoir où trouver la définition d'une police nommée HapaxTifinarCarree.

Pour employer la police dans des documents HTML, il nous reste à préciser en CSS quand l'utiliser. Le code ci-dessous précise, par exemple, que les paragraphes

(<p>) devront d'abord utiliser la police HapaxTifinarCarree, puis si des glyphes manquent d'employer la police Arial et enfin, par défaut, n'importe quelle police sans empattements.

```
 p {font-family: 'HapaxTifinarCarree', Arial, sans-
serif;}
```

Dans le cas de l'extrait du dictionnaire touareg, nous désirons d'abord utiliser des polices à empattements hautement optimisées comme Times New Roman pour le texte latin avant d'utiliser les glyphes de la police Hapax Tifinar Carree dont les glyphes latins sont moins esthétiques.

```
 body {
       font-family : Times New Roman, HapaxTifinarCarree,
serif;
       width: 650px ;
       text-align: justify;
       margin-left: auto ; margin-right: auto ;
 }
```

Ensuite dans le corps de la page HTML, il suffit d'écrire par exemple

```
 <p>— <u>téferit</u> ⵜⴼⵔⵉⵜ sf. φ (pl. <u>tiferâtîn</u>
ⵜⵉⴼⵔⴰⵜⵉⵏ), <u>daṛ tăferit</u> (<u>tĕferit</u>), <u>daṛ
tferâtîn</u> || dim. du pr.</p>
```

pour que les caractères latins s'écrivent à l'aide de la police Times New Roman et les caractères tifinaghes avec la police Hapax Tifinar Carrée comme dans la figure ci-après.

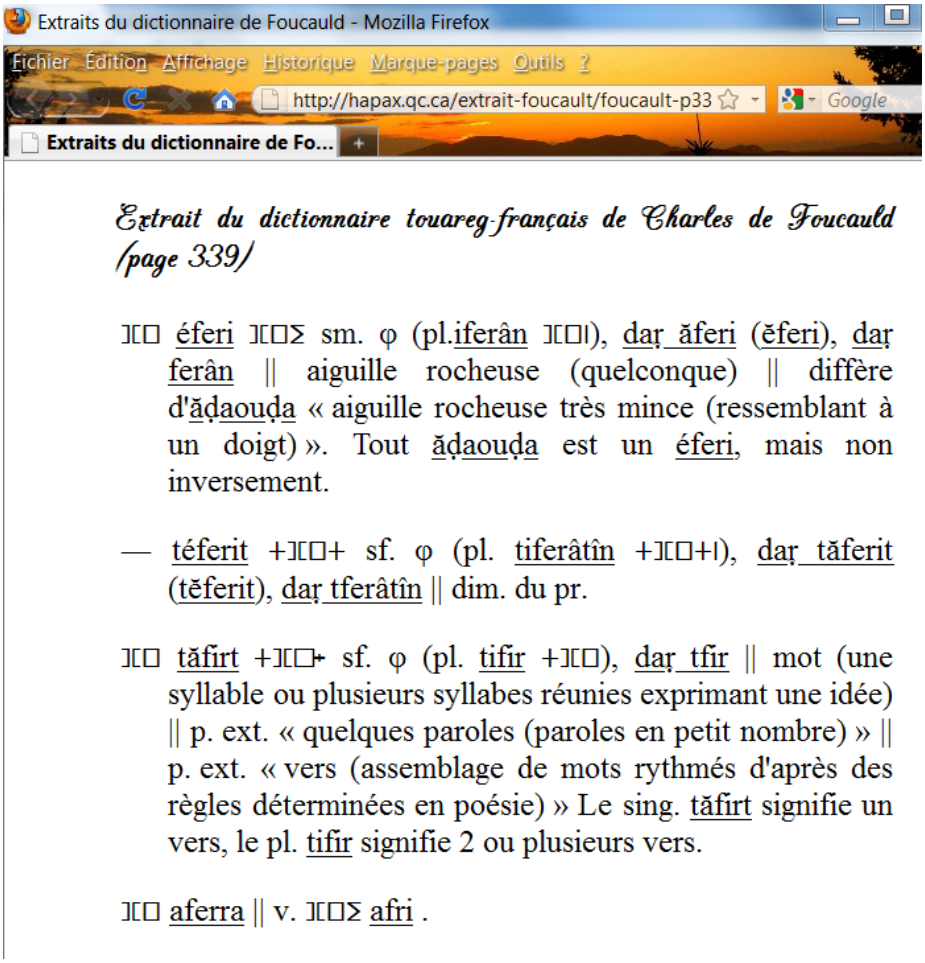


Figure 4 : Page avec des polices Internet incorporées

- ***Retour sur les différents @font-face***

Chaque @font-face permet de déclarer plusieurs fichiers alternatifs pour un même nom de police, en séparant chacun des emplacements, les `url()`, par une virgule :

```
@font-face{
   font-family : 'maPolice';
   src : url('maPolice.woff') format('woff'),
         url('maPolice.svg#abcd') format('svg'),
         url('maPolice.ttf') format('truetype') ;
}
```

La propriété `src` est suivie de l'`url` d'un fichier de police puis d'un `format` optionnel. Déclarer le format permet de préciser explicitement aux navigateurs le type de police associé à l'url qui précède et d'éviter ainsi au fureteur d'avoir à télécharger le fichier en question pour en vérifier le format.

L'url de la police SVG se termine par un nom de fragment après le croisillon (« # »). Ce nom précise où, dans le fichier SVG, se trouve la déclaration de police. À première vue, cela peut paraître superflu, il faut toutefois se rappeler qu'un fichier SVG peut contenir bien d'autres choses qu'une police de caractères ou même en contenir plusieurs.

L'ordre de déclaration des formats est également important : d'abord WOFF, puis selon la taille des fichiers, la version SVG et enfin TrueType. Il est important de considérer l'ordre de déclaration, car certains navigateurs prennent en charge plusieurs de ces formats et utiliseront le premier qu'ils comprennent dans la liste fournie. Or, dans cette liste, WOFF est – comme nous l'avons vu – nettement plus léger que TTF. Le SVG étant un format XML, les fichiers peuvent vite devenir volumineux. Pour palier cet inconvénient, il en existe une version comprimée qui porte le suffixe `.svgz`. Elle est toutefois peu prise en charge par les fureteurs. En outre, si votre serveur gère la compression des requêtes HTTP, cet artifice n'est guère utile. La version compressée d'une police SVG est, en règle générale, plus légère que sa version TrueType.

Pourquoi, dans notre code ne trouve-t-on pas de déclaration comme ci-dessous ?

```
src : url('maPolice.eot') format('eot') ;
```

La raison en est simple : Internet Explorer ignore la propriété `format`[14] et ne prend en charge qu'une valeur d'`url()`. En outre, confronté à la ligne

```
src : url('maPolice.woff') format('woff') ;
```

Internet Explorer cherche à télécharger une police du nom suivant : « `maPolice.woff') format('woff` ». MSIE envoie une requête dans ce sens au serveur et attend que le serveur lui réponde. Il le fera à l'aide d'un 404 (pas trouvé), car bien évidemment il n'existe pas de telle police sur le serveur. Cet appel au serveur est inutile et inefficace.

C'est pourquoi[15], pour les versions 4 à 8 de MSIE, on ajoute une déclaration supplémentaire :

[14] Ceci est corrigé à partir de la future version 9 d'Internet Explorer.

[15] Pour plus de détails sur les bogues d'IE en la matière, voir <http://covertprestige.info/css/font-face/>.

```
 @font-face {
       /* Déclaration destinée à Internet Explorer
*/

       font-family: 'HapaxTifinarCarree';
       src: url('hapaxtifinarcarree.eot');
 }
```

et l'on ajoute cette première ligne dans la deuxième déclaration[16] :

```
       src:     url(//:) format('pasde404'),
```

afin que la requête inutile d'IE échoue avant l'envoi de celle-ci, car aucun protocole ni aucun serveur ne correspond à l'url fourni.

Les bibliothèques logicielles

Grâce à la normalisation des tifinaghes il y a 8 ans et à leur inclusion dans Unicode, aujourd'hui les bibliothèques logicielles permettent de traiter les textes écrits en tifinaghes. Si Java 6 ne connaît pas les propriétés des caractères tifinaghes, ce n'est pas le cas d'ICU[17], une bibliothèque Unicode[18] qui sert de terrain d'essai à Java et dont les fonctionnalités sont habituellement intégrées avec un certain retard dans Java.

On trouvera ci-dessous un exemple de programme Java qui permet de découper un texte Unicode en « mots ». On remarquera que la syntaxe du BreakIterator ICU utilisée ici est identique à celle de Java. Il suffit souvent pour accéder aux fonctionnalités d'ICU de remplacer l'import de Java (ici java.text.BreakIterator) par son équivalent ICU (com.ibm.icu.text.BreakIterator) :

```
import com.ibm.icu.text.BreakIterator;

// import java.text.BreakIterator;


public class AnalyseMotsTifinaghes
{
     public static void main(String[] args)
     {
```

[16] On préférera cette astuce à celle qui consiste à ajouter une propriété `local()` qui plante le navigateur du système d'exploitation Android utilisé sur des téléphones intelligents : <readableweb.com/best-practice-for-font-face-css-takes-a-turn>.

[17] On peut télécharger ICU pour Java ici : ici <download.icu-project.org/files/icu4j/>

[18] Pour une introduction à ICU, voir le chapitre 12 d'*Unicode 5.0 en pratique*, Dunod (2008), Paris.

```
        String texteÀExaminer = "ⴰⵙⵎⵓⵔⴰⵙ ⵏ ⵜⵎⵣⵡⴰⵔⵓ ⴰⵎⴰⵣⵉⵖ ⵏ ⵍⵎⵖⵔⵉⴱ.";
        int début=0;

        BreakIterator frontière =
BreakIterator.getWordInstance();

        frontière.setText(texteÀExaminer);

        début = frontière.first();

        for (  int fin = frontière.next();
             fin != BreakIterator.DONE;
             début = fin, fin = frontière.next())
        {

     System.out.println(texteÀExaminer.substring(début,fin)
);
        }
      }
  }
```

Ce programme divise le texte en mots en se basant sur les propriétés[19] Unicode des caractères (est-ce une lettre, un signe de ponctuation, un chiffre, un espace ?) et produit le résultat suivant :

```
ⴰⵙⵎⵓⵔⴰⵙ
ⵏ
ⵜⵎⵣⵡⴰⵔⵓ
ⴰⵎⴰⵣⵉⵖ
ⵏ
ⵍⵎⵖⵔⵉⴱ
ⵏ
ⵜⵎⵣⵡⴰⵔⵓ
ⴰⵎⴰⵣⵉⵖ
ⴰⵎⵙⵉⵍ
.
```

ICU comprend bien plus que cette fonctionnalité, il permet notamment de trier des textes tifinaghes et définit des données de localisation (nom de mois, format

[19] Voir le chapitre 4 d'*Unicode 5.0 en pratique*, Dunod (2008), Paris.

des dates, etc.) en tifinaghes. .NET de Microsoft offre des fonctionnalités similaires.

Le répertoire de données de localisation CLDR

Depuis l'homologation de l'indicatif [zgh] par l'ISO 639-2, il est désormais possible de l'utiliser dans des dépôts de données numériques ou un « Répertoire de données de paramètres régionaux classiques » comme CLDR. Le CLDR contient des informations spécifiques aux paramètres régionaux (les « locales » en jargon informatique) fournies pour faciliter l'adaptation des applications informatiques à des langues ou des cultures différentes. Le CLDR comprend donc :

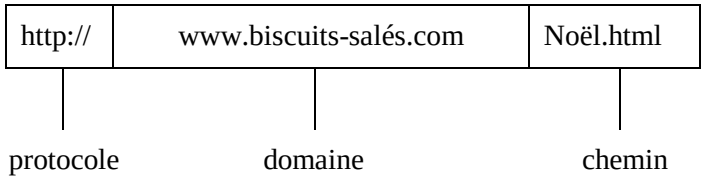
- des traductions pour les noms de langues ;
- des traductions pour les noms de territoires et de pays ;
- des traductions pour les noms de monnaies, y compris les modifications singulier/pluriel ;
- des traductions pour les jours de la semaine, les mois, dans des formes complètes et abrégées ;
- des traductions pour les fuseaux horaires et des exemples de villes (ou similaire) pour les fuseaux horaires ;
- des traductions pour les périodes du calendrier ;
- des motifs pour formater/reconnaître des dates ou des heures, etc.

Grâce à CLDR (dont les données sont notamment intégrées dans ICU) il devient facile pour un programmeur (même s'il ne connaît pas l'amazighe) de concevoir des programmes qui pourront afficher, par exemple, des dates en français et en amazighe standard du Maroc. Pour afficher la date en français, il précisera simplement la valeur ISO 639 « fr » comme « locale », pour l'afficher en amazighe standard il utilisera « zgh ».

Des données initiales pour [zgh] ont été versées dans le répertoire CLDR, elles devraient, toutefois, être complétées et révisées dans les mois à venir.

Adresses internet, courriel et noms de domaine internationalisés (NDI)

Lors du colloque de l'IRCAM en 2008 (Andries, *op.cit.*), nous avions déjà abordé le sujet des adresses internet internationalisées du type <http://www.biscuits-salés.com/Noël.html>. Techniquement on nomme ce type d'adresse des identificateurs de ressource internationalisée (IRI). Ces adresses se divisent en trois parties principales, la première avant le « :// » indique le protocole à utiliser (ici http), ensuite vient le nom de domaine proprement dit puis, après un « / », le chemin de la ressource sur le serveur identifié par le nom de domaine.



Tous les domaines de tête (le *.ma*, le *.fr* ou *.com* final du nom de domaine) n'acceptent pas ces adresses internationalisées. Certains domaines de tête comme la Suisse (*.ch*) acceptent des accents français (et plus généralement les lettres accentuées de ses langues nationales) alors que la France n'accepte pas de tels accents dans les domaines qui se terminent par *.fr* . La politique d'attribution de ces noms est décidée par l'autorité d'enregistrement responsable d'un domaine de tête particulier. Au Canada, il s'agit d'une agence sans but lucratif (l'ACEI) régie par les lois canadiennes. Les domaines de tête génériques *(.com, .org, .biz*) sont administrés aux États-Unis en vertu des lois américaines.

Comme nous le mentionnions déjà en 2008, les navigateurs modernes transformeront de manière transparente les noms de domaine comme *biscuits-salés* en des noms de domaine compatibles avec l'infrastructure actuelle prévue pour des adresses ASCII.

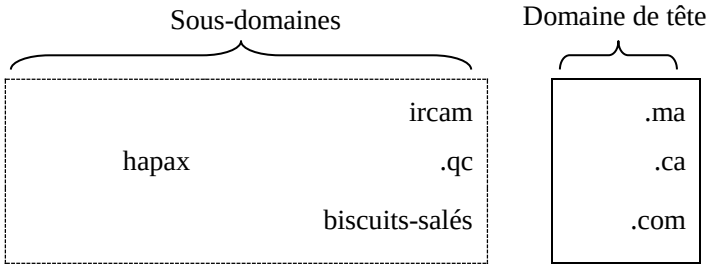


Figure 5 – Parties d'un nom de domaine

Ce qui a changé depuis ce colloque, c'est l'approbation d'IDNA 2008 qui a mis à jour le répertoire des caractères permis, passant d'Unicode 3.2 à Unicode 5.2. Par ce fait même, tous les caractères tifinaghes marocains peuvent, en théorie, être utilisés dans les adresses Internet. Il est important d'insister sur la dimension théorique de cette avancée, car l'acceptation de ces caractères dans les noms de domaine dépend des autorités d'enregistrement de noms de domaine. Pour que des sous-domaines en tifinaghes soient acceptés dans des adresses du domaine de tête `.ma`, il faudra donc que l'organisme responsable de l'attribution des noms de domaine au Maroc le permette.

Conclusion

Depuis 2005 et la normalisation du tifinagh dans Unicode et l'ISO/CEI 10646, beaucoup de choses ont changé. Il est de plus en plus facile d'utiliser des tifinaghes sur internet et dans les applications informatiques.

Il reste encore, bien sûr, dans ce domaine quelques améliorations possibles. On peut citer des pistes comme le fait de s'assurer que des claviers et des polices tifinaghes soient disponibles sur les téléphones intelligents et les tablettes informatiques, la diffusion et l'emploi de CSS 3 qui permettra une typographie fine des polices, embarquées ou non, la mise en place concrète d'IDNA 2008, la création et la diffusion de ressources de localisation en tifinaghes et amazighe, la mise à jour du clavier tifinaghe pour inclure les nouveaux caractères tifinaghes Unicode, etc. Dans ces domaines, les choses avancent et des solutions techniques existent déjà ou se dessinent pour que l'utilisation des tifinaghes et de l'amazighe dans tous ces domaines se fasse désormais sans entrave. En huit ans, que de chemin parcouru !

Références bibliographiques

Foucauld Ch.-de (1951), *Dictionnaire touareg–français, dialecte de l'Ahaggar*. 4 volumes, Imprimerie nationale, Paris.

Karl-G. Prasse (1972), *Manuel de Grammaire touarègue*, Éditions de l'Université de Copenhague.

Zenkouar L. (2004), « L'Écriture amazighe tifinagh et Unicode », *Études et Documents Berbères,* vol. 22, p. 173-175.

Andries P. (2008), *Demain encore plus de tifinaghes sur Internet*, Actes du colloque du CEISIC (NTIC), Rabat.

Asinag-Asinag, 9, 2014, p. 33-48

La promotion de l'amazighe à travers les technologies de l'info rmation et de la communication

Fadoua Ataa Allah & Siham Boulaknadel

CEISIC, IRCAM

أدرك المعهد الملكي للثقافة الأمازيغية بمدى أثر العولمة والنهضة العلمية في تغيير وضع لغات وثقافات العالم، جعله يضع سياسة تهدف إلى حوسبة اللغة الأمازيغية ابتداء من إنجاز الآليات المساعدة على كتابتها وعرضها رقميا ووصولا إلى الآليات المتطورة التي تسمح بتصحيح الأخطاء الإملائية والنحوية ودعم الترجمة الفورية.

في هذا السياق، يتناول هذا المقال الأهداف الإستراتيجية المساهمة في تعزيز مكانة اللغة والثقافة الأمازيغيتين، ويقترح على ضوء تكنولوجيا المعلومات والاتصالات خريطة طريق للنهوض بهما. كما يقدم لمحة عامة تشمل أهم الإنجازات التي حققتها اللغة الأمازيغية في هذا الصدد خلال العشرية الأخيرة.

1. Introduction

La langue amazighe constitue un élément éminent de la culture marocaine et ce, par sa richesse et son originalité. Elle a été négligée sinon écartée, pour des générations, en tant que source d'enrichissement culturel. Cependant, la création de l'Institut Royal de la Culture Amazighe (IRCAM), suivie de la constitutionnalisation lui a permis désormais de jouir d'un statut meilleur. Néanmoins, nous sommes tous conscients que l'officialisation n'est pas suffisante pour assurer la revitalisation de cette langue, particulièrement durant cette ère de globalisation et de mondialisation technologique. D'où la nécessité de penser à la promotion avec l'esprit des jeunes générations pour qui les Technologies de l'Information et de la Communication (TIC) sont devenues un support de vie social.

La notion « Technologies de l'Information et de la Communication » couvre un vaste éventail de technologies allant des techniques nécessaires pour le traitement de l'information aux technologies de transmission et de diffusion de cette information. A l'ère de la révolution numérique, ces technologies constituent une arme à double tranchant : un segment majeur dans la survie d'une langue ou un facteur aggravant sa marginalisation. En effet, la prédominance des langues des pays industrialisés risque, si aucun remède n'est pris, d'accentuer dramatiquement l'écart technologique. Toutefois, l'accès aux technologies modernes de l'information offrira de nouvelles opportunités pour la préservation et la revitalisation des langues et cultures régionales.

L'amazighe comme la majorité des langues africaines a accumulé un retard considérable dans ce domaine. D'où la nécessité de le combler à travers des actions incitatives de grande ampleur, exploitant les TICs en vue de réajuster et d'accélérer la promotion de cette langue. A cette fin, il convient de développer une méthodologie d'intégration. Cette optique nous amène à réfléchir sur l'apport et les moyens de l'utilisation des TICs sur le plan des différents axes stratégiques de la promotion de la langue amazighe :

- Quelles mesures faut-il prendre pour contribuer à la préservation du patrimoine amazighe par le biais des TICs ?
- Que peut apporter l'intégration des TICs au processus d'aménagement linguistique de l'amazighe afin de sauvegarder la diversité linguistique marocaine ?
- Comment les nouvelles technologies peuvent-elles promouvoir l'apprentissage de la langue amazighe ?
- Comment investir dans l'usage social des médias numériques au profit de la langue et la culture amazighes ?

C'est à ces questions, entre autres, que cet article souhaite apporter des éléments de réponse. Il s'agit de proposer des choix méthodologiques et des principes à retenir dans le processus de la préservation et de la modernisation de la langue amazighe. Nous étayerons notre contribution par des exemples concrets issus de recherches récentes dans le domaine des TICs, et plus spécifiquement celles en traitement automatique des langues qui consiste en la création d'outils et de ressources linguistiques par l'intégration de fonctionnalités du traitement du langage humain.

2. L'utilité des TICs pour la pérennité de la culture et la langue amazighes

Le développement et la mondialisation des technologies de l'information et de la communication offrent à toutes les langues du monde un espace globalisé de communication et d'échange. Dans cette perspective, nous introduisons quelques retombées de l'exploitation des TICs pour l'amazighe.

2.1. Les TICs au service de la sauvegarde du patrimoine amazighe

La langue amazighe comme toute langue africaine dite de tradition orale, dont le patrimoine culturel, scientifique et historique ne se transmettait de génération en génération qu'à travers la mémoire et les traditions, ce qui l'a menacée de la destruction des particularités de sa culture et de son identité nationale. D'où l'intérêt de recenser l'ensemble des éléments patrimoniaux et de pouvoir gérer leur aspect descriptif *via* une banque de données multimédias évolutive répondant à un double objectif de conservation et de valorisation de ce patrimoine vernaculaire. En effet, l'élaboration d'une telle banque de données numérique résoudra le problème de la conservation et contribuera à assurer la pérennité de cet héritage ancestral sous une forme perceptible d'une part et d'autre part reflétera la richesse et la diversité de ce patrimoine tout en garantissant une bonne structuration de ses

composants matériels et immatériels, une meilleure accessibilité universelle et une résurrection de l'intérêt du grand public.

A cette fin, une étape de numérisation est indispensable pour la création de cet espace culturel, suivie d'une phase d'archivage sous un format standard permettant la structure du contenu brut et des métadonnées nécessaires à la gestion. Cette phase repose sur un ensemble d'outils et de méthodes de traitement automatique des langues, à savoir :

- Le transcodage pour en faire un contenu intelligible dont le codage est particulièrement compatible aux évolutions technologiques ;
- L'indexation qui sert à identifier et sélectionner les métadonnées pertinentes décrivant le contexte patrimonial ;
- La classification pour regrouper et classer ce contenu numérique selon le support, le type, le thème et la région.

Par ailleurs, afin de permettre une exploitation effective de ce contenu et de diversifier les supports de stockage, une démarche de transcription automatique ou d'océrisation est essentielle à entreprendre pour accélérer la conversion des archives audio et image en documents sous format texte, suivie par la translittération afin de permettre d'unifier la graphie utilisée pour représenter le contenu.

2.2. Les TICs au service de l'aménagement linguistique de l'amazighe

Après des décennies de négligence, la création de l'Institut Royal de la Culture Amazighe a motivé la revitalisation de la langue amazighe par un processus d'aménagement linguistique qui englobe un éventail d'approches, y compris la codification graphique, l'établissement de règles d'orthographe, de grammaire et de lexiques, la terminologie et la traduction.

La réussite d'un aménagement linguistique dépend de sa visibilité. C'est pourquoi l'utilisation des TICs constitue de nos jours un levier dans ce processus permettant de valoriser les efforts fournis et d'agir sur les présentations linguistiques et ce, à travers :

- L'internet pour contribuer à la diffusion des productions de recherche et faciliter l'échange des connaissances et le suivi de l'évolution du processus d'aménagement linguistique pour d'autres langues dans des situations similaires.
- La codification graphique qui définit une norme standard et une police de caractères adéquate de la graphie afin d'assurer l'exploitation et le partage des contenus.
- La gestion des ressources linguistiques à savoir les corpus, les lexiques et la terminologie, dont le développement est hautement interdisciplinaire, exige la coopération de spécialistes de nombreux domaines. Étant donné la quantité considérable de ressources linguistiques et le nombre de personnes

qui devraient travailler ensemble en unissant leurs efforts ainsi que leurs ressources, les bases de données présentent un outil essentiel de gestion et de structuration dans le processus d'aménagement linguistique.

- L'extraction automatique de terminologie qui permettra l'attestation des termes dans différents genres textuels et l'enregistrement de l'usage des termes en contextes, et ce en exploitant la richesse des ressources langagières telles que les corpus textuels.
- La traduction automatique qui assurera une meilleure compréhension entre des locuteurs de langues différentes et contribuera à la diffusion de la culture à l'échelle internationale en utilisant des outils d'analyse, de génération et de transfert à l'aide des corpus multilingues.

2.3. Les TICs au service de l'enseignement

Malgré l'initiative très significative de l'intégration de l'amazighe au sein du système éducatif marocain en 2003, elle n'est toujours pas suivie de faits concrets permettant à cette langue d'atteindre le niveau de l'arabe ou du français. Pour tenter de remédier à cet état de fait, l'application des TICs à l'enseignement est une solution primordiale pour notre communauté pour accélérer les échanges de savoir et réduire le retard accumulé dans ce domaine.

Dans une situation dominée par l'insuffisance du nombre des enseignants et formateurs et du matériel pédagogique adapté, ainsi que par la restriction de l'expérience aux classes primaires et à un certain nombre limité d'établissements, il est judicieux de pouvoir profiter suffisamment des retombées de l'utilisation des TICs au profit de l'éducation à travers :

- L'apprentissage en ligne qui présentera une piste pertinente pour favoriser la promotion des ressources humaines et la généralisation verticale et horizontale, en résolvant le manque de formateurs par le rassemblement des enseignants dispersés à travers le pays ;
- Les outils multimédias et d'interactivité qui encourageront l'apprentissage autodirigé par la garantie de la correction informatisée, faciliteront la création et amélioreront la qualité des contenus éducatifs tout en assurant la transmission du savoir porteur de la diversité des formes d'expression culturelle ;
- L'enseignement hybride basé sur la conjonction des activités en ligne et en classe qui permettra de favoriser le potentiel motivationnel des apprenants et d'améliorer la qualité de l'enseignement ;
- Un réseau professionnel d'enseignants qui facilitera l'échange du savoir et le débat autour des besoins éducationnels et des utilitaires de production d'outils pédagogiques.

Dans cette perspective, il est important de pouvoir recenser les mécanismes et les techniques nécessaires pour mener à bien la généralisation de cette vocation et la mise en place d'une plate-forme d'apprentissage en ligne assurant la modernisation des moyens pédagogiques, l'élargissement du savoir ainsi que le partage des

connaissances et favorisant l'émergence d'un enseignement de qualité. Toutefois, l'élaboration d'une telle plate-forme dépend principalement du contenu pédagogique, dont les supports du cours sont nécessairement accompagnés par des activités pédagogiques dont la linguistique computationnelle constitue un fondement important, et ce par :

- L'exploitation de la richesse des ressources langagières telles que les corpus textuels à travers des concordanciers permettant l'exploration de la diversité des situations langagières et l'extraction d'exemples authentiques dans le but d'étudier le sens et les règles d'emploi ;
- L'utilisation de correcteurs orthographique et grammatical facilitant l'apprentissage autodirigé en assurant l'automatisation de l'analyse des productions des apprenants ;
- L'exploitation des correcteurs phonétiques pour soutenir l'acquisition de la prononciation à travers des entraînements systématiques.

2.4. Les TICs au service des médias

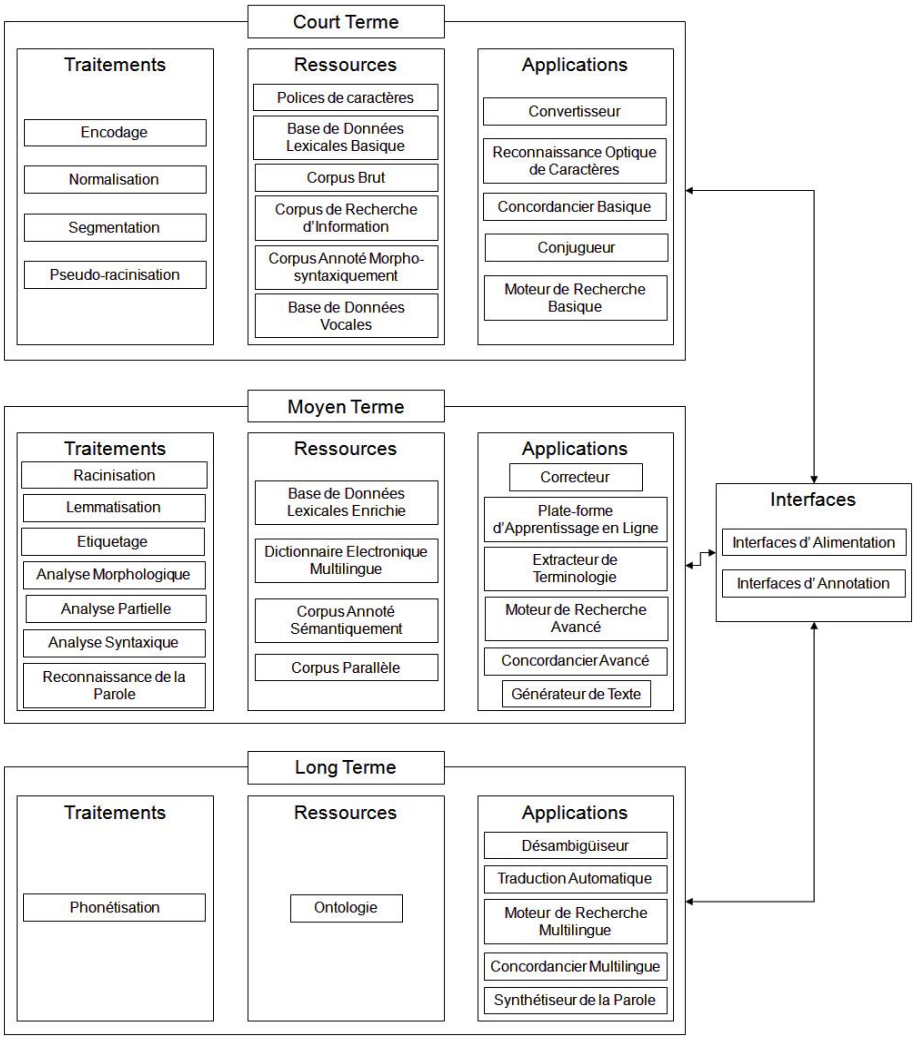
Les technologies de l'information et de la communication, principalement l'internet, ont permis de diffuser plus efficacement les moyens de communication et d'améliorer le traitement, la mise en mémoire, la diffusion et l'échange de l'information. Néanmoins, cette émergence a suscité une mutation technique des médias classiques (télévision, radio, presse, cinéma) aux niveaux de leurs chaînes de valorisation allant de la production jusqu'à la consommation de l'information, et ce à travers le passage de l'analogique au numérique et le développement des techniques du traitement de l'information, qu'elle soit écrite, orale ou multimédia, monolingue ou multilingue pour sélectionner, classer, structurer et rechercher.

Cependant, la place de l'amazighe dans les principaux médias a toujours été et reste très limitée. Dans la perspective de faire face à ce défi et de faire de l'amazighe une langue de communication, il faudrait, d'une part, adopter les outils multimédias pour valoriser et rénover son contenu médiatique et, d'autre part, exploiter les différentes plate-formes de diffusion comme les sites Web, les séminaires et les plate-formes de vidéo en ligne, les médias sociaux et les blogues pour une présentation attrayante de ce contenu.

3. Méthodologie pour l'informatisation de l'amazighe

Dans l'objectif de concrétiser l'apport des TICs aux différents axes stratégiques de l'Institut, nous avons mené une étude pour l'élaboration d'une méthodologie de mise en œuvre assurant la revitalisation et la promotion de la langue amazighe. Cette méthodologie repose sur les travaux élaborés pour l'informatisation des langues peu dotées (Berment, 2004 ; Scannell *et al.*, 2012) ainsi que sur une cartographie des différents traitements et ressources disponibles pour des langues peu, moyennement et bien dotées (Ataa Allah, 2010), afin de définir les priorités permettant de planifier une feuille de route dotant l'amazighe d'outils assurant son fonctionnement comme les langues conformément équipées (Ataa Allah et Boulaknadel, 2012).

La méthodologie proposée structure le développement du projet de l'informatisation de l'amazighe en tâches fondamentales, conçues d'une manière à ce qu'elles se déroulent en trois périodes (courte, moyenne et longue) qui peuvent s'étendre sur 5 à 10 ans selon la disponibilité humaine et les ressources linguistiques. Ces tâches sont représentées par une chaîne qui part des traitements élémentaires allant vers des applications génériques répondant généralement à des besoins et passant par la constitution de briques de ressources linguistiques (Figure ci-dessous). Cette chaîne va de l'adaptation et l'amélioration d'outils en fonction des nouvelles technologies jusqu'au développement d'applications, en se basant sur la recherche fondamentale.



Feuille de route pour l'informatisation de l'amazighe

3.1. La phase à court terme

Cette phase constitue une première étape dans le processus de l'informatisation de l'amazighe. Elle est initiée par un codage de caractères, un clavier et des polices de caractères, assurant la constitution de ressources électroniques telles que les bases de données lexicales et vocales, les corpus de recherche d'information, brut et annoté morpho-syntaxiquement. Ces ressources servent de briques élémentaires pour l'élaboration d'outils et d'applications du traitement du langage humain, en particulier celles planifiées pour cette phase, à savoir la normalisation, la segmentation, la pseudo-racinisation, le concordancier de base, le moteur de recherche, le conjugueur et le système de reconnaissance optique de caractères.

3.2. La phase à moyen terme

Après avoir établi les ressources de base et les outils fondamentaux, cette phase repose sur un traitement linguistique qui consiste en une réalisation d'une racinisation, une lemmatisation, un étiquetage morpho-syntaxique automatique, une analyse morphologique, partielle et syntaxique, et une reconnaissance de la parole. Ces traitements permettront de développer des applications avancées telles qu'un correcteur orthographique, une plate-forme d'apprentissage en ligne, un extracteur terminologique, un moteur de recherche et un concordancier avancés ainsi qu'un générateur de textes. En outre, cette phase représente une étape clef de préparation des ressources nécessaires pour la prochaine étape, y compris les bases de données enrichies, les dictionnaires multilingues, et les corpus alignés et annotés sémantiquement.

3.3. La phase à long terme

La troisième étape de la feuille de route pourrait être considérée comme une synthèse du travail réalisé. A côté d'un traitement de phonétisation et d'une élaboration d'une ontologie, d'un développement d'un désambiguïseur sémantique et d'un synthétiseur de la parole, cette phase se concentre également sur la réalisation d'applications multilingues, principalement la traduction automatique.

4. Stratégies technologiques à respecter

Dans la perspective de mener à bien le développement de ressources et d'outils nécessaires pour l'informatisation d'une langue, des études ont porté sur des stratégies technologiques qui devraient être respectées. Ainsi, nous nous sommes inspirées des travaux de Muhirwe (Muhirwe, 2007) qui suggère de prendre en considération lors du développement des applications la notion d'extensibilité et de documentation ainsi que l'adoption des technologies libres, pour dresser nos stratégies de développement préalablement nécessaires pour réussir et accélérer le processus de l'informatisation de l'amazighe.

4.1. Standardisation des ressources

L'élaboration des ressources numériques est une étape indispensable dans tout processus d'informatisation. Cependant, cette tâche est coûteuse en termes de temps et de compétences humaines. D'où l'utilité de les concevoir sous un format standard permettant à la communauté académique impliquée dans la réalisation ou l'intégration d'outils et ressources, de jouir d'un environnement garantissant la gestion et l'interopérabilité entre de tels composants.

4.2. Adaptation des technologies de langues

Une part sérieuse du temps et d'efforts nécessaires pour informatiser une langue est généralement consacrée à la réalisation de l'environnement. Cependant, cette tâche laborieuse peut être optimisée par l'intégration et l'adaptation des techniques existantes afin de profiter pleinement du potentiel qu'offrent ces technologies pour la réalisation des tâches de manière générique sans avoir recours à « réinventer la roue ».

4.3. Extensibilité

L'extensibilité est un facteur élémentaire dans la conception de tout projet en technologie des langues, assurant l'extension de ses fonctionnalités. Cette propriété permet une plus grande réutilisabilité du projet en facilitant l'ajout de nouveaux composants et l'amélioration ou l'adaptation de l'existant.

4.4. Logiciels libres

En vue de garantir l'adaptation des technologies de langues et leur extensibilité, il est nécessaire d'adopter un mode de conception et d'exploitation qui obéit à des règles de liberté susceptible d'être soumis à modification et redistribution sans restriction. Ces caractéristiques confèrent une certaine fiabilité et réactivité.

4.5. Documentation

La documentation constitue l'une des pierres angulaires dans le développement d'un projet d'informatisation d'une langue. Elle permet d'expliquer le fonctionnement du projet en déterminant ses objectifs, son architecture ou sa conception, les techniques utilisées pour son développement et son manuel d'utilisation. Ce qui peut assurer sa bonne exploitation, son extensibilité et sa réutilisation logicielle.

4.6. Evaluation des systèmes

Le recours à l'évaluation conduit à repenser les objectifs, les pratiques, voire les approches théoriques. Une évaluation peut être apportée avant la mise en œuvre d'un système, pour déterminer à l'aide d'un diagnostic les objectifs attendus et les indicateurs nécessaires à l'évaluation ; au cours de sa réalisation, afin d'ajuster le

système au besoin à travers une série d'évaluations progressives ; ou après l'achèvement du travail, et ce pour déterminer le niveau de satisfaction, la pertinence, la durabilité et l'extensibilité du système.

Par ailleurs, deux types d'évaluation sont distingués :

- L'évaluation objective fournit une mesure des performances du système en dehors de toute considération sur sa perception par les utilisateurs.
- L'évaluation subjective fondée sur le jugement des utilisateurs et qui consiste à mesurer l'adéquation du système à la tâche ainsi que son utilité en termes de convivialité, fiabilité et facilité d'utilisation.

5. Réalisations

Dans le cadre de la promotion de la langue amazighe, de nombreux travaux ont été réalisés afin de fournir à cette langue des ressources et outils permettant son traitement automatique et son intégration dans le domaine des technologies de l'information et de la communication.

Cette section introduit les différentes réalisations au niveau national [1] en les structurant en sept parties selon la nature du traitement visé, passant par le préalable à l'exploitation des TICs ; outils de recherche, d'analyse et d'assistance ; ressources langagières, localisation logicielle et reconnaissance optique des caractères.

5.1. Préalables à l'exploitation des TICs

En amont de l'exploitation des TICs dans le processus de la promotion de l'amazighe, il faut disposer de données numérisées. A cette fin, il est judicieux de définir un codage adapté, un clavier et des polices de caractères.

5.1.1. Codage de Tifinaghe

Le codage du caractère tifinaghe constitue le premier pas vers l'exploitation des TICs, permettant la transcription de l'amazighe à travers une représentation numérique pour chaque caractère. Ce processus s'est déroulé en deux étapes : la première a consisté en une adaptation de la norme ISO 8859-1 pour coder les caractères tifinaghes en codage ANSI, afin de répondre à l'urgence de l'introduction de l'amazighe dans le système éducatif. Cependant, la portée de ce codage privé de l'IRCAM a été limitée, et la gestion des textes comportant plusieurs systèmes d'écriture était difficile. D'ailleurs, les traitements des textes multilingues doivent jongler à la fois avec les différentes normes de codage et avec les polices associées. D'où la nécessité et l'importance d'intégrer le tifinaghe dans le plan multilingue c'est-à-dire un codage plus portable et facile à gérer tel que l'Unicode qui a permis d'affecter un code unique à chaque caractère (Andries, 2008).

[1] Dans cet article, nous nous sommes limitées au niveau national, néanmoins les réalisations au niveau international pourront faire l'objet d'une autre contribution.

5.1.2. Clavier

L'intégration de l'amazighe dans la norme internationale de la prescription des claviers ISO/CEI 9995 a fixé définitivement les claviers de la langue amazighe conçus pour la bureautique. En fait, cette norme a spécifié deux types de claviers : un clavier de base contenant les caractères tifinaghes préconisés par l'IRCAM et un clavier étendu comportant tous les caractères adoptés par l'ISO.

5.1.3. Polices de caractères

Dans l'objectif d'intégrer la langue amazighe dans le système éducatif et de favoriser la publication assistée par ordinateur, huit polices de caractères, associées au codage ANSI, ont été réalisées. Cette réalisation a été suivie par l'élaboration d'une nouvelle génération de polices, qui a constitué un saut qualitatif de l'universalisation de l'écriture amazighe. Cette génération inclut les polices associées à l'Unicode.

5.1.4. Convertisseur

Dès l'encodage Unicode des caractères tifinaghes, il a fallu préparer des outils pour convertir la représentation de ces caractères d'un codage à un autre. Par ailleurs, dans la perspective de promouvoir la langue amazighe et d'assurer la sauvegarde de son héritage littéraire, il apparaît intéressant de tirer profit des productions amazighes écrites en caractères arabe et latin. A cette fin, un convertisseur a été réalisé, permettant le passage du codage ANSI au codage Unicode et la translittération du caractère arabe et du caractère latin au caractère tifinaghe (Ataa Allah et *al.*, 2013).

5.2. Outils de recherche

Dans la perspective de contribuer à la diffusion de la langue et de la culture amazighes, il est intéressant de développer des applications d'exploration et de recherche.

5.2.1. Moteur de recherche

Les moteurs de recherche sont actuellement la première source d'information sur le Web pour plusieurs domaines et leur utilisation est devenue incontournable. D'ailleurs, les sondages et enquêtes qualitatives effectués en la matière démontrent bien l'importance de ces applications qui répondent aux besoins professionnels et personnels suscités par les internautes[2].

Conscients de ce fait et afin de contribuer à la promotion et la diffusion de la culture et la langue amazighes à travers le Web et d'assister les internautes à découvrir la diversité et la richesse de la culture amazighe, un moteur de recherche supportant la graphie tifinaghe a été développé (Ataa Allah et Boulaknadel, 2010a).

[2] http://www.nielsen-online.com/pr/pr_040223_us.pdf

Ce dernier est basé sur un ensemble de traitements linguistiques, à savoir, l'élimination des mots anti-dictionnaires[3] et l'exploitation de la pseudo-racinisation.

5.2.2. Concordancier

Les concordances sont un mode de présentation d'extraits de texte, contenant le même mot ou le même motif linguistique, basées sur une méthodologie d'analyse textuelle laborieuse. Cependant, le recours aux outils de concordance numérique a facilité cette tâche séculaire que nécessitaient les concordances manuelles. Dans cette perspective, un outil de concordance supportant les scripts arabes, latins et Unicode de l'amazighe a été développé. Il permet d'effectuer des recherches dans des textes numériques et répond au besoin d'exploration suscité par l'usager dans le but d'étudier le sens et les règles d'emploi (Ataa Allah et Boulaknadel, 2010a).

5.3. Outils d'analyse

Afin de doter l'amazighe d'outils d'analyse et de génération, un ensemble de travaux portant sur la morphologie flexionnelle et dérivationnelle a été effectué.

5.3.1. Pseudo-racineur

L'outil de pseudo-racinisation est basé sur une approche relevant du cas de la morphologie flexionnelle et reposant sur l'élimination d'une liste de suffixes et de préfixes de la langue amazighe dans un ordre prédéterminé. Cette approche légère permettant de regrouper les mots sémantiquement proches à partir de ressemblances pourra être facilement exploitée dans des applications telles que la recherche d'information et la classification (Ataa Allah et Boulaknadel, 2010b).

5.3.2. Conjugueur

Dans un contexte de développement d'outils de génération pour la langue amazighe, un conjugueur a été réalisé pour faciliter l'apprentissage de la langue. Cet outil est conçu à la base de modèles de comportement flexionnel augmenté de règles de régularisation morphologiques (Laabdelaoui et *al.,* 2012). Il permet la conjugaison en ligne des formes verbales simples et dérivées dans différents aspects et modes de l'amazighe.

5.3.3. Analyseur morphologique

Etant donné que toute analyse linguistique passe par une première étape d'analyse morpho-lexicale, qui consiste à tester l'appartenance de chaque mot du texte au lexique de la langue, il a été entrepris de développer des systèmes d'analyse morphologique pour l'amazighe, profitant des apports des modèles à états finis et faisant appel à une formalisation du vocabulaire et à des règles grammaticales à

[3] Les mots non significatifs ou non discriminants pour la recherche d'information, tels que les prépositions, les conjonctions et les déterminants.

large couverture. Un premier système a été développé à la base de l'environnement linguistique NooJ. Il s'est focalisé essentiellement sur la formalisation flexionnelle et dérivationnelle de la catégorie nom (Nejme et *al.*, 2013). Cependant, le deuxième système a été réalisé en exploitant l'environnement de développement Xerox en se basant principalement sur la formalisation flexionnelle de la catégorie verbe (Ataa Allah, 2014).

5.4. Ressources langagières

La réalisation de ressources langagières (jeux d'étiquettes, dictionnaires, terminologies, corpus oraux et écrits,…) est une activité de mise en œuvre d'infrastructures qui doit être perçue à un niveau amont étant donné, d'une part, le coût prohibitif d'une telle tâche, et d'autre part le fait de constituer un pré-requis indispensable pour le développement des applications de technologie de langue à haute valeur ajoutée.

5.4.1. Jeux d'étiquettes morpho-syntaxiques

Dans la perspective d'annotation de corpus amazighe, un jeu d'étiquettes morpho-syntaxiques a été élaboré à la base de la *Nouvelle grammaire de l'amazighe* (Boukhris et *al.*, 2008). Ce jeu s'inspire du projet EAGLES[4] où le choix des étiquettes repose sur une distinction entre étiquettes obligatoires, étiquettes recommandées et extension particulière. Ainsi, il est proposé deux listes d'étiquettes spécifiques aux caractéristiques de l'amazighe. L'une contient les étiquettes obligatoires ; l'autre, des étiquettes recommandées fournissant des indications plus précises (Ataa-Allah, 2011).

5.4.2. Dictionnaire

Dans la perspective d'appuyer l'enseignement et l'apprentissage de l'amazighe, un dictionnaire imagier sonore en ligne a été conçu pour les enfants âgés de 2 à 14 ans (Ataa Allah, 2011). Il se base sur l'interaction entre les connaissances linguistiques et extralinguistiques afin d'offrir à l'enfant un simple moyen pour stimuler son langage et son éveil. Ce dictionnaire comporte actuellement 550 entrées lexicales réparties sur 105 catégories regroupées en 18 thématiques. Chaque entrée est déterminée par 4 langues (anglais, amazighe, arabe, français) et représentée par une image et un son.

5.4.3. Terminologie

Dans l'objectif de doter l'amazighe d'une terminologie couvrant le plus grand nombre de champs lexicaux, une base de données terminologiques a été élaborée. Elle comporte des entrées terminologiques relevant de plusieurs thématiques, y compris les lexiques usuels, des médias et grammaticaux ainsi que leurs équivalents arabes et français (El Azrak et El Hamdaoui, 2011).

[4] http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html

5.4.4. Corpus

Dans l'objectif de développer des outils de traitement automatique de l'amazighe et pour pouvoir se fonder sur l'usage réel de la langue, un recueil de textes numérisés et codés a été réalisé.

- **Corpus brut**

  Dans le cadre d'élaboration de corpus électronique pour la langue amazighe, un corpus a été collecté, composé de 160 textes amazighes représentant différents genres littéraires, à savoir conte, conte pour enfants, poésie et articles de presse contenant les sous-genres journal, magasine et Net. Il comprend 136.093 occurrences dont 23.174 sont des mots distincts (Boulaknadel et Ataa Allah, 2011).

- **Corpus étiqueté**

  Les ressources lexicales annotées ne sont que rarement disponibles, particulièrement pour les langues peu dotées telle que l'amazighe. Dans ce contexte, une démarche d'élaboration de corpus étiqueté a été menée. Ainsi, le corpus réalisé comprend environ 20 k mots et se compose de textes extraits d'une variété de sources telles que la version amazighe du site Web de l'IRCAM, le périodique « Inghmisn n usinag » (bulletin d'information de l'IRCAM) et des manuels scolaires (Outahajala et *al.*, 2010).

5.5. Outils d'assistance

Dans l'objectif de faciliter et de mener à bien le processus d'élaboration de ressources langagières, des applications d'assistance ont été réalisées.

5.5.1. Base de données lexicales

Bien que de nombreux dictionnaires papiers de la langue amazighe soient édités, aucune base de données numérique n'est disponible. Pour faire face à ce manque, une application assurant la gestion des collectes et l'exploration des lexiques amazighes a été élaborée. Cette application permet de structurer les informations intra et inter-lexicales telles que la définition, les équivalents arabes et français, les synonymes et le classement par thèmes (Iazzi et Outahajala, 2008).

5.5.2. Outil d'assistance à l'étiquetage morpho-syntaxique

L'utilisation des corpus, notamment ceux annotés morpho-syntaxiquement, est devenue une étape indispensable dans un processus d'informatisation d'une langue. Cependant, cette tâche est d'une grande complexité et son coût en termes de temps et de ressources humaines limite la quantité et la disponibilité des corpus. Afin de minimiser les efforts nécessaires pour produire des corpus amazighes et d'améliorer leurs qualités en permettant des vérifications et en simplifiant les modifications et les mises à jour, un outil d'assistance à l'étiquetage morpho-syntaxique a été développé, permettant une automatisation partielle de la tâche

d'étiquetage et assurant une bonne cohérence entre les différents travaux réalisés par l'ensemble des chercheurs (Ataa Allah et Jaa, 2009).

5.5.3. Base de données littéraires

Dans une perspective de sauvegarde du patrimoine littéraire amazighe, une application visant la collecte et la mise en réseau d'un noyau d'une banque de corpus littéraires, contenant principalement des données textuelles et audiovisuelles amazighes, a été développée. Cette application permet la gestion et la consultation du contenu littéraire (Ait Ouguengay et *al*., 2012).

5.6. Localisation logicielle

A travers la localisation de l'interface de deux cellulaires Sony Ericsson « J110i et J120i » et du système d'exploitation Microsoft « Windows 8 », une opportunité a été offerte à l'amazighe pour assurer sa présence dans le monde technologique et offrir à l'utilisateur un environnement familier, où l'expérience de l'intégration de la langue amazighe et son système d'écriture tifinaghe a été menée.

5.7. Reconnaissance optique des caractères

Dans une vision de contribuer à la sauvegarde et la numérisation du fonds documentaire amazighe, de nombreuses études ont été menées sur la reconnaissance optique de caractères amazighes dont le taux de reconnaissance est autour de 92%. Ces études se basent sur différentes approches pour la reconnaissance du caractère imprimé et manuscrit. Elles sont regroupées généralement en trois classes, à savoir les réseaux de neurones (Ait Ouguengay, 2009), (Elyachi et *al.*, 2009), une approche syntaxique fondée sur les automates à états finis (Es Saady et *al.*, 2010) et les modèles de Markov cachés (Amrouch et *al*., 2010).

6. Conclusion

Dans un contexte général visant la pérennité et la promotion de la langue amazighe, les TICs constituent une piste prometteuse pour lutter contre la fracture numérique dont l'amazighe a souffert. Dans cette perspective, le présent article propose une méthodologie d'informatisation de l'amazighe pour agir de façon ciblée et efficace dans quatre domaines clefs, à savoir la sauvegarde du patrimoine, l'aménagement linguistique, l'éducation et les médias. Le fait de mettre l'accent sur ces domaines permettra de donner aux Marocains les moyens de s'épanouir en amazighe comme en arabe et en français et de contribuer au dévleppoment de la société. Cette méthodologie se base sur les travaux élaborés pour l'informatisation des langues peu dotées et suit un plan de développement en trois étapes partant de la réalisation d'un kit d'outils et de ressources linguistiques minimaux jusqu'à l'élaboration d'applications avancées. En outre, l'article dresse un bilan de réalisation d'outils et de ressources linguistiques amazighes permettant son traitement automatique et son intégration dans le domaine des TICs.

Références

Ait Ouguengay, Y. et *al*. (2012), « Projet GCAM- Vers une gestion informatisée du corpus amazighe à l'IRCAM », in *TICAM 2012*, 26-27 novembre 2012, Maroc.

Ait Ouguengay, Y. et Taalabi, M. (2009), « Elaboration d'un réseau de neurones artificiels pour la reconnaissance optique de la graphie amazighe: Phase d'apprentissage », in *Systèmes intelligents-théories et applications*. Europia productions.

Amrouch M. *et al*. (2010), « Handwritten Amazigh Character Recognition Based On Hidden Markov Models. International Journal on Graphics », *Vision and Image Processing*, Vol. 10, n° 5, p. 11-18.

Andries, P. (2008), *Unicode 5.0 en pratique : Codage des caractères et internationalisation des logiciels et des documents*, France, Dunod.

Ataa Allah, F. (2010), « Etude comparative au niveau de ressources et outils des langues peu, moyennement et bien dotées », Rapport interne, CEISIC, IRCAM.

Ataa Allah, F. (2011), « Construction de corpus étiqueté des documents textuels de la langue amazighe », Rapport interne, CEISIC, IRCAM.

Ataa Allah, F. (2014), «Finite-state transducer for Amazigh verbal morphology», in *Literary and Linguistic Computing*, Oxford University Press. doi: 10.1093/llc/fqu045.

Ataa Allah, F. et Jaa, H. (2009), « Etiquetage morpho-syntaxique: outil d'assistance dédié à la langue amazighe », in *SITACAM 2009*, 12-13 december 2009, Maroc.

Ataa Allah, F. et Boulaknadel, S. (2010a), « Amazigh Search Engine: Tifinaghe Character Based Approach », in *IKE 2010*, 14-16 juillet 2010, USA.

Ataa Allah, F. et Boulaknadel, S. (2010b), « Pseudo-racinisation de la langue amazighe », in *TALN 2010*, 19-23 juillet 2010, Canada.

Ataa Allah, F. et Boulaknadel, S. (2012), « Toward computational processing of less resourced languages: Primarily experiments for Moroccan Amazigh language », in *Text Mining*, (éd) Rijeka, InTech, p. 197-218.

Ataa Allah, F. *et al.* (2013), « Amazigh Language Desktop Converter », in *SITACAM 2013*, 2-4 mai 2013, Maroc.

Berment, V. (2004), *Méthodes pour informatiser des langues et des groupes de langues «peu dotées»*, Thèse de Doctorat, Université Joseph Fourier, France.

Boukhris, F. *et al*. (2008), *La nouvelle grammaire de l'amazighe*, Rabat, IRCAM.

Boulaknadel, S. et Ataa Allah, F. (2010), « Online Amazigh Concordancer », in *ISIVC 2010*, 30 septembre - 2 octobre 2010, Maroc.

Boulaknadel, S. et Ataa Allah, F. (2012), « Building a standard Amazigh corpus », in *IHCI 2011*, 29-31 août 2011, Tcheque.

ElYachi R. *et al*. (2010), « On the Recognition of Tifinaghe Scripts ». *Theoretical and Applied Information Technology*, vol. 20, n° 2, p. 61-66.

El Azrak, N. et Elhamdaoui, A. (2011), « Référentiel de la terminologie amazighe : outil d'aide à l'aménagement linguistique », in *NTIC* 2011, 24-25 février 2011, IRCAM.

Es Saady, Y. *et al*. (2010), « Printed Amazigh Character Recognition by a Syntactic Approach using Finite Automata ». *International Journal on Graphics Vision and Image Processing*, vol. 10, n° 2, p. 1-8.

Iazzi, E.M et Outahajala, M. (2008), « Amazigh Data Base », *in* HLT & NLP Workshop, 31 mai 2008, Maroc.

Laabdelaoui, R. *et al*. (2012), *Manuel de conjugaison amazighe*, Rabat, IRCAM.

Muhirwe, J. (2007), « Towards Human Language Technologies for Under-resourced languages », *Computing and ICT Research*, (éd) Joseph Kizza et al, Kampala.

Nejme, F.Z. *et al*. (2013), « Analyse Automatique de la Morphologie Nominale Amazighe », *in* TALN 2013, 17 - 21 juin 2013, France.

Outahajala, M. *et al*. (2010), « Tagging Amazigh with AnCoraPipe », *in* HLT & NLP Workshop, 17 mai 2010, Malta.

Scannell. K.P. *et al*. (2012), *The Irish Language in the Digital Age / An Ghaeilge sa Ré Dhigiteach*. White Paper Series, Springer-Verlag Publisher.

Asinag-Asinag, *9*, 2014, p. 49-73

'Smallcodes', a Unified Computational Linguistics Toolbox for Minority Languages

Carlo Zoli

Smallcodes S.r.L., Firenze, Italy

L'articolo presenta la toolbox di Smallcodes, uno strumento web indispensabile per le lingue minoritarie e valuta la sua possibile applicazione al Tamazight. Lo strumento è stato sviluppato per permettere anche alle minoranze linguistiche di incrementare la loro presenza nel cyberspace e quindi passare da una realtà solo orale e al mondo scritto di Internet. La toolbox unificata di Smallcodes è composta da diversi moduli integrati tra loro: un dizionario che tenga conto delle esigenze speciali (ad esempio i caratteri) delle lingue di minoranza; un correttore ortografico studiato per la diversità dialettale; una sezione di terminologia che aiuti la pianificazione di neologismi. Crediamo infatti che per permettere alle lingue meno usate di sopravvivere in un mondo ultra-connesso dove la maggior parte degli input sono mediati dal web e dalla lingua scritta, sia necessario dotare queste lingue di un "kit di sopravvivenza" per fornire loro gli stessi strumenti e risorse delle maggiori lingue nazionali.

Introduction

Living in a hyper-connected world means that most of the inputs we receive daily are mediated by the Web and they are thus mainly written and not orally transmitted. Therefore only those inputs with the right features, such as a good visibility, the right patterns and an understandable language are winners in the vast world of the written media. Contents in English language are therefore the most widely spread because they most often meet these parameters. It is clear that, if we want to save and preserve minority languages, we must necessarily let these lesser-used languages have access to the tools and resources of the same technological level as those of "bigger" languages. This can be done only sharing experience, expertise and costs between minorities.

We believe that a computational linguistics toolbox can offer a unique solution for minority languages to increase their presence in the written media, among which the cyberspace is and will be the most pervasive. Basically, a first set of resources that are needed to undertake the path to a complete NLP toolbox are, not necessarily in this order, lexica, morphological analyzers / synthesizers, phonetic

similitude patterns, neology / terminology thesauri, corpora and parsers (Scannel, 2011).

The aim is to create a **single** and **integrated** platform for language technology dedicated to minority languages. The big novelty in the design of this instrument is that we put all the tools together in one single highly interoperable box. This unified and comprehensive toolbox is designed for the creation and management of electronic language resources, to respond to the following needs (here we use the classic tripartition of corpus, status and acquisition planning introduced by Heinz Kloss (1976):

- Corpus planning: such a toolbox is necessary to study minority languages both in their internal variability and from a standardized point of view.
- Status planning: the tools for neology provide a rapid introduction in the world of administration and education whereas the orthographical and auto-completion tools are intended to give an easy means in order to move from the local oral variety to the standard written form.
- Acquisition planning: the toolbox aims at providing language students with a comprehensive tool made for acquiring the language and practicing its use in everyday life.

Designing such a tool for minority languages is in some ways more difficult than making it for an official national language, because only the latter has an ancient and well-established written tradition. And yet this action is even more necessary, because computational linguistics for minority languages is not an accessory "luxury", but it is a necessary (unfortunately not sufficient) condition to survive in a globalized world (Dell'Aquila, Iannàccaro, 2011).

Smallcodes platform has an explicit eco-linguistic intent because it wants to create interest around poorly investigated topics by mainstream universities. In fact, Smallcodes works as a commercial firm when working with industrial, commercial and government partners, but we also work as a non-profit organization when collaborating with non-profit, volunteer, ONG partners or when participating in co-funding of national or international projects.

A first-level toolbox

The bare minimum to ensure any language a scientific and systematic presence in the written world is made of:

- A lexicon.
- A spell-checker tool.
- A terminology module.

The final aim is the maintenance and/or re-integration of language in society and these tools are the necessary means to develop the chain of corpus planning → status planning → acquisition planning. It must be clear that these technologies are just means: the main purpose is in fact the maintenance of the language in social

life. But in the contemporary world the social use passes through the written form, and the written form passes through technology.

Minority languages lack of those fundamental IT tools that allow scientists to study other bigger languages (i.e. “terminology extractors”, or “resumé automatique”, or “question answering”). In fact, when we talk about minority, small, lesser used languages, we have to face not only their (relatively) scarce presence in the cyberspace, but also the quality of this presence. We are talking in terms of sociolinguistic quality, not about the literary or the aesthetic value.

It may be curious that an institution like ours that has devoted its life to preserve linguistic diversity is such a strong defender of standardisation. Is not standardisation an enemy of natural autochthonous languages as much as colonialism or “English glottofagy”? On the contrary, we must be realistic: there are very powerful tools that have been developed for standardised languages which have meant years of development and millions of investments. It would be crazy not to use them; it would be fool to think that Google Translator, or the incredible results of the search engines or of the semantic Web would have been achieved if English had not been… English as a world language [Zoli, 2012 (1)]. If we want to foster our small languages in the real world, and in the cyberspace (the two things will tend to be asymptotically the same) we must be as “dwarfs sitting on the shoulders of giants”, as we say in Italian. And to do this we have to pay a little price: giving the language a common standard written form. This is absolutely not sufficient, but it is terribly necessary and, in most of the contexts where we work, it is not obvious at all.

Would it be sensible to go to Microsoft and ask them to localize Windows in 3 different Sardinian languages? Would it be conceivable to go to Shēnzhèn at Apple Developers meeting and ask for 5 different Romantsch forms of iOS, or of Siri? We must make all the possible profit from these global instruments as Google or Siri and sit on the shoulders of these giants.

In this respect, having a look at Gartner’s hype cycle as of July 2012 (*Figure 1*) is of great importance. Many expectations concern technology languages, but can we think about information extraction, or integration with calendars or smart phones, if people do not agree on how to write “Thursday”.

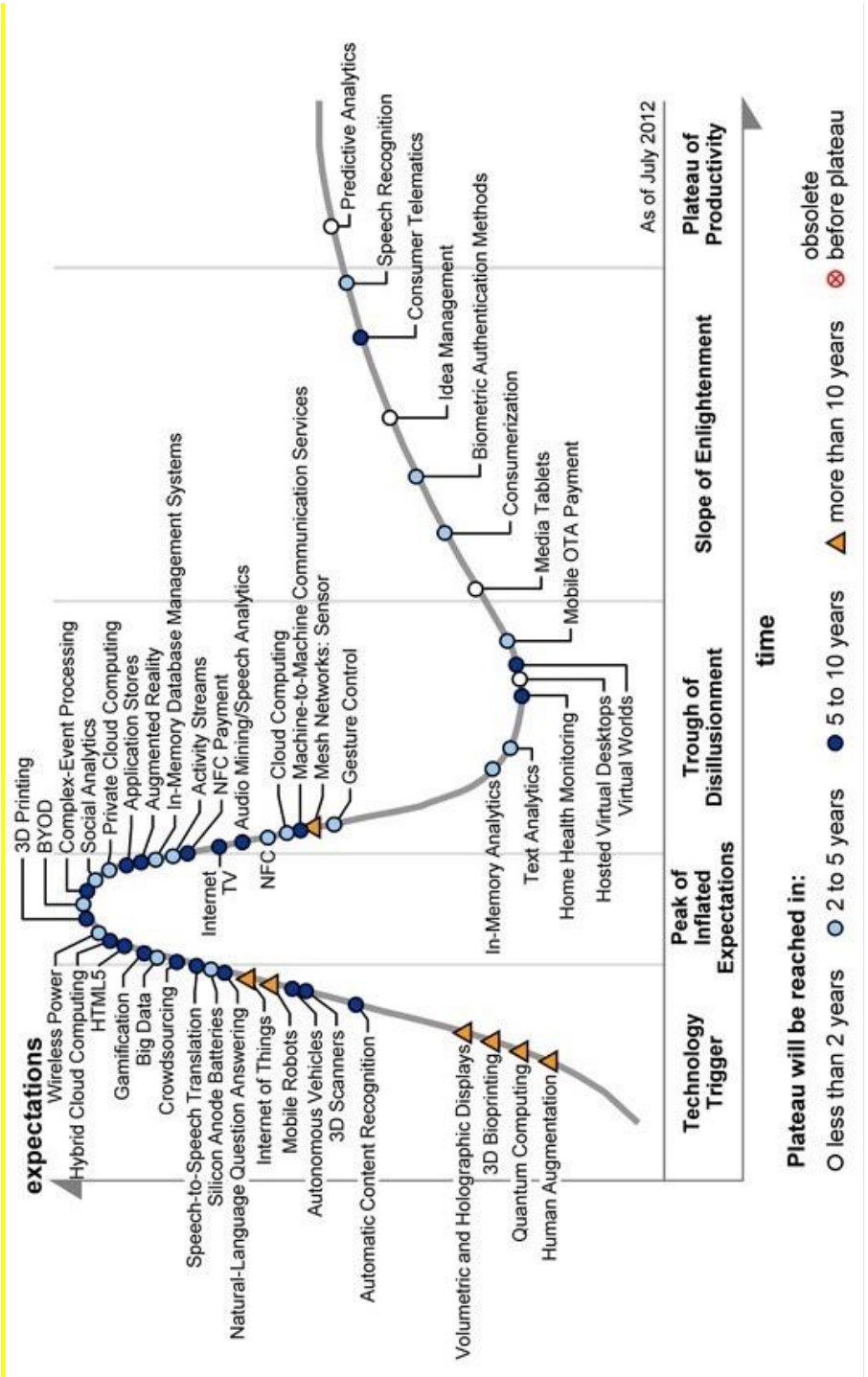


Figure 1 : Gartner's hype cycle as of July 2012

In order to meet technological expectations is therefore necessary to promote the written use of the language (which is necessarily electronic and not handwritten). The writing is - just in terms of status planning - often in the domains of administration, bureaucracy and schools (Dell'Aquila, Iannàccaro, 2011:98ff). These are the more permeable areas to language policy, while those of literary creativity are often oriented towards localisms and are more reluctant to accept a standardized script.

The written use must be then promoted and facilitated. In this sense, the advantage of minority languages is that very often the official institute for the defence of the language is unique and known by many: the Institut Royal de la Culture Amazighe for Tamazight or the Istitut Cultural Ladin for Ladin language are authority whose prestige is recognized by most of the speakers of the target minority language.

The involved fields

Not every research field of computational linguistics can be involved at the beginning of the process. At least in the first stage, it is necessary to focus on fewer and simpler areas of interest, having clear in mind that, especially for normal users, for school pupils and teachers, for a non-specialist audience *a fairly-good 'something' is much better than a perfect 'nothing'* (Scannell, 2011). Preliminarily, it is necessary to have a unifying - better than "unified" writing system (field: Writing). Then, the following step is represented by the creation of a common-use dictionary and (if this is possible with budget and workforce available) a dialectal dictionary of local varieties, plus the retrieval of studies and corpora on terminology, neologism and modernization of the lexicon (field: Dictionaries). It is then very useful to have spell-checking instruments such as online spell-checkers (available online and for Microsoft Word and/or Open Office) and automatic correction systems in all these cases (field: Writing aid). A further effort is the creation of corpora and archives of ancient texts, the production of e-books, audio books and didactic material online with downloadable and printable files off-line (fields: Digital libraries / School teaching). An optional subsequent action is the creation of a Web-TV and of free-press magazines and newspaper which is mediated and generated by the Web (field: Digital instruments of mass communication).

The chart below (*Figure2*) shows the fields of interest to be exploited for minority languages according to the urgency of the action to be taken (ranging from green: very urgent – yellow: possible – red: optional).

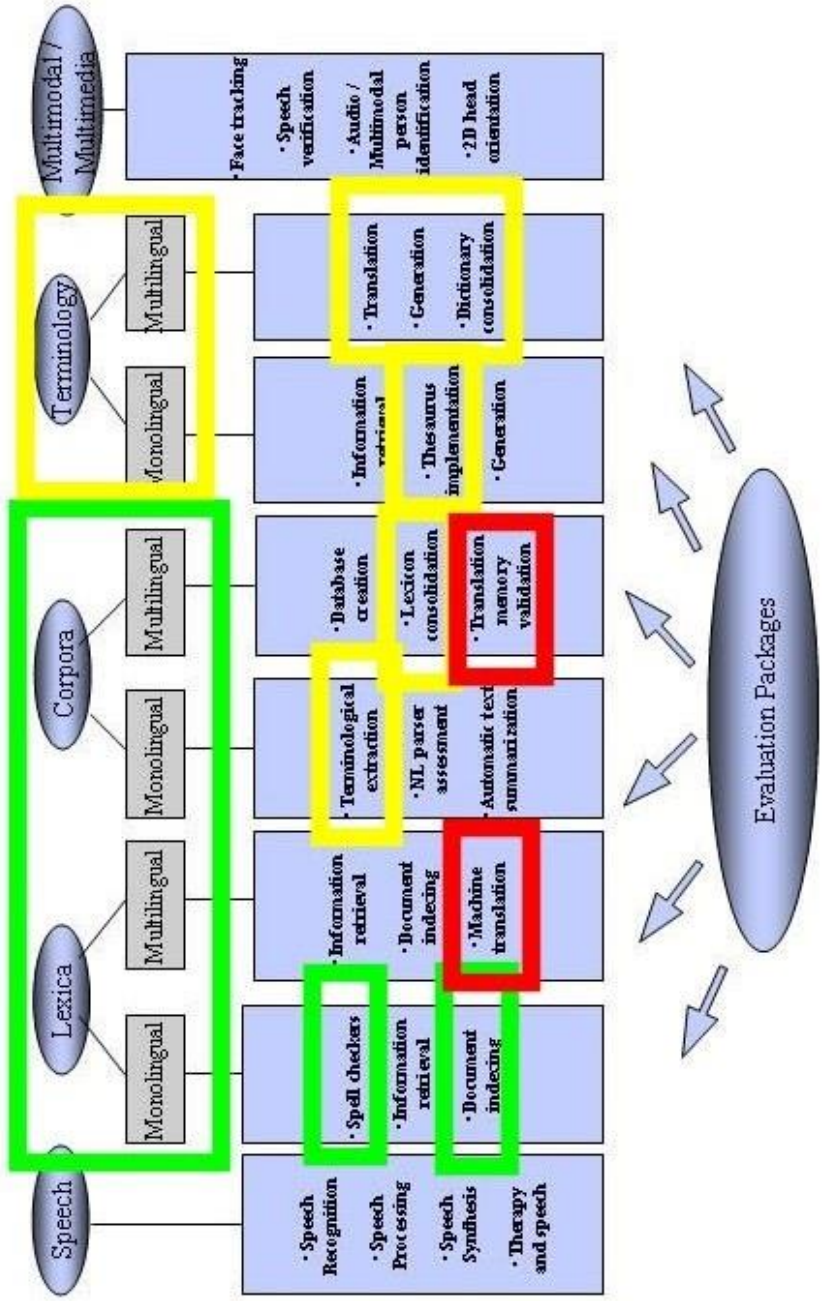


Figure 2: Computational linguistics packages for minority languages (Scannel, 2008)

After having collected enough lexical material (**preliminary step**), it is then possible to plan the dictionary (**first step**). In fact, lexical lists of various kinds are the necessary condition in order to set up the dictionary. They can be wordlists of local or global language (i.e. conforming to local varieties of the language or to the standardized spelling); they can also be imported form informal databases and being the result of an OCR or parsing of ancient dictionaries.

The figure below (*Figure 3*) shows an example of 'standardizing' dictionary with registration of local varieties. Here is the extreme case of the entry otóbro ('October') which has around 150 different phonetic realizations ascribable to three consonantal macro-phenomena (1. maintenance of etymological t; 2. palatalization of t > c. 3. loss of b). As it can be seen, the standard forms have been chosen among those forms which are more "etymologically regular" (Lurà et al., 2009). Then (*Figure 4*), we have the same entry in a human-readable form (actually an XML + CSS wich can be easily imported in a professional publishing tool as Adobe Indesign, see *Figure 6*); *Figure 5* shows the XML of fig 4 in the classic machine-readable form.

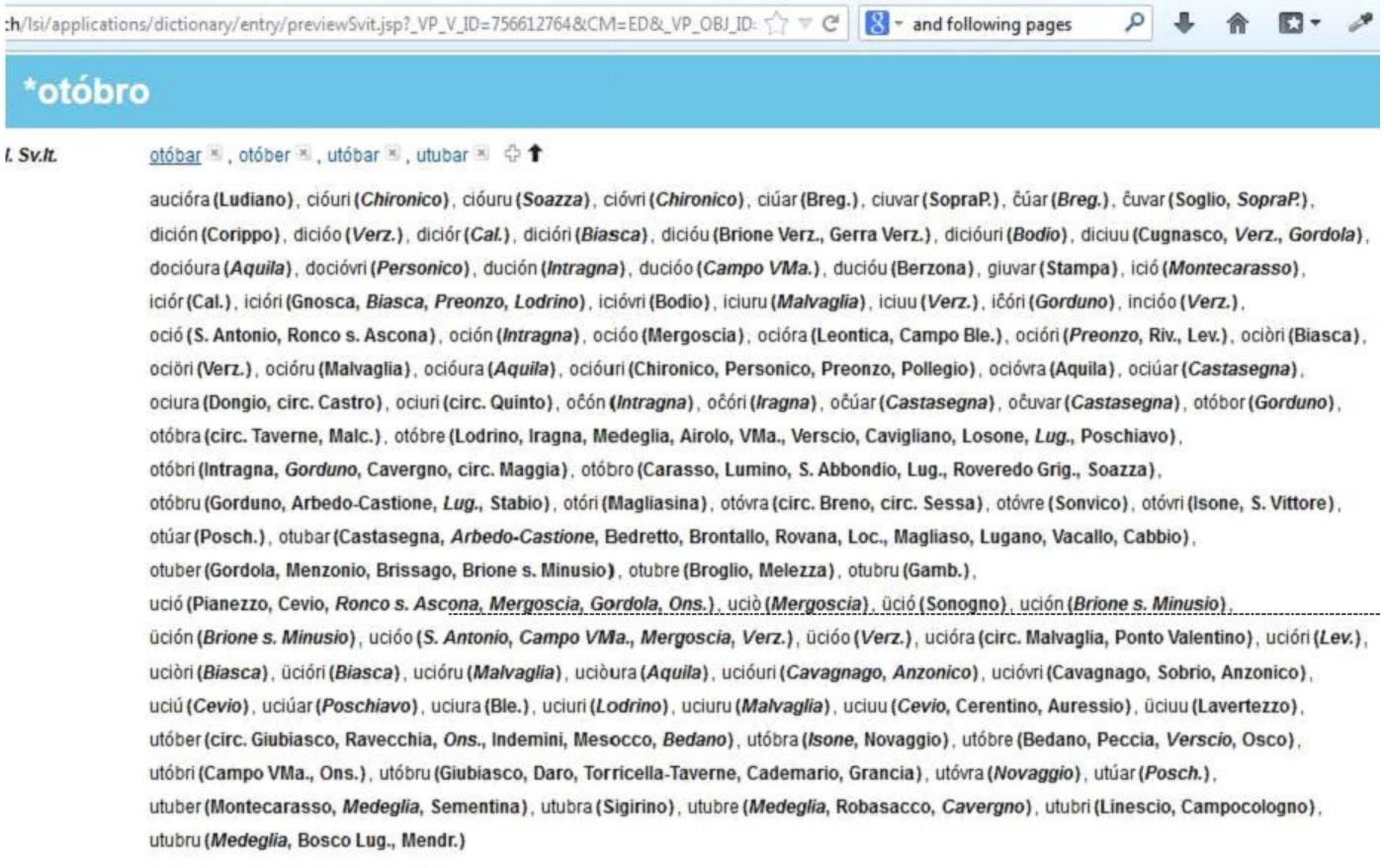


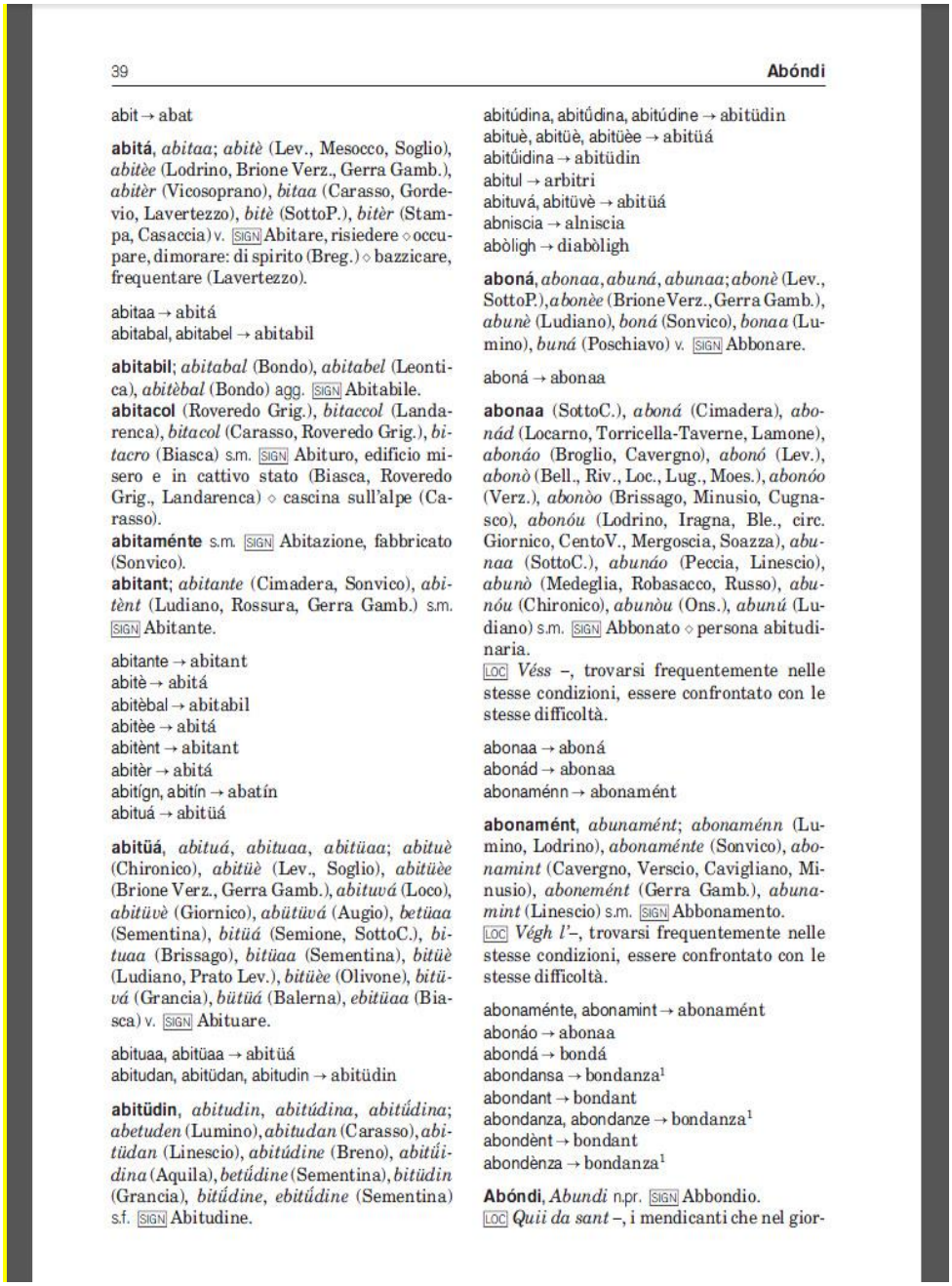
Figure 3: An example of a standardizing dictionary with registration of local varieties

Figure 4: The same entry in a human-readable form[1]

[1]Please note that the current tendency in normalization is to suggest a single graphic form but to allow free choices in local meanings and lexical types. The image shows the lexical type *otóbro* ('October') which in some places means 'autumn, fall'. Symmetrically, for the concept of "October", we could have many other lexical types, such as 'Month of St. Martin' or 'Month of chestnuts'.

```
<?xml-stylesheet type="text/css" href="../../css/xml/dictionaryFrontendXml.css"?><LEMMI xmlns:html="http://www.w3.org/1999/xhtml">
  <LEMMA ID="71671" IS_ALTERNATIVE="false">
    <DIZIONARIO_TITOLO>
      <FORMA_LE ENTRY_TYPOLOGY="NOT_ATTESTED" IS_INVERSE="false">otóbro</FORMA_LE>
      <LINGUE_LE>(dial. Sv.It.)</LINGUE_LE>
    </DIZIONARIO_TITOLO>
    <DIZIONARIO_CORPO FO="false">
      <FORMA_FOR_SEARCHING_LE>otobro</FORMA_FOR_SEARCHING_LE>
      <CAPOLEMMA>
        <CAPOLEMMA_LE_LABEL>capo-lemma:</CAPOLEMMA_LE_LABEL>
        <CAPOLEMMA_LE ENTRY_TYPOLOGY="NOT_ATTESTED">otóbro</CAPOLEMMA_LE>
      </CAPOLEMMA>
      <TUTTE_LE_FORME HIDE="false">
        <TUTTE_LE_FORME_LABEL>tutte le forme di</TUTTE_LE_FORME_LABEL>
        <TUTTE_LE_FORME_LANG>dial. Sv.It.:</TUTTE_LE_FORME_LANG>
        <TUTTE_LE_FORME_DESCR HIDE="false" ULTIMA_FORMA_VISIBILE="false">otóbar</TUTTE_LE_FORME_DESCR>
        <TUTTE_LE_FORME_DESCR HIDE="false" ULTIMA_FORMA_VISIBILE="false">otóber</TUTTE_LE_FORME_DESCR>
        <TUTTE_LE_FORME_DESCR HIDE="false" ULTIMA_FORMA_VISIBILE="false">utóbar</TUTTE_LE_FORME_DESCR>
        <TUTTE_LE_FORME_DESCR HIDE="false" ULTIMA_FORMA_VISIBILE="true">utubar</TUTTE_LE_FORME_DESCR>
      </TUTTE_LE_FORME>
      <VARIANTI_LOCALI>
        <VARIANTI_LOCALI_LABEL>varianti locali:</VARIANTI_LOCALI_LABEL>
        <VARIANTI_LOCALI_DESCR>
          <VARIANTE_LOCALE LOC_VARS_GEO_ORDER="false">
            <FORMA_VL>aucióra</FORMA_VL>
            <LINGUE_VL>(Ludiano)</LINGUE_VL>
          </VARIANTE_LOCALE>
          <VARIANTE_LOCALE LOC_VARS_GEO_ORDER="false">
            <FORMA_VL>cióuri</FORMA_VL>
            <LINGUE_VL>(Chironico)</LINGUE_VL>
          </VARIANTE_LOCALE>
          <VARIANTE_LOCALE LOC_VARS_GEO_ORDER="false">
            <FORMA_VL>cióuru</FORMA_VL>
            <LINGUE_VL>(Soazza)</LINGUE_VL>
          </VARIANTE_LOCALE>
          <VARIANTE_LOCALE LOC_VARS_GEO_ORDER="false">
            <FORMA_VL>cióvri</FORMA_VL>
            <LINGUE_VL>(Chironico)</LINGUE_VL>
          </VARIANTE_LOCALE>
          <VARIANTE_LOCALE LOC_VARS_GEO_ORDER="false">
            <FORMA_VL>ciúar</FORMA_VL>
            <LINGUE_VL>(Breg.)</LINGUE_VL>
          </VARIANTE_LOCALE>
          <VARIANTE_LOCALE LOC_VARS_GEO_ORDER="false">
            <FORMA_VL>ciuvar</FORMA_VL>
            <LINGUE_VL>(SopraP.)</LINGUE_VL>
```


Figure 5: Machine-readable output of the same entry in a LMF (Francopoulo et al., 2006), compliant XML-schema



39 **Abóndi**

abit → abat

abitá, *abitaa*; *abitè* (Lev., Mesocco, Soglio), *abitèe* (Lodrino, Brione Verz., Gerra Gamb.), *abitèr* (Vicosoprano), *bitaa* (Carasso, Gordevio, Lavertezzo), *bitè* (SottoP.), *bitèr* (Stampa, Casaccia) v. SIGN Abitare, risiedere ◇ occupare, dimorare: di spirito (Breg.) ◇ bazzicare, frequentare (Lavertezzo).

abitaa → abitá
abitabal, abitabel → abitabil

abitabil; *abitabal* (Bondo), *abitabel* (Leontica), *abitèbal* (Bondo) agg. SIGN Abitabile.
abitacol (Roveredo Grig.), *bitaccol* (Landarenca), *bitacol* (Carasso, Roveredo Grig.), *bitacro* (Biasca) s.m. SIGN Abituro, edificio misero e in cattivo stato (Biasca, Roveredo Grig., Landarenca) ◇ cascina sull'alpe (Carasso).
abitaménte s.m. SIGN Abitazione, fabbricato (Sonvico).
abitant; *abitante* (Cimadera, Sonvico), *abitènt* (Ludiano, Rossura, Gerra Gamb.) s.m. SIGN Abitante.

abitante → abitant
abitè → abitá
abitèbal → abitabil
abitèe → abitá
abitènt → abitant
abitèr → abitá
abitígn, abitín → abatín
abituá → abitüá

abitüá, *abituá*, *abituaa*, *abitüaa*; *abituè* (Chironico), *abitüè* (Lev., Soglio), *abitüèe* (Brione Verz., Gerra Gamb.), *abituvá* (Loco), *abitüvè* (Giornico), *abütüvá* (Augio), *betüaa* (Sementina), *bitüá* (Semione, SottoC.), *bituaa* (Brissago), *bitüaa* (Sementina), *bitüè* (Ludiano, Prato Lev.), *bitüèe* (Olivone), *bitüvá* (Grancia), *bütüá* (Balerna), *ebitüaa* (Biasca) v. SIGN Abituare.

abituaa, abitüaa → abitüá
abitudan, abitüdan, abitudin → abitüdin

abitüdin, *abitudin*, *abitúdina*, *abitűdina*; *abetuden* (Lumino), *abitudan* (Carasso), *abitüdan* (Linescio), *abitúdine* (Breno), *abitűidina* (Aquila), *betűdine* (Sementina), *bitüdin* (Grancia), *bitűdine*, *ebitűdine* (Sementina) s.f. SIGN Abitudine.

abitúdina, abitűdina, abitúdine → abitüdin
abituè, abitüè, abitüèe → abitüá
abitűidina → abitüdin
abitul → arbitri
abituvá, abitüvè → abitüá
abniscia → alniscia
abòligh → diabòligh

aboná, *abonaa*, *abuná*, *abunaa*; *abonè* (Lev., SottoP.), *abonèe* (Brione Verz., Gerra Gamb.), *abunè* (Ludiano), *boná* (Sonvico), *bonaa* (Lumino), *buná* (Poschiavo) v. SIGN Abbonare.

aboná → abonaa

abonaa (SottoC.), *aboná* (Cimadera), *abonád* (Locarno, Torricella-Taverne, Lamone), *abonáo* (Broglio, Cavergno), *abonó* (Lev.), *abonò* (Bell., Riv., Loc., Lug., Moes.), *abonóo* (Verz.), *abonòo* (Brissago, Minusio, Cugnasco), *abonóu* (Lodrino, Iragna, Ble., circ. Giornico, CentoV., Mergoscia, Soazza), *abunaa* (SottoC.), *abunáo* (Peccia, Linescio), *abunò* (Medeglia, Robasacco, Russo), *abunóu* (Chironico), *abunòu* (Ons.), *abunú* (Ludiano) s.m. SIGN Abbonato ◇ persona abitudinaria.
LOC *Véss* –, trovarsi frequentemente nelle stesse condizioni, essere confrontato con le stesse difficoltà.

abonaa → aboná
abonád → abonaa
abonaménn → abonamént

abonamént, *abunamént*; *abonaménn* (Lumino, Lodrino), *abonaménte* (Sonvico), *abonamint* (Cavergno, Verscio, Cavigliano, Minusio), *abonemént* (Gerra Gamb.), *abunamint* (Linescio) s.m. SIGN Abbonamento.
LOC *Végh l'*–, trovarsi frequentemente nelle stesse condizioni, essere confrontato con le stesse difficoltà.

abonaménte, abonamint → abonamént
abonáo → abonaa
abondá → bondá
abondansa → bondanza[1]
abondant → bondant
abondanza, abondanze → bondanza[1]
abondènt → bondant
abondènza → bondanza[1]

Abóndi, *Abundi* n.pr. SIGN Abbondio.
LOC *Quii da sant* –, i mendicanti che nel gior-

Figure 6: XML above imported automatically into Adobe InDesign for automatic layout for printing.

The **second step** should be the integration of a morphological analyzer-synthesizer within the dictionary, in order to develop a fully integrated spell-checker for the minority language. The majority of spell-checking systems (e.g. HunSpell which is the base of LibreOffice, Firefox, Chrome, etc. proofing tools) are fed with wordlists which are not integrated and often not even exported from a coherent dictionary authoring system (Németh 2011); the same can be said for morphological engines or corpus analysis software, such as NOOJ (Ben Hamadou, Mesfar, Silberztein, 2010): they may provide powerful tools, but they are never integrated with a dictionary authoring and publishing system, and their use is normally confined to NLP specialists, and often well beyond the reach of traditional linguists not to say general public, school teachers or public administration staff. In fact, having an integrated system means that every change is reported automatically in both modules of the system and that the spell-checker is always up to date, and so is authoring, Web publication, Smartphone app generation, and even traditional paper publishing are all steps of a highly integrated procedure. This is especially useful in treating minority or lesser-used language, where the fieldwork is always active and new additions, changes, creation of neology and terminology, and even spell reforms are frequent events. As modern spell-checkers, our module works with a "best-guess" pattern of the rule, based on statistic algorithms, on *Levenshtein distance* (Levenshtein, 1966) and on *double metaphone* (Philips, 1990).

In addition, it includes dialectal-driven error patterns, which are fundamental for minority languages. In fact, every correction system sets up its guesses upon similarities of words. Our system adds to this method the awareness that, for semi- or recently standardized languages where the overwhelming majority of writers are *de facto* illiterate in their language, most errors can be caused by the knowledge of a word in one particular language variety that is not the standard form: in minority languages people do not only misspell: they simply can't write, even if they can perfectly speak (and write in the dominant language). The two word forms (standard and non-standard) may differ a lot sometimes: the non-standard word can be, for example, more similar to a word with a complete different meaning than to its standard equivalent; or it can also be so graphically far from the standard form that the system is not able to find the equivalence using the statistic algorithm or the standard pattern matching. The system must then know that there can be odd correspondences. We can offer a typical example from Sardinian language (the first language for which we developed the spell-checker): the word *berbeghe* (sheep) is pronounced /brebei/ in South Sardinia. If we analyze the differences among the two words, we can understand that a simple system would not be able to guess the standard form (*berbeghe*) starting from the non-standard one (*brebei*) (Corongiu, 2013). Conversely, our dialect-oriented spell-checker knows these odd correspondences and the rules that allow to guess them. Our system uses therefore two guess pattern, shown in the table below (*Figure 7*): the simple one detects "*soundslike* typical mistakes"; the advanced one detects "linguistic-background driven mistakes". See Figure 8 and Figure 9 for MS Word and web interface of the "dialectal" spellchecker (Zoli, 2008).

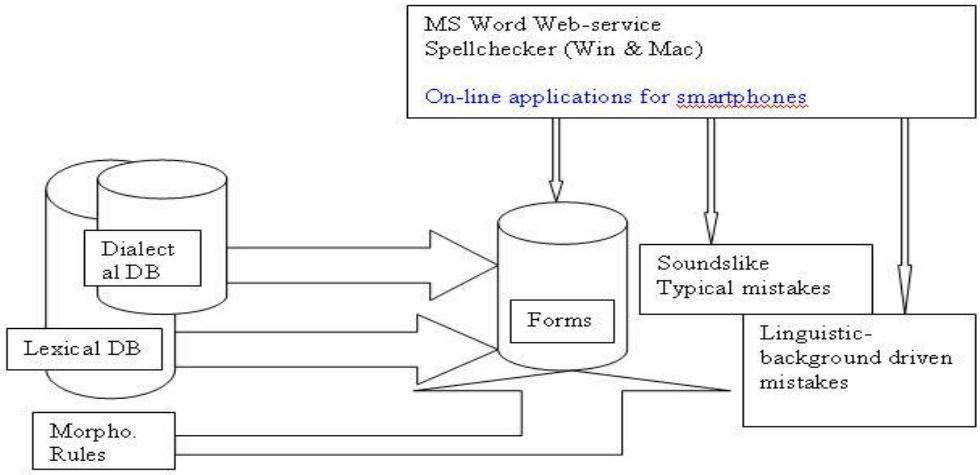


Figure 7: Functioning of an advanced spell-checking system

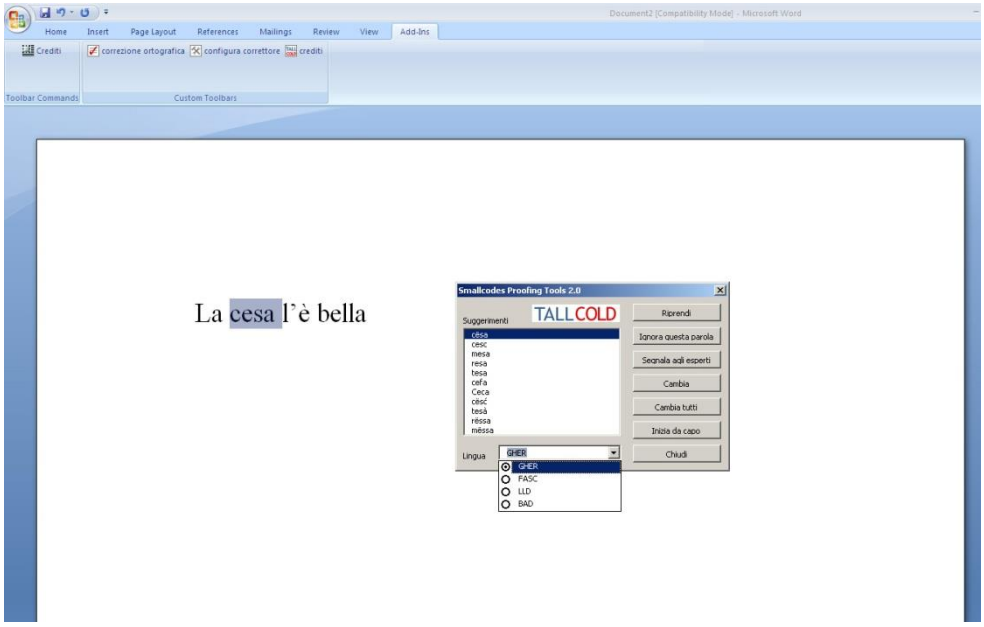


Figure 8: Spell-checking of standard Ladin language with correction based on the typical errors caused by the three main dialectal backgrounds (corresponding to the three major oral dialects spoken in the respective alpine valleys: Gherdëina, Badiot, Fascian

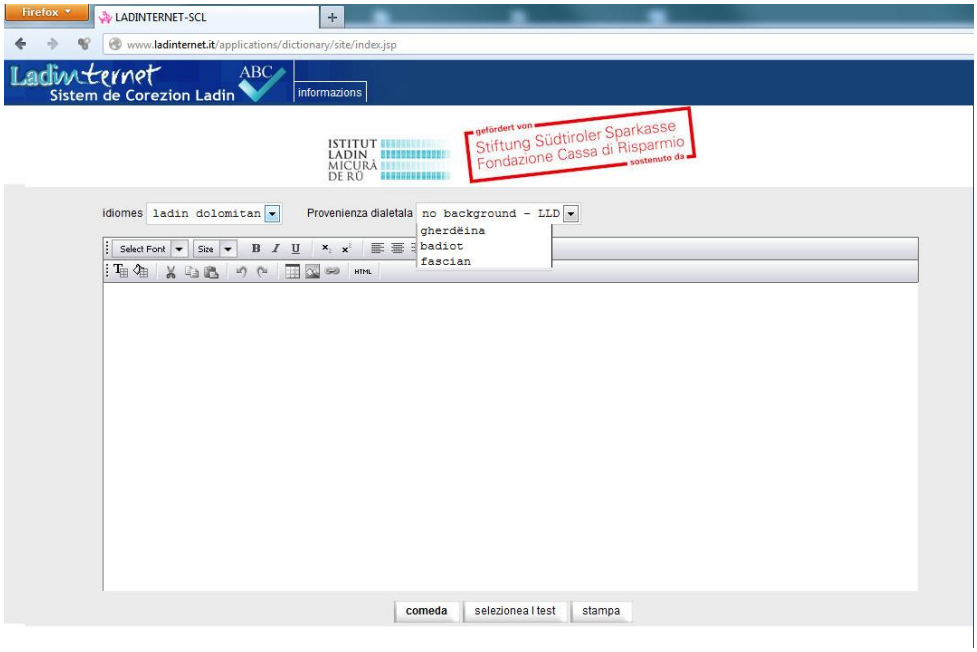


Figure 9: Same system as above accessing exactly the same data via the same SOAP Web-service but for use in a text area within a Web browser.

The **third step** is the terminology module. The creation of the terminology is a fundamental procedure if we want the language to be employed, for example, in school teaching (see for example Figure 10, which shows a collaborative webTool for neology, used by the authors of schoolbooks in Ladin Dolomitan), and administrative / official translation (see fig 11 & 12 for a tool of computer-aided technical translation for Sardinian languages, used by various public bodies). Languages which do not have a written tradition normally lack of technical lexicon. These new words need therefore to be created and the method for their creation already exists: the sources are the other international languages that have made this procedure before and the other minority languages that have already solved these issues. Another possibility is to re-use old words whose original meaning is losing importance in today's life and make these words express new meanings. A typical example is the vocabulary used for cars nowadays in Italian: this is nothing more than the recovered lexicon for horse carriages; similarly, the lexicon of Air Navigation is directly taken from Maritime Navigation vocabulary. English typically uses this strategy for neologisms, exploiting metaphors and meaning extensions of pre-existing words. Romance languages, on the other hand, favour the use of loan words, drawing inspiration from present or past prestigious languages.

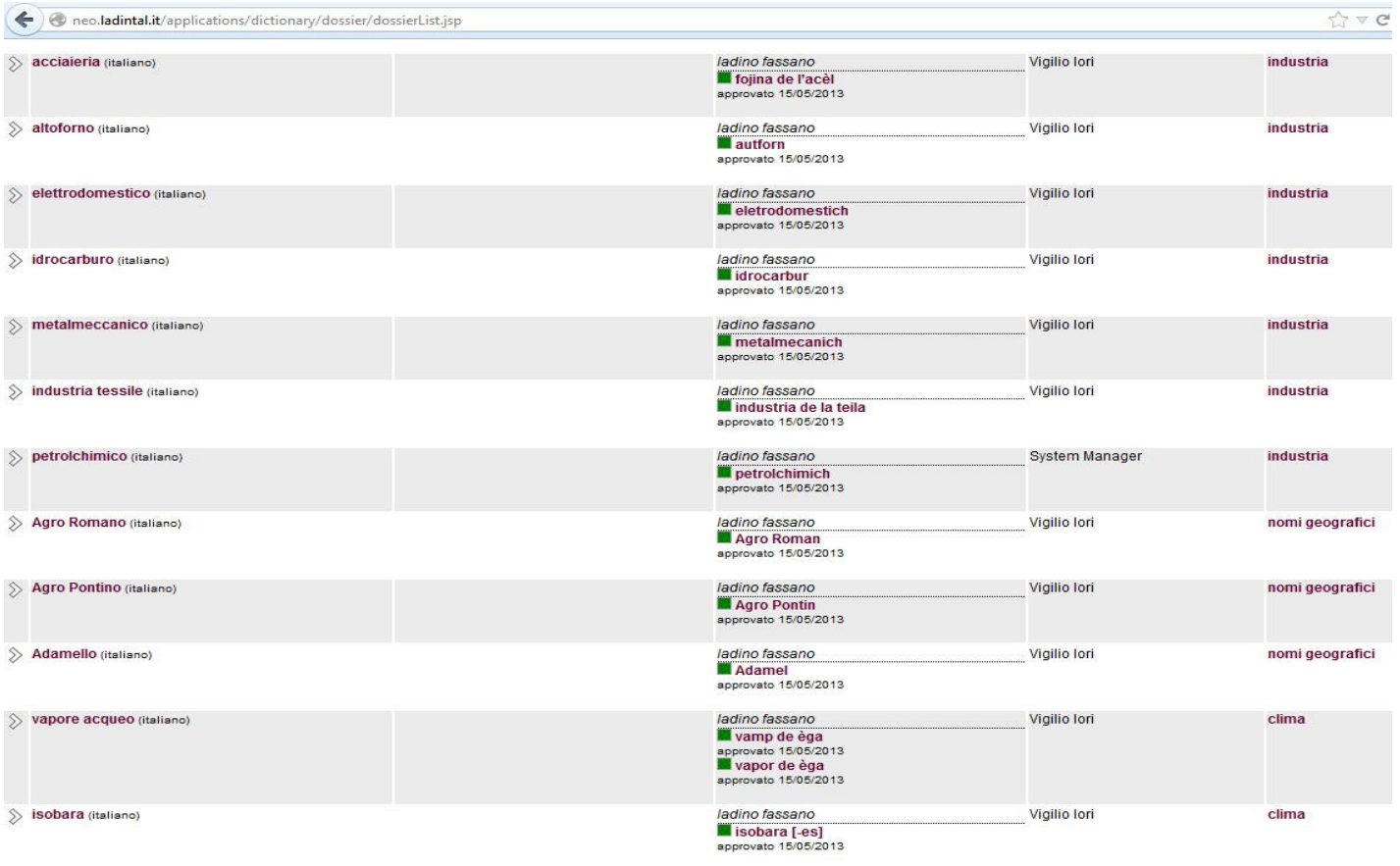


Figure 10: An example of the work flow (with various status of approval) for the creation and consolidation of terminology in Ladin language: please note that the system is fully integrated with the dictionary module so that specific word-lists can be included or excluded from the general dictionary, exported for the Web or via Web-service for use within other applications

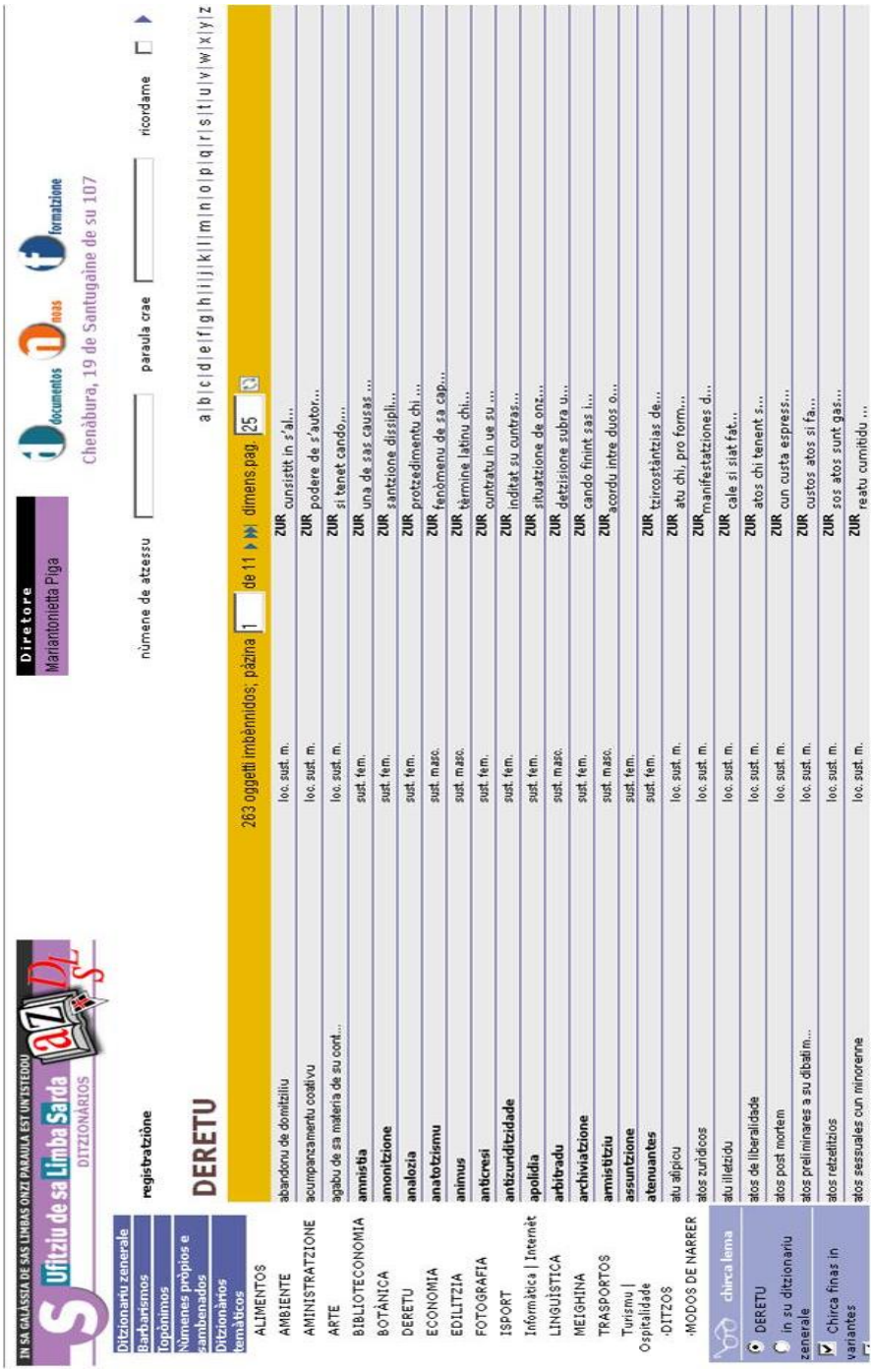


Figure 11: Web page output of terminology module

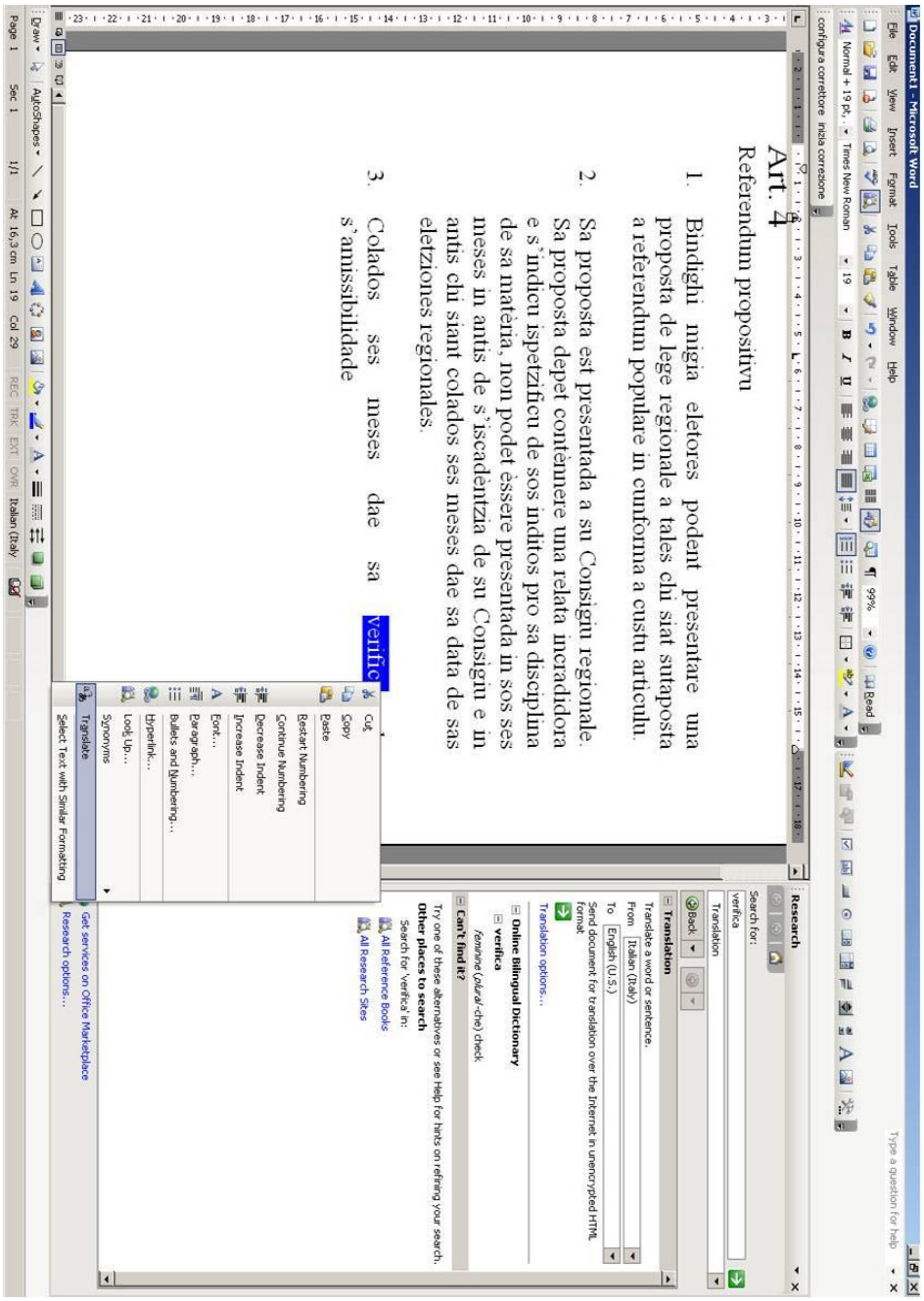


Figure 12: Web-service output for word-to-word terminology translation

The **fourth** (optional) **step** is the creation of a reference thesaurus, namely the collection of written material of any literary kind and historical period which offers a precious help in the consolidation of lexicon. A first-level thesaurus does not need to be exaggeratedly ambitious: actions such as pos tagging, machine-translation or stylistic analysis can be left out of the first phase of our project, not because they are not interesting or important but because they do not belong to the very first set of tools that every language needs to start its digital preservation (Soria, Zoli, 2012). What is very useful, at the beginning, is the possibility of having phrase quotations for literary dictionaries, the lemmatization and the *aligning of old orthography with new orthography*.

The material collected and organized in the corpus allows to compare old and new texts and wordlists and represents an authoritative support to be consulted at any time. The thesaurus, thanks to frequency parameters, can offer guarantees on the effective use of a word or on its register / linguistic style.

The choice of literary texts is done in order to let the speakers' community recognize them as "properly written" and trustable. Old literary texts often have, in fact, a special prestige and are regarded as models in the lexical and semantic field (Videsott, 2011). But very often, if not always, literary texts in minority languages are written in different, incoherent spellings. We must preserve the original script and at the same time re-publish the text in modern / standardized scripts as far as it is possible. The automatic statistical alignment of the two version of the same text (old and modern) allows the users and the researches to quote literature both "as it was written" and "as it would be written today".

Our Ladin thesaurus (*Corpus dl Laden leterar* / *Wörterbuch des literarischen Ladinisch / Corpus letterario del ladino*)[2] is a electronic corpus of literary works written in Ladin language. It currently stores more than 1,200 texts from Ladin valleys (Val Badia, Val Gardena, Val di Fassa, Fodom, Ampezzo): the material has been completely scanned and digitalized and it consists of an archive of more than 250,000 different words.

Such a thesaurus represents the last step of the complete system and its strong connection with the dictionary module contributes to show the importance of an integrated system for the study and filing of lexical material.

Some possible objections

In most contexts where a minority language is struggling to be recognized and protected, standardization is feared by many people. These people, especially when they master the lesser-used language, are afraid that a major standard form might hide away their native varieties, which have a strong identity value for them. At the same time, the speakers who fear standardization, also reject the use of tools such as electronic instruments for spell-checking (according to the belief that everyone writes in his or her own way, or in a way which is totally respectful of local pronunciations cfr. Vitali 2008). This attitude contribute to relegate minority

[2] http://corpuslad.ladintal.it/

languages such as Tamazight to the status of dialects and prevent them to evolve and flourish.

Instead, it must be clear that standardized spelling only makes sense for a written language. If there were, for example, a talk show in one minority language, the titles and explanatory signs would be in standard, but the presenter and the guests would talk in their own dialects (as it happens in German Switzerland or in Norway) [Zoli, 2012 (2)]. The spell-checker we developed is aimed at appointing languages such as Tamazight a written authority in which every rule is fixed and scientifically established. Only in this way, we believe, small local languages can be protected by the unified bigger standard form (which is, again, only a standard script that tries to enables everyone to read in his or her own dialect, as long as a small effort is made to find regularities and correspondences between every vernacular variety and the standard form).

Moreover, some speakers might challenge the effectiveness of the terminological research we support with our third module of the integrated system. In fact, some people do not accept the creation of neologisms because they are alien to the traditional language these speakers learned as children ("*my grandma would have never said that!*"). As a matter of fact, no language, at an early stage, has the words to express novelties or brand new concepts, but the school has made us believe that certain languages are rich for some sort of divine predestination (Pellegrini, 1977). If we take any Italian or French vocabulary, we can see that about 4000-5000 words are derived directly from Latin: these words are the most frequent and they concern concepts or things which are the backbone of the language. Another 20,000 are also derived from Latin, but these other words were created later, invented by the humanist scholars and writers. When a new concept was needed, scholars used to draw it form the inexhaustible mines of Latin or Greek and superficially make the word fit the graphics system and the phonetics of the target language. This explains why French and Italian basic words directly derived from Latin only remotely resemble each other (occhio / œil, bocca / bouche, casa / chez), while the scientific terminology is virtually identical (oculare / oculaire, orale / oral, domestico / domestique). The former have come straightforwardly from Latin and their form is the result of phonetic modification. The latter have been reinserted later and belong to scientific or academic (i.e. terminological) vocabulary. The creation of new terminology is therefore at the basis of the rehabilitation of a language and it is a necessary step for this language to acquire prestige and be adopted in the public sphere (e.g. school or public administration).

A quick comparison with possibly similar tools

It has to be said that similar ideas have been around for a while: efforts like BLARK (Krauwer, 2003) and LCTL at the Linguistic Data Consortium rely on somewhat similar ideas [3].

The main difference is that this projects aims to be industry-standard: the idea, as it is expressed in the manifesto below, is to give long digital life not only to data, but

[3] http:// projects.ldc.upenn.edu/LCTL/index.html

also to applications, source-code, etc. A limit of software tools that come from the academic and pure-research world is that they often cannot be maintained by "big" teams of professional software developers, but often are either quickly abandoned (not by the users, by the developers when the research project, and consequently the funding, is over) or suffer an inevitable technical obsolescence (the case of E-Meld is paradigmatic).

We could say that the Smallcodes project, stemming from the private industry sector and approaching the research world (rather than vice versa) has a, so to say, *different business model*.

The business model is not that the language experts or researchers adopt the system as users, basically using it "at their own risk" or contributing to the development, in a classical open-source fashion.

On the contrary, the Smallcodes business model is that the software is centrally developed, and partnerships and funding opportunities are established every time a new language group enters the "community". Every new language expert group adds new expertise, new funding, requests new features, but development is pursued in an industrial fashion, with attention to the latest web technologies, with highly resourced staff in an a "web 2.0 commerciale way"; then, the business itself is basically non-profit , but all the same this is different from software development done inside the linguistic academic world, which cannot have the structure and the attitude of a commercial software house.

Finally it is more common to find a commitment for sharing language resources (see for example OLAC [4], DoBeS[5]), whereas Smallcodes focuses more on the sharing of *software tools*.

A possible employment of the toolbox for Tamazight

Our aim is to have one integrated tool which will be multi-accessible and will give multiple simultaneous outputs. These are as follows:

A) "human readable data": a web-app which will provide (not necessarily all, and not necessarily at the same time):

- Online authoring of dictionary of standard language (with synonyms, antonyms, WordNet-like synset relationships (Fellbaum 1998).
- Online authoring of dictionary of dialectal variation
- Online authoring (with collaborative discussion and workflow) of neology/terminology
- Web publishing for public consultation of dictionary
- Web publishing of conjugation / declination tables (paradigms / schemas)

[4] http://www.language-archives.org/ 21/07/2013

[5] http://dobes.mpi.nl/ 21/07/2013

- Integrated output of XML files for paper publishing (Adobe InDesign format)
- Integrated output of e-books (ePub format)
- Integrated output of XML files for Android / iOS dictionary apps.

B) "machine-readable data": a Web-service which will provide (not necessarily all, and not necessarily at the same time):

- spell-checking
- dictionary look-up
- thesaurus look-up
- glossary look-up → encyclopaedic information
- terminological word-to-word translation
- morphological analysis and synthesis.

The Web-service will provide data to many different applications: for use in a browser, or integrated in a word-processor (via XML SOAP web-service) or, again, integrated in e-Books for dictionary / terminology lookup.

Every language has its own peculiarities in terms of phonology and morphology. A comprehensive tool must take account of a very large number of possible differences among languages and anticipate the changes that each language will require to the system.

The case of Tamazight is more complicated: we must be able to search for an entry using the Tamazight script but also with the corresponding Latin characters. We must therefore create the correspondences between graphemes and insert them into the system. We must also take into account all possible variations in transliteration and design a list of interchangeable graphemes. All this will be accomplished with multiple Lucene indexes.[6]

We also know that the Tamazight, as every language of recent standardisation can have oscillation in writing, and event different realisations of the same phoneme: □ / □; □ / □; □ / □; □ / □; □ / □ (Boukous, 2009). The dictionary module and the attached spell-checker must take account of all these possibilities.

In addition to these mutation processes, Tamazight language also possesses many assimilation processes, such as the propagation of emphasis or the assimilation of voiced and unvoiced consonants. These aspects concern instead the spell-checker, which must then conceive special rules for words formation which reckon with these phenomena. Only a strong language-aware spellchecker and metaphone algorithm can achieve good results: in the situation of non-latin, recently-standardized and highly diatopically variable languages standard spell-checking simply does not work.

[6] http://lucene.apache.org/ 29/05/2013.

Again, with regard to the spell-checking module, we must have a look at morphology: there are typical functions of Tamazight language which do not concern, for example, Romance languages, such as the discontinuous affixation. Therefore, we have to formulate a set of rules in order to automate the process of spelling correction. In this particular case, we must take account, for example, of the incredible variety of patterns in plural formation. Moreover, we must add the categories of grammatical cases and consider the morphological changes that words undergo in this inflection (in addition to gender and number inflection). Yet again, there are several morphological changes in verbal inflection, such as personal endings, aspect, derivative morphemes (causative, reciprocal and passive), noun agreement (for the participle form) (Boukhris, 2008).

These examples are useful to show that an effective comprehensive system must be able to adapt to the needs of every language. Tamazight has a complex grammar, even if, when compared to other distant languages (such as Mexican languages, which whom we work with), it has some sort of similarities with European languages. We are struggling in order to reach the best results in the consideration of the largest amount of lexical and grammatical possibilities. Every new language we introduce in our system is an important piece of the puzzle that allow us to test the capabilities of our system, add new concepts, discard old beliefs. This expectation, we think, is our way to put into practice the principle of cooperation among minority languages of the world.

Our manifesto (Zoli, 2008)

As we have seen, language technology offers significant opportunities for minority languages and can be a major force in addressing and alleviating some of the difficulties they face. Speech and language technologies are in fact a powerful means to bring together speakers' communities, to have a major impact on language learning support, to promote inclusion of elderly or impaired people and to foster widespread use of a language through digital means (Soria, Zoli 2012).

In developing the integrated system we describe here we have been inspired by some beliefs. First of all, we believe that any serious project of cultural defence should start from the defence of the language, and that modernization is to be achieved through a written form of the language, as coherent and as widely accepted as possible. We firmly think that digital technologies can play a crucial role in this process of language modernization and in that of promotion and diffusion of the language among younger generations. Finally, speaking of technology, we believe that the highest possible degree of standardization (in file formats, in communication protocols, in programming languages, in DBMS's) is mandatory. Only so it is possible to guarantee "long digital life" to language resources and only so we can allow a real exchange of information, data and technologies.

Several years of experiences have allowed us to reach different results:

- Our platform supports the standard Unicode (diacritics and all sorts of characters are accepted).
- The interface language can be changed very simply at any desired moment.
- There is a high-parameterization (nothing is hard-coded).
- Our software meets industrial standards.
- All modules have achieved a real interoperability.

The final aim was to develop a unique tool which can be integrated in the main writing systems (Word, Libre Office, Web browser, etc.) and which can operate at all the different levels (or modules) of the toolbox. This, we believe, has shown to be one of the most complete and effective "survival kits" for all endangered minority languages such as Tamazight.

Bibliography

Ben Hamadou, Abdelmajid; Mesfar, Slim; Silberztein, Max. "Finite State Language Engineering: NooJ 2009". International Conference and Workshop. Touzeur: Centre de Publication Universitaire, 2010.

Boukous, Ahmed. *Phonologie de l'Amazighe*. Rabat: Institut Royal de la Culture Amazighe, 2009.

Boukhris, Fatima. *La Nouvelle Grammaire de l'Amazighe*. Rabat: Institut Royal de la Culture Amazighe, 2008.

Corongiu, Giuseppe. *Il sardo: una lingua normale*. Cagliari: Condaghes, 2013.

Dell'Aquila, Vittorio; Iannàccaro, Gabriele. *La pianificazione linguistica*. Roma: Carocci Editore, 2011.

Kloss, Heinz "Abstandsprachen und Ausbausprachen". In Göschel, Joachim; Nail, Norbert; Van der Els, Gaston. Zur Theorie des Dialekts: Aufsätze aus 100 Jahren Forschung. Zeitschrift fur Dialektologie and Linguistik, 1976.

Krauwer, Stevem. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap *Proceedings of SPECOM 2003,* Moscow, 2013.

Fellbaum, Christiane, ed. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

Francopoulo, Gil et al. *Lexical markup framework (LMK) Genoa*: LREC, 2006.

Levenshtein, Vladimir I. *Binary codes capable of correcting deletions, insertions, and reversals*. Soviet Physics Doklady, 1966.

Lurà Franco et al. "Dalla carta al web: la versione informatica del lessico dialettale della Svizzera italiana", In: Ruffino G., D'Agostino M. Storia della lingua italiana e dialettologia, atti del VIII Convegno Internazionale dell'Associazione per la Storia della Lingua Italiana, Palermo, 2009.

Németh, László http://hunspell.sourceforge.net/ (27/05/2013).

Pellegrini, Giovan Battista. *Carta dei dialetti d'Italia.* Pisa: Pacini editore, 1977.

Philips Lawrence. "Hanging on the Metaphone", In *Computer Language*, Vol. 7, No. 12 (December), 1990.

Scannel, Kevin. "New computational resources for indigenous and minority languages", 17th annual NAACLT conference. Isle of Man, 2011.

Scannel, Kevin. "Semi-automated construction of semantic networks using web corpora", Words, Texts and Dictionaries conference. University of Wales Centre for Advanced Welsh and Celtic Studies, Aberystwyth, 2008.

Vitali, Daniele. "Appello ai romagnoli per studiare la diversità dialettale" La Ludla XII, 2008.

Soria, Claudia, Zoli, Carlo. “New markets for Language Technology for minority languages”, Maaya Conference. Paris, 2012.

Videsott, Paul. Vocabolar dl Ladin Leterar / Wörterbuch des literarischen Ladinisch / Vocabolario del Ladino letterario (VLL). Projektbeschreibung, 2011.

Zoli, Carlo. “Encouraging the presence in the cyberspace of the lesser used languages through writing and proofing tools: the case of Sardinian language”, Maaya Conference. Paris, 2012.

Zoli, Carlo. “La scrittura standard del romagnolo: un’urgenza non rimandabile” La Ludla IX, 2012.

Zoli, Carlo. “Trattamento digitale delle lingue al servizio delle lingue meno usate”, Corongiu G., Romagnino C. Sa Diversidade de sas Limbas in Europa, Itàlia e Sardigna. Atos de sa cunferèntzia regionale de sa limba sarda, Macumere/Macomer, 2008.

Appendix: Institutions whom which we work

- Institut national des langues et civilisations orientales (**INALCO** - Paris)
- **Rromani** Baxt - Paris
- PARIS 3 (prof. J.-L. Léonard – **Meso-American** languages)
- Chubri, institu d'inventérr e d'valantaij du **Galo**
- Università Orientale di Napoli (prof. M. Gnerre - **Meso-American** languages)
- Chambra d’Òc – **Occitan**, **Francoprovençal**
- Regione Piemonte – Minority Department (**Walser**, **Occitan***)*
- Bureau Régional Ethnographie et Linguistique – Val D’Aosta (**Francoprovençal language**)
- Istituto di Dialettologia ed Etnografia della Svizzera Italiana (**Lombard = north Italian dialects of Italian Switzerland**)
- Ufitziu pro sa **Limba Sarda** – Regione Autònoma de sa Sardigna
- Istitut **ladin** “Micurà de Rü” – Val Gardena-Val Badia
- Istituto culturale **ladino** “Majon di Fascegn” – Val di Fassa
- Union Generèla di **Ladins** dla Dolomites - SPELL
- Istituto Culturale **Mòcheno** Palù TN
- Istituto Culturale **Cimbro** Luserna TN
- Ufici Lenghe **Furlane** – Provincia di Udine
- Agjenzie regjonâl pe lenghe **furlane** (ArLeF)

Asinag-Asinag, 9, 2014, p. 75-90

La réalisation de grands corpus linguistiques berbères normalisés interopérables : enjeux culturels et enjeux d'ingénierie linguistique

Noura Tigziri (1) & Henri Hudrisier (2)

(1) Université de Tizi Ouzou
(2) Laboratoire Paragraphe, Université de Paris 8

We describe our involvement in projects aimed at the production of French and Franco arabo berber digital resources : the BNFB (a project of the OIF FFI [1]) and HumanitéDigitMaghreb (a project of the CNRS ISCC).

In this paper, we focus particularly on the methods used in HumanitéDigitMaghreb (the TEI, specifically applied to the structuration of speech corpora and corpora of poetry and folk tales). The link with the ethnomusicological TEI markup is expected but will be considered later.

We will also examine the practical and future issues of very large corpora, linguistically annotated in accordance with a common standard and designed to constitute, for the linguistic community (for us, the Berber world), the context necessary to interact with the future tools translation and e-semantics

On this last point, for written or oral (audio signal or transcription) corpora, it is essential that the research community about Berber cooperate to promptly equip Berber languages of modern tools for digital processing.

Introduction

La mise en place de corpus numériques est devenue une exigence si on veut sauvegarder notre patrimoine culturel et identitaire mais la création de corpus oraux, par exemple, avec leur transcription, leurs traductions, leur annotation nécessite des méthodes modernes qui puissent faciliter leur exploitation et leur accessibilité. Aussi dans notre article, présentons- nous l'une de ces méthodes, la TEI (Text Encoding Initiative) appliquée à la structuration d'un corpus en kabyle.

En lien avec d'autres collègues, nous sommes impliqués à des niveaux différents dans des projets de bibliothèques numériques, de production d'e-learning, de normalisation des TIC, de recueil linguistique et d'organisation de ressources de documents dans la dynamique des Humanités digitales. Sans pour autant en tirer

prétention, nous considérons que la sélection et le financement conséquents de ces actions de recherches dans des appels d'offres proposés par des instances comme l'OIF ou le CNRS, confortent la pertinence de nos choix méthodologiques. La large assiette des partenariats rassemblés [1] nous permet aussi de disposer de compétences interdisciplinaires (recherche littéraire, ethnologie, linguistique, bibliothéconomie, musicologie, ingénierie linguistique du document et des réseaux, expertise en normalisation) mais aussi de diversité géographique et multilingue. Il nous semble, en effet, primordial que cette diversité des points de vue, des langues, des modes d'expressions, des médias, des genres (écrit, oral, image, musique, théâtre, contes, poésie, etc.) soit pris en compte dans une collégialité numérique véritablement communicante.

Telles sont, en effet, les ambitions primordiales des travaux dans lesquels nous sommes engagés qui nécessitent cette large palette de disciplines, de langues, de métiers et de diversité internationale et institutionnelle :

- Faire communiquer des langues entre elles (la famille des langues berbères, les langues du Maghreb mais aussi à termes les langues mortes qui fondent son patrimoine mais encore rendre accessible les ressources que nous rassemblons de façon mondiale)
- Construire en synergie des corpus de documents numériques en prenant la précaution de négocier leur intercompatibilité normative de façon cohérente et concertée pour que quantité d'utilisateurs [2] (mais aussi de créateurs) de ressources puisse les réutiliser selon des diversités de facettes d'approche. Nous pensons, en effet, que rassembler des ressources numériques doit obligatoirement se faire en ayant le souci de déployer un maximum de scientificité mais en ayant le souci constant que ces travaux soient utilisables par d'autres disciplines, mais aussi puissent participer de prospérité numérique des communautés concernées par ces patrimoines[3].

Si nous affichons ces ambitions c'est parce que nous savons qu'à l'égal de la mutation de la « Galaxie Gutenberg[4] », la mutation de la « Galaxie Digitale » nous impose de prendre en compte les recompositions de collégialité interdisciplinaire et bien sûr la globalisation internationale et interlinguistique.

La Galaxie Gutenberg avait refondé les sciences, l'industrie et l'économie. La Galaxie Digitale nous impose elle aussi de revoir fondamentalement nos méthodes

[1] AUF, ISO, Télécom Paris Sud, Alliance Cartago, EHESS, Université de Paris 8, de Bordeaux 3, de Paris 10, d'Evry, de Tunis, de Tizi–Ouzou, d'Oujda, d'Agadir, de Niamey, Conservatoire de Rimouski-Quebec.

[2] Non obligatoirement prévu à l'origine des projets.

[3] Si le monde académique a d'année en année besoin de plus d'outils, d'équipement, de missions internationales et s'il ne peut raisonnablement augmenter « en pourcentage » sa part du budget national cela impose obligatoirement à prêter attention et à s'inscrire dans des synergies interdisciplinaires, des synergies internationales, des coopérations sciences–industrie et à être attentif aux retombées d'usage pour la prospérité de la société civile.

[4] Cf M. Mac Luhan. Sa thèse impliquait des hypothèses de mutations similaires que l'auteur pointait avec les mass médias des années 60.

selon de multiples impératifs, notamment : interdisciplinarité, approche multimédia, multilinguisme, synergies sciences-industrie, pluralité des usages, mondialisation numérique des ressources (cloud computing), e-sémantique.

Nous ne développons pas dans cet article la facette Bibliothèque numérique de nos travaux. Cette facette sera exposée dans d'autres publications et a été d'ailleurs soumise au comité du TICAM 2012. Il est cependant indispensable de signaler que, bien sûr, la réalisation de grands corpus de documents impose de s'inscrire au minimum dans les recommandations et les bonnes pratiques proposées par l'OCLC[5] et notamment s'appuyer sur le Dublin Core[6] partagé par la plupart des bibliothèques numériques dans le monde. L'avantage majeur de ce respect des recommandations de l'OCLC et du DC étant que, si on en donne l'autorisation, toutes les bibliothèques compatibles dans le monde peuvent venir « moissonner » nos ressources berbères, et que réciproquement nous pouvons enrichir nos propres corpus en venant moissonner nous-mêmes automatiquement toutes les bibliothèques numériques du monde grâce aux mots-clefs ou aux termes « tagués » qui représentent les problématiques qui traversent nos corpus.

Baliser des documents sous plusieurs facettes

Les tâches spécifiques décrites par les auteurs dans ce papier se focalisent plus spécifiquement sur la TEI et son importance grandissante pour rendre disponibles, interopérables, réutilisables et normalisées des ressources linguistiques qui peuvent être indifféremment des corpus oraux, des chansons, ou de la littérature. L'avantage de la TEI pour des ressources numériques, c'est qu'elle autorise des traitements par balisages successifs, facette par facette, et permet ensuite leur alignement multifacette, multisupport, multidisciplinaire et multilingue. Concrètement, une ressource sonore chantée et parlée kabyle pourra être analysée linguistiquement et transcrite, elle pourra être liée et alignée avec sa transcription, son analyse ethnomusicologique, puis ses transcriptions (par ex. en d'autres langues berbères et en fr., ar., es., en. …). Le texte lui-même pourra être l'objet d'un balisage correspondant à des analyses littéraire et poétique, elles aussi, alignées avec les autres facettes.

[5] Le Online Computer Library Center (OCLC), fondé en 1967, nommé à l'origine *Ohio College Library Center*, est une organisation à but non lucratif mondiale au service des bibliothèques dont le but est d'offrir un meilleur accès public aux informations et d'en réduire le coût. Plus de 60 000 bibliothèques dans le monde utilisent les services de l'OCLC afin de trouver, de cataloguer ou de conserver leurs ouvrages. Les bureaux de l'organisation sont situés à Dublin, Ohio (USA).

[6] Le Dublin Core est un schéma de métadonnées générique qui permet de décrire des ressources numériques ou physiques et d'établir des relations avec d'autres ressources. Il comprend officiellement 15 éléments de description formels (titre, créateur, éditeur), intellectuels (sujet, description, langue, …) et relatifs à la propriété intellectuelle. Le Dublin Core fait l'objet de la norme internationale ISO 15836, disponible en anglais et en français depuis 2003. (6 pages, c'est donc une norme extrêmement concise et facile à s'approprier). Il est employé par l'Organisation mondiale de la santé (OMS), ainsi que dans de très nombreuses institutions, états et entreprises. Le Dublin Core a un statut officiel au sein du W3C et bien sûr de l'ISO.

C'est en cela que les exigences de multidisciplinarité, de multilinguisme, de synergie sciences-industrie (notamment industrie de langue) ne sont pas des vains mots. Le cœur de la mutation cognitive de la Galaxie numérique se situe là. Dans une opérabilité synergique numérique d'analyses scientifiques en sciences humaines qui s'additionnent, se recomposent, s'interfécondent de façon croisée. C'est dans ce but que nous nous inscrivons dans le projet HumanitéDigitMaghreb qui a précisément pour objet de pousser plus loin les ambitions du projet BNFB et d'aider les participants francophones, arabophones et berbérophones du projet à s'approprier des méthodes de la TEI. Le but final sera, donc, que nous ne disposions pas seulement de bibliothèques numériques uniquement référentielles (ce qui est déjà bien !), mais que nous mettions en œuvre graduellement un « balisage savant » des ressources (linguistique, littéraire, musicologique) qui donnera une véritable valeur ajoutée notamment aux ressources berbères.

Le courant des Humanités digitales et la TEI

Pour nous, le travail de formalisation des documents linguistiques s'opère sur deux versants complémentaires :

1. Celui de la numérisation des documents (Dublin Core) pour qu'ils puissent devenir disponibles, de façon normalisée et interopérable sur une plate-forme partagée en commun par les participants des projets (la plate-forme OMEKA[7]), mais qu'ils puissent aussi être moissonnés partout dans le monde sur des plates-formes répondant aux spécifications de l'OCLC et qu'à l'inverse, les participants des projets berbères et arabo-berbères précités puissent « moissonner eux aussi des documents dans toutes les bibliothèques numériques ».
2. Celui d'un balisage interne des documents déjà numérisés et référencés pour ce qui est de leur structure formelle, de leur morphologie, de leur signification, de l'ajout de gloses ou de notes, d'hypothèses explicatives ou encore du balisage de leur alignement avec des fichiers associés

[7] Omeka est un logiciel flexible et open source, conçu pour la publication sur le web de collections de documents numériques provenant de bibliothèques, de musées ou d'archives. Le logiciel est développé par le "Roy Rosenzweig Center for History and New Media". L'interface standard permet de parcourir la liste des documents (ou items), d'afficher les fichiers associés à chaque document, de filtrer par mot-clé, de parcourir les collections. Une recherche simple et avancée complète les possibilités de navigation. Des extensions permettent l'ajout de fonctionnalités, facilitant par exemple la création d'expositions électronique. L'administration du site, intuitive et fonctionnelle permet la gestion des collections, des documents et des fichiers associés à chaque document. Le type des documents peut être précisé : texte, image, son, vidéo, cours, histoire orale, email, site web, lien hypertexte, évènement ou personne. Les documents ont le statut public ou privé et peuvent être mis en avant sur la page d'accueil du site. Les informations descriptives de chaque document sont renseignées au format Dublin Core. Des métadonnées supplémentaires peuvent être ajoutées, dépendant du type du document, notamment la TEI. Les fichiers associés peuvent être du type texte (TXT, DOC, PDF, XML, JPG, TIFF), image (GIF, JPEG, PNG, TIFF), son (AIFF, MIDI, MP3, OGG, QT, RA, WAV) ou vidéo (AVI, MPEG, MP4, QT, SWF, WMV).

(transcription, traduction, interprétation ou autre versus médiatique comme des fichiers sonores ou vidéos associés à des textes, des partitions musicales, des photographies, des cartes, des réseaux ou des schémas).

C'est donc sur le point 2 que nous insistons dans cet article.

Le problème posé : la formalisation numérique normalisée des travaux en SHS

Les SHS travaillent globalement sur une matière plus « floue » que les sciences expérimentales et bien sûr que les sciences exactes : leur matériel principal est le document (souvent linguistique), leur outil d'analyse est le plus souvent l'argumentation textuelle et leurs résultats sont globalement des textes.

Evidemment certaines sciences humaines (notamment la linguistique) pratiquent depuis longtemps la formalisation d'un grand nombre de leur description (qu'elles soient morphologiques, syntaxiques, argumentatives, etc.). Cette pratique de formalisation a grandement facilité leur collaboration avec les informaticiens et explique pour partie les progrès en ingénierie linguistique. D'autres sciences humaines, les études littéraires par exemple ont été longtemps et sont aujourd'hui encore globalement rétives à la formalisation de leurs analyses. La recherche littéraire est de ce fait une science qui travaille sur le langage naturel, analyse, pose des hypothèses, les formalise sous forme d'énoncés en langage naturel et communique ses résultats sous la forme quasi exclusive de textes argumentés en langage naturel.

Cependant, notre objectif n'est pas de distribuer des bonnes ou mauvaises notes à telle ou telle catégorie de chercheurs en SHS mais de comprendre la mutation de méthode et d'habitus des chercheurs. De fait, la question qui devient récurrente est celle de la normalisation des pratiques face à une relative prolifération des outils d'aide informatique dans certains segments du travail linguistique.

Si nous nous appuyons sur une typologie grossière des travaux scientifiques et industriels sur le langage, nous pouvons distinguer :

→ Le travail sur les corpus oraux pour lequel il existe une relative prolifération des outils d'aide à la transcription, mais sur lequel il est urgent de s'entendre sur des standards.

→ Le travail d'analyse littéraire qui n'a longtemps connu que des outils très rustiques et limités comme les analyses statistiques de vocabulaire.

→ Les travaux terminologiques et lexicographiques dont les principes et méthodes ont été normalisés très tôt grâce notamment à Eugen Wüster qui a perçu très vite l'obligatoire nécessité de normaliser les pratiques en fondant dès 1937 ce qui allait devenir le Comité Technique 37 de l'ISO (ISO TC37).

→ Les travaux sur l'informatisation de l'écriture : la question est largement connue à l'IRCAM. Notons cependant que le scénario historique de ce qui s'est passé entre les années 1960 et aujourd'hui est une excellente leçon d'évolution technologique et de la longue durée d'appropriation technique, de l'impérieuse nécessité de

s'inscrire dans la normalisation et de la nécessité de comprendre le lien entre les progrès de l'environnement technique[8]. Pour des raisons historiques, l'écriture latine non accentuée a été dès le début prise en compte et normalisée. On connaît ensuite les normes successives et notamment la famille ISO 8859 qui prenait en compte les grandes écritures alphabétiques (latine, cyrilliques, arabe, grec, hébraïque…) mais qui ne pouvait coder ni les écritures idéographiques, ni les écritures sans intérêt industriel évident comme le tifinagh ou les écritures archéologiques (cunéiformes, hiéroglyphes). Sur ces derniers segments, on a bien sûr assisté à une relative prolifération de standards propriétaires « bricolés » par des laboratoires ou de petites sociétés informatiques. C'est ensuite grâce aux efforts des équipes des « chercheurs de terrain » (notamment à l'IRCAM) que la normalisation de ces technologies de transition a pu se faire.

Nous avons insisté sur cette question de la numérisation des écritures (qui pose désormais peu de problèmes) parce qu'elle est emblématique de l'obligatoire normalisation pour passer du foisonnement antiproductif des « standards propriétaires » comme c'était le cas avant Unicode et comme c'est encore le cas pour la transcription des corpus oraux.

En effet, comme le signale Thomas Schmidt[9], il existe aujourd'hui un choix relatif pour des outils d'aide à la transcription (de transcription informatique des corpus oraux[10]) et complémentairement une relative profusion de standards de formats pour coder de façon inutilement distincte les objets, les évènements, le contexte et les textes résultant de la transcription de ces corpus oraux. Certes, ces outils et formats présentent des différences mineures dues au contexte de leur développement et de leur production.

Par exemple, CLAN/CHAT a été développé pour transcrire et coder des corpus oraux d'enfants dans la base CHILDES alors que EXMARaLDA Partitur-Editor a été développé dans un contexte d'étude du multilinguisme et de la dialectologie. Tous ces outils ont des fonctionnalités similaires qui leur permettent simultanément de disposer d'un « player son » visualisant « l'enveloppe des productions sonores » et de zones de capture textuelle présentée en lignes parallèles (voir ci-dessous figure 1, une capture d'écran d'EXMARaLDA).

Pour compliquer encore la situation, les grands corpus mondiaux de transcription orale, ont bien naturellement développé des « conventions propriétaires » de codage des résultats dans leurs bases (voir ci-dessous figure 2, un tableau récapitulatif selon Thomas Schmidt).

[8] Dans ce cas particulier la nécessité d'attendre les progrès d'une informatique à 8 bits puis des processeurs à 16, 32, 64 bits et plus qui permettent maintenant de travailler directement en Unicode ce qui était plus problématique quand les processeurs étaient à 8 bits.

[9] Thomas Schmidt, « A TEI-based Approach to Standardising Spoken Language Transcription », *Journal of the Text Encoding Initiative* [Online], Issue 1 | June 2011, Online since 08 June 2011, connection on 08 July 2012. URL : http://jtei.revues.org/142 ; DOI : 10.4000/jtei.142

[10] ANVIL, CLAN/CHAT, ELAN, EXMARaLDA Partitur-Editor, FOLKER, Praat, Transcriber

File formats and transcription conventions for different spoken language corpora

Corpus (Language) [URL]	**File format**	**Transcription convention**
SBCSAE (American English) [http://projects.ldc.upenn.edu/ SBCSAE/]	SBCSAE text format	DT1 (DuBois et al. 1993)
BNC spoken (British English) [http://www.natcorp.ox.ac.uk/]	BNC XML (TEI variant 1)	BNC Guidelines (Crowdy 1995)
CallFriend (American English) [http://talkbank.org/]	CHAT text format	CA-CHAT (MacWhinney 2000)
METU Spoken Turkish Corpus (Turkish) [http://std.metu.edu.tr/ en]	EXMARaLDA (XML format)	HIAT (Rehbein et al. 2004)
Corpus Gesproken Nederlands (CGN, Dutch) [http://lands.let.kun.nl/ cgn/ ehome.htm]	Praat text format	CGN conventions (Goedertier et al. 2000)
Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK, German) [http://agd.ids-mannheim.de/ html/ folk.shtml]	FOLKER (XML format)	cGAT (Selting et al. 2009)
Corpus de Langues Parlées en Interaction (CLAPI, French) [http://clapi.univ-lyon2.fr/]	CLAPI XML (TEI variant 2)	ICOR (Groupe Icor 2007)
Swedish Spoken Language Corpus (Swedish) [http://www.ling.gu.se/ projekt/ old_tal/ SLcorpus.html]	Göteborg text format	GTS (Nivre et al. 1999)

Pour ce qui est du traitement numérique des corpus oraux, nous sommes donc confrontés à une situation tout à fait similaire à celle de la codification des écritures avant leur normalisation convergente avec Unicode. Il existe une relative anarchie des outils d'aide à la capture et à la transcription ainsi que des formats de codages. Comme le fait remarquer Lou Burnard[11], « Le constat est récurrent : à la variété des formats utilisés se superpose l'incohérence des pratiques conventionnelles de transcription des données orales. En dépit de plus de vingt années de pratiques convergentes, les communautés intéressées préfèrent travailler avec leurs propres

[11] Lou Burnard : « Encoder l'oral en TEI : démarches, avantages, défis…. » Conférence à la Bibliothèque Nationale de France, prononcée le 10 mai 2012, Publié le 19/06/2012 par Abigaël Pesses.

outils et conventions “maison”. Pourtant, l’intérêt de se servir d’un format commun, voire pivot, est un sujet qui a été abordé dans la littérature académique à de multiples reprises : Edwards & Lampert (1993), MacWhinney (2007), Schmidt (2011). Ne serait-il pas finalement temps d’établir un format d’échange normalisé pour les données orales? …[La TEI grâce à son format fédérateur TEI transcription of speech et aussi grâce à son alliance en consortium avec l’ISO TC37 est actuellement en situation de devenir la norme[12]] ».

Dans les projets auxquels nous faisons références (BNBF et HumanitéDigitMaghreb) les participants se réclament globalement de ces deux disciplines, souvent des deux ensembles, mais aussi de la bibliothéconomie, l’histoire, la pétrographie, l’ethnologie, l’ethnolinguistique, la musicologie.

En nous inscrivant dans l’école de pensée des Humanités digitales et de la TEI qui est sa norme et son outil technique, nous voulons explicitement donner non seulement une réalité tangible et numérique à nos travaux, mais aussi les rendre facilement échangeables, cumulables, améliorables, modifiables partout dans le monde. Nous voulons que nos travaux linguistiques participent « pré-industriellement de l’ingénierie linguistique. Nous voulons que nos travaux de recherche littéraire soient non seulement visibles dans le monde entier, mais encore qu’ils s’inscrivent dans la synergie mondiale des études littéraires computorisables. Nous voulons aussi sur un plan plus spécifiquement pan berbère que nos travaux soient déjà facilement échangeables et cumulable entre nous et avec nos trois langues partenaires maghrébines (arabe, français, espagnol auxquelles il convient de rajouter l’anglais). C’est la raison primordiale de notre implication dans le projet HumanitéDigitMaghreb.

Qu’est-ce que HumanitéDigitMaghreb ?

HumanitéDigitMaghreb est un projet du CNRS-ISCC[13]. C’est une recherche-action dans laquelle sont engagés des acteurs de terrain (linguistes, chercheurs en littérature, culture, histoire tant française qu’arabe ou berbère) soucieux d’inscrire leurs pratiques dans l’organisation rationnelle de corpus numériques répondant aux recommandations mondiales des réseaux de bibliothèques numériques (OCLC) et des Humanités digitales. Ces acteurs de terrain sont associés avec des praticiens de la coopération francophone numérique, des spécialistes de l’information et de la communication, des fondateurs de la TEI et des Humanités digitales en France, des spécialistes de l’appropriation des usages du numérique et plus spécifiquement des patrimoines numériques. Parallèlement à son étude d’appropriation, cette recherche s’appuiera sur des réalisations en cours de structuration de corpus patrimoniaux franco-arabo-berbères.

[12] Cette conclusion entre crochets est d’Henri Hudrisier mais correspond à l’action des leaders de la TEI effectivement liés aux actions de l’iso TC37, notamment Laurent Romary convener de l’ISO TC37-SC4.

[13] Institut des sciences de la communication du CNRS (Centre National de la Recherche Scientifique).

Les partenaires historiques de la TEI

La TEI (Text Encoding Initiative) a été fondée à la suite d'une conférence sponsorisée par l'ACH (Association for Computers and the Humanities)[14] et financée par le NEH (U.S. National Endowment for the Humanities)[15]. Cette conférence avait lieu au Vassar College (Poughkeepsie, N.Y. - USA) en novembre 1987. Environ trente représentants du monde des bibliothèques, des sociétés savantes et de projets de recherche intéressés par le codage des textes et la recherche littéraire ainsi que d'informaticiens spécialisés en SGML étaient invités à cette conférence pour discuter la faisabilité d'un codage standard et élaborer des recommandations. Pendant la conférence, l'ACL (Association for Computational Linguistics)[16] et l'ALLC (Association for Literary and Linguistic Computing)[17] ont décidé de rejoindre l'ACH comme les sponsors d'un projet pour développer les Directives de la TEI (TEI Guidelines). En 1988, ils ont été rejoints par la Commission de la Communauté Européenne, l'Andrew W. Mellon Foundation[18] et le Social Science and Humanities Research Council of Canada[19].

[14] L'ACH (Association for Computers and the Humanities) a été fondée en 1978, une époque où la relation entre informatique et humanités, était encore très confidentielle. La plupart des grands universitaires du domaine jugeaient même qu'il s'agissait d'une alliance contre nature. En une trentaine d'année, le paysage a bien changé. L'ACH a mis en place un forum pour la recherche, des discussions et les explorations techniques qui ont alimenté cette transformation. L'ACH est devenue une association beaucoup plus vaste. Elle patronnait chaque année ; la conférence d'Humanités Numériques annuelle (maintenant patronnée par ADHO.

[15] Le NEH (U.S. National Endowment for the Humanities) est une agence fédérale américaine indépendante fondée en 1965 par le Président Lyndon Johnson. C'est le plus important organisme de financement dans le secteur des Humanités aux USA. Il intervient pour financer l'excellence culturelle, muséale, académique mais aussi la radio et la télévision, voire des bourses de recherches individuelles.

[16] L'ACL (Association for Computational Linguistics) est l'Association de référence mondiale pour les professionnels et les scientifiques travaillant sur les questions liant langages naturels et traitement informatique. L'ACL édite Computational Linguistics et organise des conférences annuelles (la 51ème conférence est prévue en 2013 à Sofia).

[17] L'ALLC (Association for Literary and Linguistic Computing) a été fondée en 1973 dans le but de favoriser des applications d'informatisation de l'étude du langage et de la littérature. L'ALLC s'intéresse à l'analyse des textes, aux corpus textuels, à l'histoire, l'histoire de l'Art, la musique, l'étude des manuscrits et à l'édition électronique.

[18] La Fondation Andrew W. Mellon de New-City et Princeton est une fondation privée, dotée de richesses accumulées par Andrew W. Mellon de la famille Mellon (Pittsburg, Pennsylvanie). C'est une fondation prestigieuse qui intervient dans l'enseignement supérieur, les bibliothèques et la communication savante, les musées et la conservation de l'Art, les arts de la scène, et les TIC. Plus précisément le développement de logiciels interréssant ses principaux champs d'intérêts ci-dessus.

[19] Le Social Science and Humanities Research Council of Canada en français Conseil de recherche en sciences humaines du Canada (SSHRCC- CRSHC) est un organisme du gouvernement fédéral canadien ayant pour mission d'appuyer la recherche et la formation avancée en milieu universitaire dans le secteur des sciences humaines.

On voit bien que ces associations fondatrices ne sont pas d'obscurs partenaires, ces diverses institutions opéraient une importante jonction synergique en fondant la TEI. Certes, la TEI a contribué à ce qu'un vaste public savant s'approprie des « standards bonnes pratiques » en matière de traitement et de communication pour l'étude savante des textes. Parallèlement, les institutions fondatrices n'oubliaient pas leurs objectifs fondateurs réciproques éminemment complémentaires : Humanités computationelles ; recherche littéraire par ordinateur ; linguistique computationelle ; recherche littéraire computationelle ; développement de logiciel pour le traitement de corpus culturels numériques et bibliothèques.

C'est d'ailleurs dans la perspective de ces objectifs que l'ALLC, en coopération avec l'ACH et la SDH-SEMI[20] ont préfiguré (dès 2002) puis fondé en 2005 l'ADHO[21].

On voit bien ainsi la synergie qui peut exister entre la TEI qui définit des standards et des bonnes pratiques et les Humanités digitales qui permettent que se socialisent ses usagers, qu'ils adaptent les outils (notamment ceux des bibliothèques numériques) à des besoins spécifiques, qu'ils échangent des méthodes, des modèles de structuration et de balisage de leurs corpus (en fait des TEI-DTD adaptées aux besoins de leurs corpus et de leurs pratiques d'analyse savantes et d'échanges).

Une synergie TEI, Humanités digitales et bibliothèques numériques

Toutes les universités héritières de ces premières universités européennes travaillant en latin utilisent le terme « Humanitas » pour désigner les disciplines de sciences humaines et sociales, ainsi que la recherche en Art et littéraire. Pour des raisons historiques, les institutions académiques anglophones gardent toujours vivant la désignation « Humanités » qui constitue toujours une sorte de métadiscipline recouvrant pratiquement ce que les francophones nommeraient « Arts et Lettres » parce qu'en français, l'expression «, ' les Humanités » est devenue un peu désuète. Quels que soient les termes, le monde anglo-saxon puis l'Europe du Nord et, avec un certain retard, la Francophonie s'emparent maintenant de l'expression Humanités digitales, ce qui redonne du sens à l'ancienne expression « les Humanités ».

En fait, on pourrait dire que de la rencontre de ces institutions et de leur convergence synergique sont nés deux axes de dynamique d'action fonctionnellement complémentaires qui rentrent en résonance avec une réalité de l'environnement technologique : les bibliothèques numériques. L'ADHO a adopté comme publication principale, le journal officiel de l'ALLC « Journal of Digital Scholarship in the Humanities » publié par les Oxford University Press. Deux

[20] Society for Digital Humanities-Société d'étude des médias interactifs (CAN).

[21] L'ADHO (Alliance of Digital Humanities Organizations) est donc une alliance internationale qui a pour objectif de soutenir les applications informatiques pour l'étude du langage et de la littérature : en fait les Humanités digitales. Elle le fait en soutenant des publications, des ateliers spécialisés (classes d'été), à travers aussi des groupes de travail thématiques répondant notamment à des disciplines et des sous-diciplines.

autres publications ont une portée mondiale en la matière : « DHQ, Digital Humanities Quartely » et « Digital Studies / Le champ numérique[22] tous deux publiés sous la responsabilité de l'ADHO.

Citons encore pour mémoire afin de survoler la problématique des Humanités digitales : *Humanist:* Un séminaire électronique sur les applications de l'informatique aux Humanités http://www.allc.org/publications/humanist

Mind Map of the Digital Humanities: Une cartographie conceptuelle de l'univers des Humanités digitales disponibles sur http://www.allc.org/publications/mind-map-digital-humanities

Cela permet d'avoir une vision synoptique et facile d'accès mise à jour en permanence par l'ensemble des communautés TEI, des publications, des outils disponibles. Notre ambition serait que TEI berbère y figure bientôt.

Dans cette partie, nous appliquons le codage TEI sur un corpus kabyle. Nous avons choisi de travailler sur un écrit parce que l'oral présuppose un certain nombre de décisions à prendre en ce qui concerne la définition de certains concepts tels par exemple la phrase, l'énoncé, le paragraphe... C'est pour cela que pour cette première application, nous avons jugé plus pratique de travailler sur un corpus écrit. Il s'agit de la traduction en kabyle de Kamal Bouamara de "Jours de Kabylie" de Mouloud Feraoun. L'oeuvre contient un certain nombre de parties. Nous avons travaillé sur deux parties pour montrer comment se fait le codage en TEI.

La première partie est "Taddart-iw" (mon village), la deuxième est "Tajmaɛt n At Flan" (la djemaa de Flan (un tel)).

Tout codage en TEI commence par la définition des éléments à mettre dans le <TeiHeader>. Pour notre part, nous avons le <TeiHeader>, comme ceci :

[22] *Digital Studies / Le champ numérique* (ISSN 1918-3666) est une publication universitaire spécialisée paraissant trois fois par an, destinée aux chercheurs dans le domaine des sciences sociales numériques et ayant pour objectif de leur offrir une ressource de niveau universitaire et de fournir un cadre formel à leurs activités de recherche. DS/CN est publiée par la Society for Digital Humanities / Société pour l'étude des médias interactifs (SDH/SEMI), un organisme affilié à l'Association for Computers and the Humanities (ACH) et à l'Association for Literary and Linguistic Computing (ALLC), via l'Alliance of Digital Humanities Organisations (ADHO).

```
<TeiHeader>
        <fileDesk>
                <
                <publicationStmt>
                        <publisher> ENAG </publisher>
                        <pubPlace> Alger </pubPlace>
                        <date>1998 </date>
                </publicationSmt>
        </fileDesk>
</teiHeader>
```

Avec deux attributs dans le <fileDesk>, le titre <titleStmt> qui précise tout ce qui est relatif au titre avec l'intitulé de l'ouvrage, l'auteur et on a ajouté la balise <editor> pour spécifier que c'est une traduction et la balise <relatedItemtype="translatedFrom"> pour donner la traduction.

```
        <titleStmt>
                        <title> Ussan di Tmurt </title>
                        <author> Mouloud Feraoun </author>
                        <editor role="translator"> Kamal Bouamara </editor>
                        <relatedItemtype="translatedFrom">
                          <bibl>
                                <author>Mouloud Feraoun</author>
                                 <title>Jours de Kabylie</title>.
                                   <date>1954</date>
                        </bibl>
                         </relatedItem
                </tileStmat>
```

et les données concernant la publication de cet ouvrage comme, l'éditeur la date, le lieu de publication.

```
                <publicationStmt>
                        <publisher> ENAG </publisher>
                        <pubPlace> Alger </pubPlace>
                        <date>1998 </date>
                </publicationSmt>
```

Une fois ces données introductives définies, nous passons au codage du texte lui-même.

```
<text>
        <body>
                <div n=1>
                        <head> taddart-iw </head>
```

Avec le corps du texte <body> comprenant un attribut qui lui-même est subdivisé en <head> (entête) et <p> (paragraphe).

Dans cette dernière balise nous définissons une balise <S> (phrase). Nous définissons la phrase au sens large du mot. Un segment compris entre deux points. Comme nous avons à prendre par moment des décisions par rapport à la définition du « mot » et du « mot composé » qui a deux parties reliées par un trait d'union, phénomène assez courant en kabyle, nous avons opté pour considérer que le mot composé de n éléments est un seul mot avec comme la définition du mot « tout élément compris entre deux blancs », dans ce cas, le codage en TEI se fait de cette manière

Exemple :

d win d-yettalsen (mot composé : d-yettalsen)

<w>d</w>

<w>win</w>

<w>d</w>

<hyphen>-</hyphen>

<w>yettalsen</w>

Si on considère le mot composé de n parties comme étant un seul mot, dans ce cas, le codage se fait ainsi :

Exemple : d win d-yettalsen (mot composé : d-yettalsen)

<w>d</w>

<w>win</w>

<w>d<hyphen>-</hyphen>yettalsen</w>

Ainsi le codage TEI du paragraphe et de la phrase est donné par ce qui suit :

```
<p>
    <s>
        <w>Ur</w>
        <w>lliɣ</w>
        <w>ara</w>
        <w>seg</w>
        <w>wid</w>
        <hyphen>-</hyphen>
        <w>nni</w>
        <w>yettɣuddun</w>
        <w>taddart</w>
        <hyphen>-</hyphen>
        <w>nsen</w>
        <pc>.</pc>
```

```
</s>
<s>
        <w>Mi</w>
        <w>ur</w>
        <w>fneṭzeɣ</w>
        <w>ara</w>
        <w>s</w>
        <w>waya</w>
        <w>nezzeh</w>
        <pc>,</pc>
        <w>ẓriɣ</w>
        w>acuɣer</w>
        <pc>.</pc>
</s>
```

Le deuxième corpus est un corpus oral. Il s'agit d'un meeting tenu par le président du parti du Rassemblement pour la culture et la démocratie, le docteur Saïd Saadi, lors des élections législatives de 2002.

Codage du TEI HEADER

```
<teiHeader>
   <fileDesc>
    <titleStmt>
       <title>Meeting politique</title>
          <author> Said Sadi="Président du  RCD"  </author>
    </titleStmt>
    <publicationStmt>
       <pubPlace>type="Tizi Ouzou">Stade Oukil Ramdane</pubPlace>
          <date> 2002.05.02</date>
    </publicationStmt>
    <sourceDesc>
       <recordingStmt>
        <recording type="audio" dur="P30M">
        <equipment>
           <p> audio tape, réalisé par B. S.</p>
         </equipment>
        </recording>
       </recordingStmt>
      </sourceDesc>
    </fileDesc>
 </teiHeader>
```

Codage du corps du texte

```
<text>
   <body>
     <incident>
       <desc>Applaudissement</desc>
     </incident>
     <u>who="# Saïd  Sadi ">
       <seg>Azul</seg>
     </u>
     <incident>
       <desc>Applaudissement</desc>
     </incident>
     <u>who="# Saïd  Saadi ">
       <seg>Azul <pause dur="PT10S"/> Azul d ameqran</seg>
     </u>
     <incident>
       <desc>Applaudissement</desc>
     </incident>
     <u>who="# Saïd  Saadi ">
       <seg>Tsellem-d deffir kunwi i  igubrentili</seg>
     </u>
     <incident>
       <desc>Applaudissement</desc>
     </incident>
     <u>who="# Saïd  Sadi ">
       <seg>
         Azul<pause dur="PT20S"></pause>
       </seg>
       <seg>Yidwen am yiḍelli am assa am uzekka wer ttagadut<pause
dur="PT20S"/></seg>
     </u>
     <incident>
       <desc>App</desc>
     </incident>
     <u>who="# Saïd  Saadi ">
       <seg>Qqaren- as imezwura- nneɣ<pause dur="PT10S"/> isers uḥeddad
tafḍist<pause dur="PT10S"/>irfed-itt mmi-s<pause dur="PT20S"/></seg>
     </u>
     <incident>
       <desc>Applaudissement</desc>
     </incident>
     <u>who="# Saïd  Sadi ">
       <seg>D ayen igellan di tiɣri n nouvembre rebɛa uxemsin<pause
dur="PT10S"/>Dayen igellan di la plate forme n la soumam<pause dur="PT05S"/>
ayen i gellan<pause dur="PT05S"/>deg dusyi –nni i d nexdem deg Σekkuren deg
seggasen n tmanyin<pause dur="PT05S"/> i gellan di la plate forme Lleqsar<pause
dur="PT20S"/> </seg>
```

```
    </u>
  </body>
</text>
```

Conclusion

La TEI a l'avantage de rendre disponibles, interopérables, réutilisables et normalisées des ressources linguistiques. Il est vrai que le travail de codage est fastidieux surtout quand il s'agit de travailler sur des corpus de grandes tailles mais vu les avantages que présente cette méthode, nous avons tout intérêt à l'exploiter et nous l'approprier à cause de l'un de tous ses avantages dont l'interopérabilité.

Quel que soit le corpus choisi, la disponibilité des balises de codage rend la tâche de balisage plus facile à appréhender. En effet, les membres du consortium qui ont établi la TEI ont prévu absolument toutes les balises utiles dans le codage de ressources linguistiques quelles qu'elles soient : corpus écrit, corpus oral et de quelle que soit la discipline : linguistique, littérature, musique

Bibliographie

Ben Henda M., (2012), *Vision historique, technique et prospective des systèmes d'information et de communication interopérabilité normative globalisée*. Mémoire de HDR sous la dir. De Roland Ducasse, Université Bordeaux III.

Ben Henda M. et Hudrissier H., (2000), Normalisation et terminologies multilingues pour les TICE in *Forum Terminologique International*, Université de Sousse 20 au 23 novembre 2000.

Gille B., (1978), *Histoire des techniques*, Encyclopédie de la Pléiade, Gallimard, Paris.

Hudrissier H. et Romary L., (2003), Le balisage normalisé des concepts et documents en liaison avec les normes de l'CAD. In *Normes et standards pour l'apprentissage en ligne*, Versailles, 18 mars 2003. http://www.initiatives.refer.org/inititaives, 2012.

Hudrissier. H. (2009), La nécessité d'adapter internet à la mondialisation linguistique, *in Critique de la société de l'information* (coordonné par J.P. Lafrance). Les Essentiels d'Hermès, CNRS éditions, Paris, p. 115-134.

Asinag-Asinag, 9, 2014, p. 91-104

Utilisation des CACs et des ressources externes pour l'amélioration des performances de l'étiquetage morphosyntaxique[1]

Mohamed Outahajala (1), Lahbib Zenkouar (1),
Yassine Benajiba (2), Paolo Rosso (3)

(1)[2] LEC–EMI, Université Mohammed V- Agdal, Maroc
(2) Thomson Reuters, New York, USA
(3)[3] Universidad Politécnica de Valencia, Spain

La langue amazighe, comme la plupart des langues de moindre diffusion, souffre encore de la pénurie d'outils et des ressources pour son traitement automatique, en particulier les corpus annotés. Ces derniers sont plus difficiles à construire que les corpus bruts qui à leur tour nécessitent, dans la majorité des cas, des prétraitements. L'objectif de cet article est de présenter une approche basée sur l'apprentissage semi-supervisé visant l'utilisation d'un corpus de textes bruts, sélectionnés sur la base de la mesure de confiance des Champs Aléatoires Conditionnels(CACs), conjointement avec un corpus annoté manuellement de 20k morphèmes. Les résultats des expérimentations préliminaires montrent une réduction du taux d'erreur de l'étiqueteur morphosyntaxique de 1,3%. De même, la réduction du taux d'erreur est-elle de 5,9%, entre 60% et 90% du corpus, lorsque le modèle est entrainé par les phrases du corpus brut annotées automatiquement.

Amazigh language, and like most of the languages which have only recently started being investigated for the Natural Language Processing (NLP) tasks, lacks annotated corpora and tools and still suffers from the scarcity of linguistic tools and resources and especially annotated corpora. Creating labeled data is a hard task. However, obtaining unlabeled data, although needing most time preprocessing for languages with scarce resources, is less difficult. The aim of

[1] Le premier auteur exprime sa gratitude à la CODESRIA. Les travaux du quatrième auteur ont été financés dans le cadre des projets de recherche: VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems, la commission européenne WIQ-EI IRSES (no. 269180) et DIANA-APPLICATIONS (TIN2012-38603-C02-01).

[2] Laboratoire Electronique et Communication, Ecole Mohammadia d'Ingénieurs (EMI).
[3] Natural Language Engineering Lab – EliRF, DSIC.

this paper is to present a semi-supervised based approach using labeled and unlabeled data. Preliminary results show an error reduction of 1,3%, when training our POS tagger with Conditional Random Fields(CRFs), with chosen automatically annotated texts and a small manually annotated corpus of about 20k tokens. Also, when trained with automatically annotated data, the achieved improvement between 60% and 90% of the trained data is 5.9%.

1. Introduction

L'étiquetage morphosyntaxique consiste en l'annotation de chaque mot d'une phrase avec une étiquette récapitulant une information morphosyntaxique selon le contexte. Il augmente l'information des mots étiquetés pour les couches supérieures pour le traitement automatique des langues(TAL). Il s'agit de la première couche au-dessus du niveau lexical et le niveau le plus bas de l'analyse syntaxique. Ainsi, toutes les tâches traitant des niveaux linguistiques supérieurs, utilisent le POS tagging, par exemple : l'analyse partielle ; la désambiguïsation des sens des mots; l'affectation des fonctions grammaticales, la reconnaissance d'entités nommées, etc. (Manning & Schütze, 1999, Cutting et *al*., 1992, Benajiba et *al*., 2010).

Dans la littérature, il a été démontré que les approches basées sur l'apprentissage supervisé sont les plus efficaces pour construire les étiqueteurs grammaticaux, en s'appuyant sur un corpus annoté manuellement et souvent d'autres ressources, telles que des dictionnaires et des outils de segmentation. Dans l'approche que nous proposons dans ce papier, nous utilisons des techniques de classification des séquences, basées sur les CACs et conjointement des données étiquetées et non étiquetées, pour construire notre étiqueteur grammatical. D'une part, nous utilisons un corpus de ~20k mots annoté manuellement (Outahajala et *al*., 2011a) pour former nos modèles et les caractéristiques n-grammes lexicales pour aider à augmenter la performance ainsi que des ressources externes qui consistent en un ensemble de textes bruts.

Le papier est organisé comme suit : en section 2, nous présenterons les travaux connexes sur les techniques d'étiquetage morphosyntaxique. Puis, dans la section 3 nous donnerons le cadre théorique des CACs. Dans la section 4, nous décrirons les expériences et nous discuterons les résultats. Enfin, dans la section 5, nous dresserons quelques conclusions et nous présenterons les travaux à effectuer dans le futur proche.

2. Etat de l'art

De nombreux systèmes pour l'étiquetage automatique des parties du discours ont été développés pour un large éventail de langues. Parmi ces systèmes, certains s'appuient sur les règles linguistiques et d'autres sur des techniques d'apprentissage automatique (Manning & Schütze, 1999, Jurafsky & Martin, 2009). Les premiers étiqueteurs morphosyntaxiques étaient principalement à base de

règles. La construction de tels systèmes nécessite un travail considérable afin d'écrire manuellement les règles et de coder les connaissances linguistiques qui régissent l'ordre de leur application. Un exemple d'étiqueteur à base de règles est TAGGIT, développé par Green et Robin (Greene & Rubin, 1971) et contenant environ 3300 règles, ce système atteint une précision de 77%. Par la suite, l'apprentissage automatique des étiqueteurs s'est avéré à la fois moins pénible et plus efficace que ceux à base de règles. Dans la littérature, de nombreuses méthodes d'apprentissage ont été appliquées avec succès pour réaliser des POS taggeurs, tels que les Modèles de Markov Cachés (HMM) (Charniak, 1993), la transformation système basé sur la réduction du taux d'erreur (Brill, 1995), le modèle d'entropie maximale (Ratnaparkhi, 1996), les arbres de décision permettent de construire (Schmid, 1999), sur la base d'un corpus de référence, un outil d'aide à la décision qui utilise ce modèle. Les méthodes d'apprentissage automatique permettent de construire des modèles complexes (comportant de très nombreux paramètres), chose qui est difficile à faire manuellement. La qualité des modèles est souvent liée à la quantité de données utilisées dans l'apprentissage. Ainsi, à partir d'exemples appris précédemment, les programmes s'appuyant sur ces méthodes affectent l'étiquette aux mots selon le contexte. Parmi les travaux basés sur l'apprentissage qui ont donné de bon résultats, on cite ceux de Kudo & Matsumoto (2000) et de Lafferty et *al*. (2001).

Bien que ces méthodes aient une bonne performance, la précision des mots inconnus, mots hors vocabulaire du corpus de test par rapport au corpus d'apprentissage, est beaucoup plus faible que celle des mots connus, ce qui est problématique lorsque le corpus d'apprentissage est de petite taille.

Dans la pratique, la plupart des analyseurs limitent le nombre d'étiquettes en ignorant certaines distinctions difficiles à désambiguïser automatiquement, ou sujettes à discussion du point de vue linguistique.

En raison de sa morphologie complexe (Chafiq, 1991 ; Ameur et *al*. 2004; Ameur et *al*. 2006; Boukhris et *al*. 2008) ainsi que l'utilisation des différents dialectes dans sa normalisation, la langue amazighe présente des défis intéressants, pour les chercheurs en TAL, qui doivent être pris en compte. Concernant la tâche d'étiquetage morphosyntaxique, certains défis du TAL pour l'amazighe sont les suivants :

1. L'amazighe dispose de sa propre graphie : le Tifinaghe, qui s'écrit de gauche à droite ;
2. Il ne contient pas de majuscules ;
3. Les noms, les noms de qualité, les verbes, les pronoms, les adverbes, les prépositions, les focaliseurs, les interjections, les conjonctions, les pronoms, les particules et les déterminants consistent en un seul mot entre deux blancs ou des signes de ponctuation. Toutefois, si une préposition ou un nom de parenté est suivi par un pronom personnel, à la fois la préposition/nom de parenté et le pronom qui suit, forment chaîne unique délimitée par des espaces ou des signes de ponctuation. Par exemple : □□ (ɣr) signifiant « pour, au » + □ (i) qui signifie « moi » (pronom personnel première personne du singulier) donnent «□□□□/□□□□ (ɣari/ɣuri) » ;

4. Les signes de ponctuation amazighe sont semblables aux signes de ponctuation adoptés au niveau international et ont les mêmes fonctions[4]. Les lettres majuscules, néanmoins, ne se produisent ni au début ni à l'initiale des noms propres ;
5. A l'instar d'autres langues naturelles, l'amazighe peut présenter des ambiguïtés au niveau des classes grammaticales. En effet, la même forme de surface peut appartenir à plusieurs catégories grammaticales selon le contexte dans la phrase. Par exemple, □□□□ (illi) peut fonctionner comme verbe à l'accompli négatif, il signifie « il n'existe pas », ou comme nom de parenté « ma fille ». Quelques mots tel que « □ » (d) peuvent fonctionner comme préposition ou une conjonction de coordination ou particule de prédication ou d'orientation ;
6. De même que la majorité des langues dont les recherches en TAL ont récemment commencé, l'amazighe est peu doté en ressources langagières et outils du TAL.

3. Les Champs Aléatoires conditionnels

Les CACs ou CRFs sont des processus stochastiques qui modélisent les dépendances entre un ensemble d'observations discrètes réalisées sur une séquence discrète et un ensemble d'étiquettes. Dans le cas de l'analyse morphosyntaxique la suite des mots est la séquence discrète. En comparaison avec les Modèles de Markov Cachés, un CAC ne repose pas sur l'hypothèse forte d'indépendance des observations entre elles conditionnellement aux états associés.

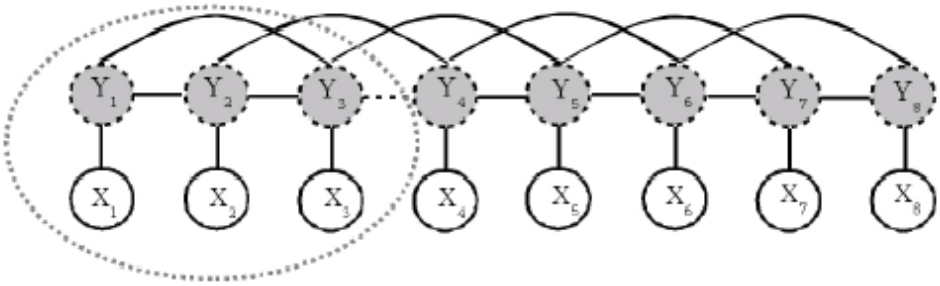


Figure 1 : Exemple d'un graphe des CACs, la partie encerclée est une clique.

Les CACs sont des modèles graphiques probabilistes se basant et sur la théorie des graphes et sur la théorie des probabilités. Ces deux théories permettent de modéliser le problème de classification des séquences : la théorie des graphes

[4] Les deux caractères : ⵯ (2D70) et ⵿ (2D7F) sont deux signes de ponctuation supplémentaires utilisés par les Touaregs. Ils font désormais suite à un amendement du standard Unicode, partie des caractères tifinaghes dont la liste actualisée est sur : http://www.unicode.org/charts/PDF/U2D30.pdf .

permet la modélisation des structures de séquence des étiquettes des phrases, quant à la théorie des probabilités, elle permet de gérer les ambiguïtés causées par les séquences des étiquettes. Les CACs sont avec les Modèles de Markov à Entropie Maximale(MMEMs) les deux principaux modèles discriminants. Bien que les MMEMs aient obtenu de bons résultats sur les tâches d'extraction d'information et de segmentation (MCallum, 2000), ils souffrent du problème du biais du label. En effet, si le graphe est tel qu'un nœud *i* n'a qu'un successeur *i+1*, alors la masse de probabilité est entièrement transmise à y_{i+1} indépendamment des observations x, appelé biais du label. Les CACs permettent de palier à ce problème et cela en calculant les poids de transition non normalisée et en calculant un facteur de normalisation sur l'ensemble de la séquence y conditionnellement à x.

Définition : Soit *G= (V, E)*, où *V* est l'ensemble des sommets et *E* l'ensemble des arcs, un graphe non orienté et soient *X* et *Y* deux champs aléatoires décrivant respectivement l'ensemble des étiquettes, de sorte que pour chaque nœud *i* appartenant à *V*, il existe une variable aléatoire y_i dans *Y*. Nous désignons *(X, Y)* comme étant un champ aléatoire conditionnel si chaque variable aléatoire Y_i respecte la propriété de Markov suivante : $p(Yi|X, Yj\ i \neq j) = p(Yi|X, Yj, i \sim j)$, où $i \sim j$ signifie que *i* et *j* sont voisins dans *G.* La figure 1 présente un exemple d'un graphe de CACs.

Cette propriété n'est par conséquent satisfaite que si chaque variable aléatoire ne dépend que de ses voisins : Y_i ne dépend que de *X* et des Y_j ses voisins dans le graphe d'indépendance.

D'après le théorème de Hammersely-Clifford (Hammersly et *al*., 1971), la distribution de probabilité *p* d'un champ de Markov est décomposable comme un produit de fonctions φ_c définies sur cliques, sous graphes complets, maximales c de l'ensemble des cliques C de G. Ainsi, la probabilité d'un étiquetage y étant donnée une réalisation d'observations *x* s'écrit :

$$p(y|x) = \frac{1}{Z(x)} \prod_{c \in C} \varphi_c(y_c, x)$$

Où y_c est la réalisation des variables aléatoires de la clique *c* et *Z(x)* est un coefficient de normalisation défini comme suit :

$$Z(x) = \sum_{y} \prod_{c \in C} \varphi_c(y_c, x)$$

Le coefficient *Z(x)* est un coefficient de normalisation égal au produit des fonctions de potentiel de tous les étiquetages possibles sachant la séquence d'observation *x*.

Lafferty et ses co-auteurs (Lafferty et *al*., 2001) ont proposé de définir la forme de la fonction φ_c comme l'exponentiel de sommes pondérées des fonctions caractéristiques f_k ayant des poids w_k.

$$\varphi_c(y_c, x, W) = \exp(\sum_{k=1}^{K} w_k f_k(y_c, x))$$

La forme de ces fonctions dépend du domaine d'application. Par exemple, dans le TAL, il s'agit généralement de fonctions binaires qui testent la présence ou l'absence de certaines caractéristiques. Concernant les poids w_k, ils permettent d'accorder plus ou moins d'importance à chacune des fonctions caractéristiques. Ils sont fixés lors de la phase d'apprentissage en cherchant à maximiser la log-vraisemblance sur un ensemble d'exemples déjà annotés formant le corpus de référence. La probabilité d'un étiquetage sachant une réalisation d'observations s'exprime ainsi par :

$$p(y|x) = \frac{1}{Z(x, W)} \exp(\sum_{c \in C} \sum_{k=1}^{K} w_k f_k(y_c, x))$$

Les CACs sont appliquées à de nombreuses tâches du TAL, à titre indicatif l'analyse syntaxique partielle (Sha, Pereira, 2003), l'extraction d'informations à partir des tables (Pinto et *al*., 2003), la reconnaissance d'entités nommées (Li & McCallum, 2003 ; Benajiba et *al*., 2010) et l'étiquetage morphosyntaxique (Outahajala et *al*., 2011b). Les CACs ont été utilisés pour de nombreuses langues pour l'étiquetage morphosyntaxique, tel que l'amharique (Adafre, 2005), le tamoul (Lakshmana & Geetha, 2009), etc.

Dans les expérimentations présentées dans la section suivante, nous avons utilisé l'outil CRF++[5], une implémentation open source des CACs pour la segmentation et l'étiquetage des données.

4. Expérimentations et résultats

Dans cette section, nous présentons une description du corpus brut ainsi que son prétraitement, ensuite les modèles de référence et enfin les expérimentations relatives à l'utilisation de la mesure de confiance, le choix des données aléatoirement pour un apprentissage semi-supèrvisé et l'apprentissage de notre étiqueteur morphosyntxique.

4.1. Description du corpus brut utilisé

Le corpus utilisé dans ces expérimentations a été puisé dans quelques romans amazighes, une partie des données collectées par le Linguistic Data Consortium en collaboration avec l'IRCAM (Cieri et Liberman, 2008), textes brut des sites web de l'IRCAM[6] et de l'Agence Marocaine de Presse[7]ainsi que certaines phrases traduites

[5] http://crfpp.sourceforge.net/

[6] http://www.ircam.ma/amz/index.php

[7] http://www.mapamazighe.ma/am/

en amazighe de divers sources. Le corpus collecté a subi de multiples prétraitements, à savoir :

- révision des textes collectés selon les règles orthographiques adoptées par l'IRCAM. Aussi, la correction de certaines erreurs fréquentes telles que le mauvais placement du e muet "□". Dans ce sens, un script écrit en PERL a été réalisé afin de fixer cette erreur. En effet, l'utilisation du e muet s'impose dans les deux cas suivants :

- Succession de plus de trois consonnes radicales identiques à l'intérieur du même mot, par exemple □□□□□ (zmmem) "inscrire", □□□□□ (tettu) "elle a oublié" ;

- Radicaux verbaux se terminant par deux consonnes identiques, par exemple □□□□(mlel) "être blanc",

- Pour les textes rédigés en utilisant la police Tifinaghe-IRCAM (Tifinaghe-IRCAM fait usage de glyphes tifinaghes mais caractères latins), afin de corriger certains éléments comme le caractère "^" qui existe dans certains textes dû à une erreur en saisissant les lettres emphatiques ;

- Translittération des textes écrits en Tifinaghe-IRCAM et des textes écrits en utilisant la transcription officielle tifinaghe de la langue amazighe, vers le système d'écriture choisi ;

- Segmentation, en utilisant le segmenteur amazighe réalisé pour cet effet (Outahajala et *al*. 2013) ;

Le nombre total des morphèmes à partir du corpus recueilli est d'environ un quart de million.

4.2. Modèles de références

En ce qui concerne les modèles de références utilisés, nous avons choisi d'adopter deux lignes de base comme références dans ces expériences. En outre, nous avons utilisé le dernier jeu d'étiquettes disponible, composé de 28 étiquettes (Outahajala et *al*., 2013), et les CACs comme modèles de classification des séquences pour les générations des modèles de classification. Un jeu d'étiquettes de taille presque similaire a été utilisé pour l'étiquetage morphosyntaxique de l'arabe (Diab et *al*., 2004). Les modèles de références utilisés comme lignes de base dans les éxpérimentations des sous sections 4.3 et 4.4 sont :

1 – Modèle de référence basé sur la fréquence des mots (Freq-Base.) : il s'agit d'un algorithme basé sur la fréquence des étiquettes des mots. L'étiquette prévue pour un mot est tout simplement l'étiquette la plus fréquente qui a été associée dans les données de formation. Ainsi, cette base ignore totalement le contexte environnant et résout les cas ambigus utilisant uniquement les fréquences des étiquettes. Une telle référence a été utilisée dans la tâche de reconnaissance d'entités nommées

dans CoNLL. Le code source de ce modèle basé sur les fréquences est librement disponible[8].

2 – Modèle de référence du meilleur cas (Best-Base.) : pour étudier le meilleur des cas, on a commencé par la génération d'un modèle initial M_{init} à partir de 60% des données étiquetées. Les 30% des données étiquetées restantes on été subdivisées en blocks de 2k jetons. Ceci, dans le but d'étudier la performance des modèles générés à partir des données annotées automatiquement. Le choix des données pour la génération de M_{init} n'est pas aléatoire. En effet, nous avons effectué la validation croisée de 60% du corpus et nous avons pris le modèle qui a donné la meilleure précision.

Le choix de l'ensemble des caractéristiques a été obtenu suite à des résultats empiriques. Ils sont les mêmes que ceux employés dans (Outahajala et *al*., 2012) à savoir :

1. Le jeton actuel ;
2. Les propriétés lexicales n-grammes : consistant en les i premiers et dernier n-grammes du jeton, avec i variant de 1 à 4. Les caractéristiques n-grammes servent comme caractéristiques représentant les suffixes et les préfixes des jetons ;
3. Le contexte lexical : s'agissant des jetons voisinant plus leurs propriétés n-grammes définies dans le point 2 ci-dessus ;
4. Etiquettes de contexte qui consistent en les balises prévues pour les deux mots précédents.

4.3. Expérimentation : choix des données selon la mesure de confiance

Le but de ces expérimentations préliminaires est d'évaluer le critère de confiance dans la sélection des phrases pour l'auto-apprentissage de notre modèle. Nous avons part de l'hypothèse que notre modèle apprend plus quand la confiance est élevée. Pour évaluer notre approche, nous commenceons par un modèle initial M_{init} entraîné sur la base de Tr_1 (voir Figure 2), contenant l'équivalent de 60% du corpus de référence.

Pour ce faire, nous étudierons la corrélation entre la mesure de confiance et la probabilité d'obtenir un étiquetage correct. C'est l'estimation des chances d'assigner une étiquette correcte à un mot automatiquement quand la probabilité de l'étiquette affectée au mot par le système est élevée. Nous pensons que cette estimation est importante car lorsque la corrélation observée tend vers 1, la probabilité des données sélectionnées tend à améliorer le système et, lorsque cette probabilité tend vers 0,5, l'amélioration est aléatoire. D'un point de vue de filtrage du bruit, on peut dire que dans le cas d'absence de corrélation entre les deux termes en question, il n'est pas possible de filtrer le bruit en se basant sur la mesure de confiance générée par le système.

[8] http://www.outamed.com/downloads/baseline.txt

Afin d'obtenir l'information requise, nous avons automatiquement annoté 10% du corpus de test (nous n'avons intentionnellement pas utilisé le corpus de test lors du du calcul de la corrélation) utilisant M_{init}. Les étiquettes obtenues ont servi comme données de base dans le calcul de la corrélation.

La corrélation entre la mesure de confiance et la probabilité d'avoir un étiquetage correct est 0,78. On a ainsi une nette régression positive.

Pour étudier l'utilité de la mesure de confiance du système pour les mots dans la sélection des données, nous avons effectué des expérimentations utilisant M_{init} et les données brutes présentées dans la sous-section 4.1. Les données non étiquetées ont été annotées automatiquement et nous avons gardé les meilleures : 1295 phrases, soit l'équivalent de 90% des données annotées manuellement, selon la mesure de confiance.

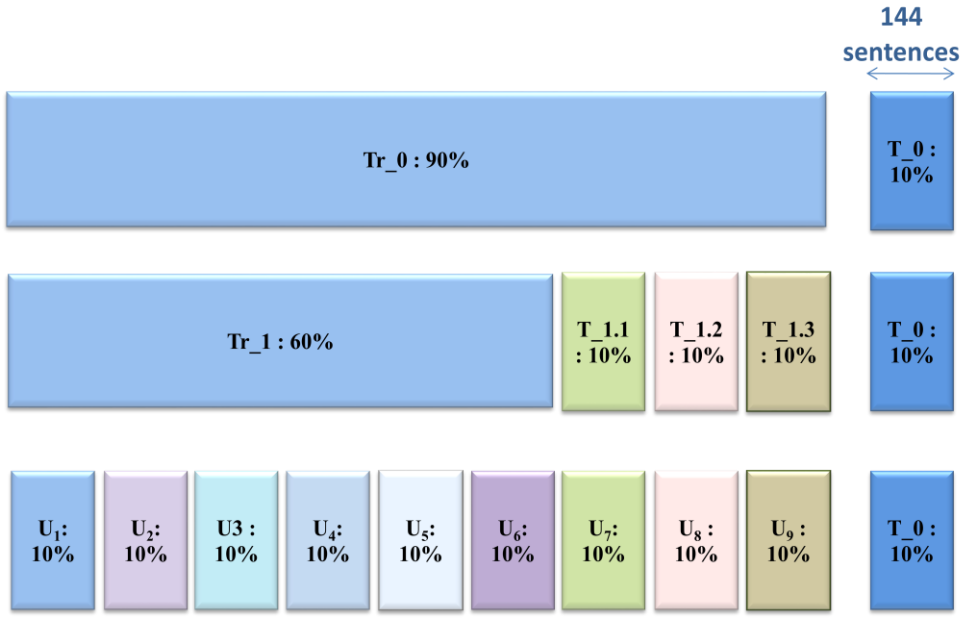


Figure 2 : Subdivision des données pour les expérimentations sur l'apprentissage semi-supervisé

Dans cette expérimentation, le critère de sélection est basé sur la mesure de la confiance donnée par le système. Après, ce corpus a été subdivisé en 9 parties U_1, U_2, U_3, U_4, U_5, U_6, U_7, U_8, et U_9, où chacune des parties U_i contient 144 phrases avec i variant de 1 à 9 (l'équivalent de 10% du nombre total des phrases du corpus annoté manuellement). La subdivision du corpus est présentée dans la figure 2.

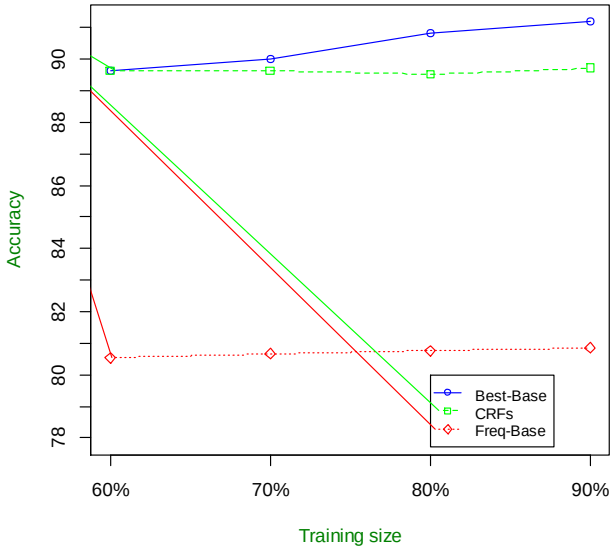


Figure 3 : Apprentissage du modèle en utilisant la mesure de confiance du mot comme moyen de sélection

Nous avons remarqué qu'au fur et à mesure que la performance augmente, elle devient difficile à améliorer. Néanmoins, la différence d'amélioration ne diminue pas de façon régulière, elle fluctue légèrement. Par exemple, le taux d'amélioration entre 70 et 80% (0.81) est supérieur au taux d'amélioration entre 60 et 70% (0.66) lorsqu'on fait l'entraînement des modèles à l'aide des données annotées manuellement. A l'analyse des fichiers en sortie de l'étiqueteur, il s'avère que les mots hors vocabulaire constituent un facteur important dans l'amélioration de la précision de l'étiqueteur. Aussi, la performance des modèles basés sur les CACs est-elle nettement supérieure à celle du modèle à base des fréquences.

Pour ce qui est des résultats du modèle utilisant et les données du corpus de référence et les données brutes, l'amélioration est légère. Les résultats de l'expérimentation montrent qu'il y a une réduction du taux d'erreur de 1,3% (voir figure 2).

4.4. Expérimentation : choix aléatoire des données pour l'apprentissage

Pour étudier l'effet d'ignorer la confiance et voir si ce critère est important ou non, nous avons conduit une expérimentation où nous commenceons par M_{init} et à chaque itération de l'opération d'apprentissage nous ajoutons 144 phrases de U annotées automatiquement par M_{init} et choisies aléatoirement.

Tels que montrés dans la figure 4, les résultats de l'apprentissage à partir des données choisies aléatoirement sont moins précis que ceux qui se basent sur la sélection des données en utilisant la mesure de la confiance. Ceci confirme l'utilité

de cette mesure dans la sélection des phrases dans l'auto-apprentissage de notre étiqueteur morphosyntaxique.

Dans la figure 4, CRFs-R représente le modèle généré à partir des données sélectionnées aléatoirement et CRF-BS le modèle généré en utilisant la mesure de confiance du mot comme moyen de sélection des données pour l'apprentissage.

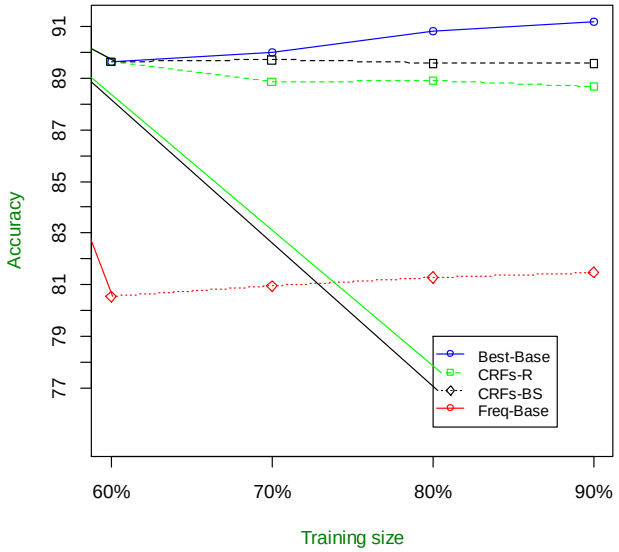


Figure 4 : Apprentissage à partir de données sélectionnées aléatoirement

Afin de vérifier l'hypothèse que le bruit de l'auto-apprentissage n'empêche pas la réduction du taux d'erreur lors de l'entraînement de notre modèle, nous avons conduit l'expérimentation suivante :

- génération de M_{init} à partir des parties U_1, U_2,...U_6 constituant 60% de la taille du corpus de référence ;

- ajout à chaque itération de l'apprentissage de 144 phrases au corpus d'apprentissage jusqu'à ce que le corpus d'apprentissage atteigne l'équivalent de 90% du corpus de référence.

Les résultats de l'expérimentation montrent qu'il y a une réduction du taux d'erreur de 5,9% entre M_{init} et le modèle appris en utilisant U_1, U_2,…U_9. Ce qui montre que, même si le bruit existe, le système continue d'apprendre.

5. Conclusions

La langue amazighe, comme la plupart des langues de moindre diffusion, souffre encore de la pénurie d'outils et des ressources pour son traitement automatique, en particulier les corpus annotés. Dans ce papier, nous avons présenté les expérimentations préliminaires d'utilisation de ressources externes, consistant en un corpus de textes bruts de 225.240 morphèmes et d'un corpus manuellement

annoté d'environ 20k morphèmes et leur impact sur la performance de la tâche d'étiquetage morphosyntaxique de la langue amazighe.

Les résultats des expérimentations montrent une réduction du taux d'erreur de 1,3%. Aussi la réduction du taux d'erreur est-elle de 5,9%, lorsque le modèle est complètement entraîné par les phrases du corpus brut annotées automatiquement.

Dans le futur proche, nous étudierons l'impact de l'utilisation du caractère informatif des MHVs et la mesure de confiance lors de l'utilisation des méthodes d'apprentissage semi-supervisé sur l'amélioration de la performance de l'étiqueteur morphosyntaxique.

Références

Ameur, M., Bouhjar, A., Boukhris, F., Boukouss, A., Boumalk, A., Elmedlaoui, M., Iazzi, E. et Souifi, H. (2004), *Initiation à la langue Amazighe*, Rabat, Publications de l'IRCAM.

Ameur, M., Bouhjar, A., Boukhris, F., Boukouss, A., Boumalk, A., Elmedlaoui, M. et Iazzi, E. (2006),*Graphie et orthographe de l'Amazighe*, Rabat, Publications de l'IRCAM.

Adafre, S. F. (2005). Part of Speech Tagging for Amharic using Conditional Random Fields.In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pp. 47-54.

Benajiba, Y., Diab M., and Rosso P. (2010). Arabic Named Entity Recognition: A Feature-Driven Study. *IEEE Transactions on Audio, Speech and Language Processing, vol. 15, num. 5.*Special Issue on Processing Morphologically Rich Languages, pp. 926-934. DOI: 10.1109/TASL.2009.2019927.

Boukhris, F. Boumalk, A. El moujahid, E. et Souifi, H. (2008), *La nouvelle grammaire de l'Amazighe*, Rabat, Publications de l'IRCAM.

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics.In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, Menlo Park.AAAI Press/MIT Press.

Chafiq, M. (1991). *أربعة وأربعون درسا في الأمازيغية*. éd. Arabo-africaines.

Cieri, C., and Liberman, M. (2008). 15 Years of Language Resource Creation and Sharing. A Progress Report on LDC Activities. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, LREC'08, Marrakech.

Diab, M., Hacioglu, K., Jurafsky, D. (2004). Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. In *Proceedings of Human Language Technology-North American Association for Computational Linguistics (HLT-NAACL).*

Jurafsky, D., and Martin, J.H. (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, 2nd Ed. New Jersey: Prentice Hall.

Hammersley, J.M. et Clifford, P. (1971). Markov Fields on finite graphs and lattices. *Manuscrit non publié*.

Kudo, T., and Yuji Matsumoto, Y. (2000). Use of Support Vector Learning for Chunk Identification. In *Proceedings of CoNLL-2000 and LLL-2000*.

Lafferty, J. McCallum, A. and Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML-01*, pp. 282--289.

Lakshmana Pandian S., and Geetha, T. V. (2009).CRF Models for Tamil Part of Speech Tagging and Chunking. In *Proceeding ICCPOL '09*. Springer-Verlag Berlin, Heidelberg.

Li W., and McCallum, A. (2003). Rapid Development of Hindi Named Entity Recognition Using Conditional Random Fields and Feature Induction. In *ACM Transactions on Computational Logic*, pp 290--294.

Manning, C., And Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. *The MIT Press*.

McCallum, A., Freitag, D. and Pereira, F. (2000). Maximum entropy Markov models for information extraction and segmentation. *International Conference on Machine Learning*, pages 591—598.

Greene, B.B., and Rubin, G.M. (1971). Automatic Grammatical Tagging of English. Providence, R.I.: Department of Linguistics, Brown University.

Outahajala, M., Zenkouar, L., and Rosso, P. (2011a). Building an annotated corpus for Amazighe. In *Proceedings of 4th International Conference on Amazigh and ICT*. Rabat, Morocco.

Outahajala, M., Benajiba, Y., Rosso, P., and Zenkouar, L. (2011b). POS tagging in Amazigh using Support Vector Machines and Conditional Random Fields. In *Proceedings of 16th International Conference on Applications of Natural Language to Information Systems*, NLDB 2011, LNCS(6716), Springer-Verlag, pp, 238--241.

Outahajala, M., Benajiba, Y., Rosso, P. et Zenkouar, L. (2012), « L'étiquetage grammatical de l'amazighe en utilisant les propriétés n-grammes et un prétraitement de segmentation », *e-TI - la revue électronique des technologies d'information*, N° 6.

Outahajala, M., Zenkouar, L, Benajiba, Y., and Rosso, P. (2013). The Development of a Fine Grained Class Set for Amazigh POS Tagging. In *proceedings of ACS/IEEE 10th conference*. AICCSA 2013. Fes, Morocco.

Pinto, D., McCallum, A., Wei, X., and Croft, W. B. (2003). Table Extraction using conditional random fields. In *Proceedings of the 26th annual international of SIGIR'03*, pp. 235-242, New York, USA.

Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-Of-Speech Tagging. In *Proceedings of EMNLP*, Philadelphia, USA.

Schmid, H. (1999). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*. Academic Publishers, Dordrecht, 13--26.

Sha, F., and Pereira F. (2003). Shallow Parsing with Conditional Random Fields. In *Proceedings of Human Language Technology*.

Asinag-Asinag, 9, 2014, p. 105-118

Characterizing the Evolution of Arabic Learners' Texts: A Mostly Lexical Perspective

Violetta Cavalli-Sforza & Mariam El Mezouar

Al Akhawayn University

Nous étudions l'évolution d'une série de textes conçus pour les apprenants de la langue arabe, langue seconde, le long d'un cursus en considérant leur contenu lexical en termes de vocabulaire soi-disant acquis ou en cours d'acquisition par les apprenants auxquels sont destinés ces textes. Nous examinons aussi l'évolution d'autres variables de texte communément utilisés pour mesurer la lisibilité d'un texte. L'objectif est de déterminer les traits des textes qui peuvent être utilisés pour construire un modèle prédictif de la pertinence d'un texte à un apprenant, à un stade d'apprentissage donné, tel que défini principalement par le vocabulaire appris. Nous concluons en examinant si l'approche et les résultats peuvent être appliqués à l'amazighe.

Introduction and Motivation

Reading is one of the four fundamental competences that are targeted when learning a new language, the others being writing, listening, and speaking. The activity of reading serves multiple purposes, of which recognition and understanding of written words in context is a primary one, as it aids in the development of a rich vocabulary and mastery of the nuances of its use. However, a learner cannot read just any text: when creating or choosing a text for language learners, an instructor must consider different goals and constraints. The text must aid in practicing newly learned language concepts—vocabulary and grammar, among others—while at the same time being sufficiently accessible to the learner by containing enough familiar concepts and covering, ideally, an interesting topic. Although the topic of the text can indeed be a motivating or demotivating factor for a learner, except in the case of fully independent and/or rather advanced learners, the choice of topic is usually determined by the instructor in view of the goal of developing specific vocabulary competence. On the other hand, the constraint of "containing enough familiar concepts" cannot be overlooked. In language learning, as in other areas of learning, what is known provides the framework for comprehending and anchoring novelty. A text that contains too many new terms and unfamiliar structures can only be comprehended with great effort, if at all, and will mostly serve to frustrate and demotivate all but the most tenacious of learners.

Novel language concepts are only one aspect of text difficulty; another is the complexity of the text due to the style in which it is written. A text may contain familiar concepts, but they may be expressed in a complex fashion, for example, using difficult words and intricately structured sentences whose relationship to each other is not clearly signaled. Such a text may also require more cognitive effort for successful processing and may lead to frustration and/or failure of comprehension. The problem of readability of a text was addressed initially in the context of learning to read in one's first language, driven by the need to create accessible schoolbooks. Since not all students in a given school grade have the same reading skills, and in some case are weak readers, the complexity of a text can interfere with the learning of subject matter content and work against students success across different subjects. More recently, with extensive information about topics of public interest (e.g., medical conditions and care) becoming available to everybody on the internet, the concern with producing generally and easily readable materials has spread to a much wider range of topics and audiences. Correspondingly, as more text material becomes accessible in electronic format, and as natural language processing (NLP) technology advances, researchers have started turning their attention to the use of computational tools to help evaluate texts for readability and their appropriateness for learners at different levels.

Pioneering efforts in the use of NLP technology for assessing text readability and appropriateness to a learner's level focused on English, followed by other European and oriental languages. Very little work has been done to date on Arabic and, not surprisingly, none on Amazigh, as far as we know. Yet, these two languages are of interest for Morocco today and one characteristic they share is that, in the context of reading, they can be thought of as combining aspects of first and second language learning. Modern Standard Arabic (MSA or فصحة) is no one's mother tongue. Students in Arabic-speaking countries learn MSA at school and, while undoubtedly aided by the dialect in some respects, they are still learning a different and more complex language. The place of MSA as a second language is more evident for individuals who are born in Amazigh-speaking households, and the difficulty of learning to read in MSA is even more acute when considering literacy programs for adults who need to develop character and word-decoding skills for a language that is not entirely the one they usually speak (Maamouri, 2005). Learning to read Amazigh (whether using Tifinagh or the Latin alphabet) poses similar challenges: it is clearly a second language for Moroccans coming from Arabic-speaking households, and the many variants of Amazigh spoken in Morocco alone make the standard form of the Amazigh language used in written educational materials something of a second language even for native speakers.

In view of the dual first-language/second-language position held by a language such as MSA, the work we describe in this paper draws on research performed in both first and second language contexts, and it examines both absolute text readability measures and the appropriateness of a text to a learning stage. Our research focused on MSA, so some of its conclusions are specifically relevant to that language and pertain to linguistic features that are not necessarily present in the Amazigh language. Nonetheless, the general approach remains valid and the results obtained to date for Arabic can inform similar work on Amazigh, thus providing further encouragement to develop NLP technology for this language.

The work reported herein should be viewed as an intriguing but shallow excursion into largely unexplored territory. We are fully aware that there are several aspects of text appropriateness and complexity could not be investigated due to time and resource constraints. We are confident, however, that the work performed and the results obtained to date provide a sound basis for future work.

Readability of a Text vs. Appropriateness of a Text to a Learner

Readability assessment is defined in the literature as the estimation of how 'difficult' a piece of writing is. The assessment is based on the characteristic of the text itself and so is independent of the learner. On the other hand, determining whether a text is appropriate for a specific learner or class of learners requires characterizing the learner(s) and the text in terms of common features and defining criteria that relate the two. In this section we begin by describing a few of the most common and well known readability assessment approaches and measures—which number in the hundreds—and how they have been applied to Arabic. We then examine how a text and a student can generally be characterized in order to establish a way of measuring the appropriateness of the former to the latter. We review some important previous work performed in that area before proceeding to describing our own study.

Measuring Readability

Readability of a text can be defined, as well as measured, in different ways. In 1963, Klare defined it as "a measure of the ease of understanding due the style of writing" (DuBay, 2004); this is a sweeping definition that focuses on the text itself while not precisely referring to any of its specific linguistic or stylistic features. McLaughlin, in 1969, emphasized the relationship between the material and the reader, defining readability as "the degree to which a given class of people finds certain reading matter compelling and comprehensible". A comprehensive definition of readability, which also relates the reader to the text, is given by Dale & Chall (DuBay, 2004) as: "The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting." The complexity and comprehensiveness of these definitions, however, is not reflected in the more quantitative measurement methods developed in the educational literature.

How is readability measured? A widespread method is cloze tests, in which a reader must fill in missing words in the text. Human judgments of readability are also used. However, the most common approaches to measuring readability are formulas that compute a numeric score from some text characteristics, which vary based on the formula. The Fleish-Kincaid Reading Ease Score (FRES), for example, is a linear combination of average word length (measured in syllables per word), and average sentence length (measured in words per sentence) (DuBay, 2004). The Dale-Chall formula uses average sentence length and the ratio of "difficult" words to total words; difficult words are those not present on a list of words expected to be known by a fourth grader in the American school system

(DuBay, 2004). The SMOG Formula uses a somewhat more complex algebraic formulation involving the number of polysyllable words (words with more than 3 syllables) and the number of sentences (Mc Laughlin, 1969). The values obtained from these readability formulas are used to link a text to a grade level, either directly, as in the case of the Flesh-Kincaid Grade Level formula (a variation on the FRES) or the SMOG formula, or by going through a table, as in the case of the Dale-Chall formula. For example, a score of 4.9 and below for Dale-Chall indicates that a text is appropriate for the 4th grade and below, whereas a score of 9 to 9.9 places a text at the college level and a score of 10 and above is for texts suitable for graduate school. These formulas have varying degrees of accuracy but do not typically transfer to languages other than English.

Outside of English, work on readability formulas has been carried out for European languages, such as French, Spanish, Swedish and Danish (Al-Tamimi *et al*., 2013), and for Japanese (Tateisi *et al*., 1988), among others. In addition to the readability formula approach to readability assessment, other approaches are found in the literature for different languages. For example, for Chinese, Pang (2008) describes a method based on Support Vector Machine for regression problems, where key text features are selected and used to predict readability. A study of Hebrew, a Semitic language closely related to Arabic, examined the correlation of 50 features (lexical, semantic, morphological, statistical and syntactic characteristics of the texts) and the difficulty score (1-easiest to 10-hardest) of a set of 70 texts, as rated by language experts (Ben-Simon and Cohen, 2011).

Measuring Readability for Arabic

For Arabic, two formulas were found in the literature: the Dawood Formula (Dawood, 1977), which includes five text features (average word length, average sentence length, word frequency, percentage of nominal clauses, and percentage number of frequency of definite nouns) and the Al Heeti formula (Al-Ajlan *et al.*, 2008), which takes into consideration only the average word length. The selection of these specific features has not been thoroughly justified in the literature, nor do the formulas achieve good results.

We also found two studies in which Saudi researchers used machine learning techniques to learn features for mapping texts to grade levels. In the first of these studies, researchers trained a Support Vector Machine classifier to assign texts, hand-selected from the Saudi school curriculum, to three difficulty levels: easy, medium and hard (Al-Khalifa and Al-Ajlan, 2010). The candidate text features chosen as input to the classifier for each text were: average sentence length, average word length, average number of syllables, word frequencies and perplexity scores for bigram language model. The trained classifier was then tested against unseen texts and expert ratings. The authors concluded that average sentence length was the best single feature in determining Arabic text readability, with a 66.67% accuracy rate; the best combination of features was average sentence length, statistical language model, and term frequency. However, while the prediction accuracy of the model was excellent on easy texts (100%), it dropped to 70% for hard texts and did not achieve any good results on medium texts. The authors attributed the poor results obtained for the latter category to some confounding

factors in the texts themselves. As in previous work (Ajlan *et al.*, 2008), they questioned the validity of the assumption that texts from the Saudi curriculum are correctly distributed among the three difficulty levels, and argued that, to be able to give more accurate predictions from such a system, there is a need for an Arabic corpus correctly labeled with readability levels.

A research effort along similar lines aimed at determining the factors that affect readability of an Arabic text and its mapping to the 10 grades of the Jordanian school curriculum (Al-Tamimi *et al*., 2013). The study was conducted using factor analysis on features that included word length, word frequency, vocabulary load, number of difficult words, average sentence length, sentence complexity, the clarity of the text idea, the use of topology or metaphors, and the grammatical structure complexity. The features were later grouped to remove some redundancy using principal component analysis to determine the most salient features. These were in turn used to create the AARI Base formula, which is then used in another formula to map a text to a grade level. The best performance (with accuracy of over 83%) was obtained when assigning a text to one of three clusters (1st to 3rd grade, 4th to 5th grade, and 6th to 10th grade). The accuracy dropped to under 50% when assigning to individual grades.

A third study took a much simpler approach to assessing the difficulty of texts (Daud *et al*., 2013). It was based on summing the score of words and dividing it by the number of words in the text. The score of individual words is drawn from the frequency of the words in the King Abdulaziz City for Science and Technology Arabic Corpus (KACSTAC) corpus, a general corpus whose texts are derived from magazines, books, newspapers, referred journals, dissertations, government circulation, school curriculums, newswire and the Internet. The authors assume that the more frequent a word is, the easier it is, and so the word score is its reversed ranking in the corpus. Consequently, the lower the overall score of the text, the easier the text is assumed to be. Unfortunately this work appeared to be rather preliminary and did not provide any specific conclusions about the effectiveness of the readability estimation.

Measuring Appropriateness of a Text to a Learner

The readability indices described in the previous section either assign an absolute readability score to a text or relate the text to a grade level in a school curriculum. They presuppose that the grade level is a sufficient, if vague, indicator of an average level of reading skill expected of students in that grade. The actual skills that are expected of a student are characterized, in general, by the learning objectives at each grade level. This characterization is generally taken to be adequate for a gross mapping of texts to curriculum level and for the population of students at that level. It does not really permit any finer level analysis of what the student actually knows, nor what learning is afforded by a specific reading material, and it does not support any adaptation of text selection to the needs of the individual student. In contrast, with advances in artificial intelligence techniques, user-adaptive tools and interfaces are now possible and have become the hallmarks of effective man-machine interaction, whether applied to educational experiences in intelligent tutoring systems or to general information access over the internet.

The development of a more user-adapted interaction between the human user and the machine involves collecting data on users and then selecting from a range of possible information to offer to each user, or group of users, materials or services that seems to best fit their needs at a given point in time. It was with this perspective in mind that the REAP project (http://reap.cs.cmu.edu) developed an approach to text selection addressed at characterizing the student's knowledge and interests, the contents of a text, and the contents of a learning curriculum in such a way as to allow a reading practice environment to select a text to support reader-specific lexical practice (Brown and Eskenazi, 2004). REAP focused on supporting learning English as a second language, but many of its core ideas carry over to learning to read complex texts in one's first language and in literacy training as well. The initial focus on English and vocabulary learning was later extended to cover some grammatical concepts and versions of the system were built for French and European Portuguese.

According to the initial REAP approach, the selection of an appropriate text depends on the vocabulary the student already knows (the *student model*), the vocabulary the student is trying to learn within a curriculum of study (the *curriculum model*), and the vocabulary content of a text (the *text model*). In a nutshell, a text supports reader-specific lexical practice if it contains at least some of the words targeted for learning at a particular curriculum stage, while also containing a sufficient number of known words—and correspondingly a sufficiently small number of unknown words—so as to make the text accessible for the learner. It is this view of what makes a text appropriate or adapted to a learner that underlies the research that we conducted for Arabic and describe below.

Description of the Study

Goals and Approach of the Study

The ultimate goal of our research was to develop a characterization of what makes a text suitable for a learner studying MSA as a second language, so as to be able to predict the suitability of a text to a given *skill stage* and, eventually, a specific learner and dynamically select texts to enrich available learning materials. To our knowledge, there are no freely available corpora of texts for Arabic learners tagged by level of difficulty. So, we used as an implicit measure of increasing difficulty the curriculum presented by part of two volumes of the commonly used Al-Kitaab textbook series (Brustad *et al.*, 2004 and 2007). The small collection of texts contained in this textbook series, each tied to a specific lesson with its vocabulary and grammar concepts, served as our basic data for characterizing the appropriateness of texts to a given skill stage. This characterization or model, once developed, would then serve to inform the selection of further texts to enrich the supply of reading materials available for learners at a specific skill stage. Using the terminology adopted by project REAP, we framed the problem as follows:

- Chapters 1-20 of Al-Kitaab Volume 1 and Chapters 1-5 of Al-Kitaab Volume 2 formed the 25 skill stages of our *curriculum model*.
- Each skill stage had an associated set of vocabulary items explicitly targeted by the chapter as lexical items the student should learn.
- The lexical items presented at each skill stage cumulatively represent the content of a *student model* for the perfect student—the one who does not forget anything—who has reached that skill stage.
- At any one stage in the curriculum, the perfect student is expected to have learned and be able to recognize all words in previous skill stages (the *known* words), is in the process of learning new words for the current skill stage (the *target* words), and is not expected to know any of the words presented in later skill stages (*unknown* words).
- The words contained in the texts and their classification as *known*, *target*, and *unknown* words for a given skill stage are elements of the *text model*, to which we later added other readability variables.

The research examined, qualitatively and quantitatively, trends in the proportions of known, target and unknown words in texts through the 25 skill stages. These variables and some of the standard readability measures were used to train a model to predict which stage an unseen text was appropriate for.

Data Used in the Study

In addition to the Al-Kitaab texts and vocabulary lists that served as the training data for the model, three additional sets of texts were used:

- A set of 23 texts, created or chosen by an Arabic language instructor familiar with the Al-Kitaab book to match specific chapters/skill stages; a few of these were transcription of online videos. These labeled texts were used to test the accuracy of the predictive model.
- A set of 10 texts, hand selected from the Syrian school curriculum (Syrian, 2013) spanning grade levels from primary to high school.
- A set of 10 texts manually collected from the internet from the online newspaper Hespress (http://hespress.com/).

These additional texts, collected using the results of the analyses performed on the Al-Kitaab texts, were intended to be used for prediction from a model. The model was still being built at the time of writing.

Tools and Processing Used in the Study

None of the texts used in the study had any associated information to help identify the words contained in the text. As a result, each text was initially processed by running it through MADA (Habash *et al.*, 2009), which performs several

operations—tokenization, diacritization, morphological disambiguation, part-of-speech tagging, stemming and lemmatization—at the same time. It provides as output a list of ranked analyses for each token in its input. MADA's analyses were aggregated by Buckwalter lemma-ID, a feature that conveys the sense of the word, thereby ignoring analyses that differed only in inflectional features or in the presence of different clitics. MADA is trained on an extensive corpus of texts, of which the Al-Kitaab texts or the other texts we used are not necessarily representative. Evaluation of MADA by its authors on the same type of data on which it was trained shows that MADA's first analysis picks the right part of speech and the right lemma-ID over 96% of the time. Manual examination of MADA's output on 100 randomly selected words in the Al-Kitaab texts showed that MADA was able to pick the right word sense over 94% of the times. Additional analyses we performed using the second MADA result (after aggregation by lemma-ID), shows similar patterns to the first analysis. It was therefore decided that it was safe to base further work on the first MADA analysis.

A MADA analysis was picked for each word in the text and classified as a *known*, *target*, or *unknown* word from the perspective of the perfect student model. A word may have been introduced at different times and/or with different information or expected *competence* in the vocabulary curriculum. For example, the learner may have seen the word in a list of vocabulary targeted by a specific skill stage; if so, this word is labeled as having *high competence*, since the (perfect) learner will supposedly have acquired it before moving on to the next skill stage and will not forget it thereafter. Alternatively, the learner may have been exposed to a word without actually being expected to learn it; for example, the word may have been used in an example with an inline gloss, or in a previous text with no gloss. In this case the learner would not be expected to display high competence. After a number of exploratory analyses, it was decided to use the following procedure to classify words. Let **t** stand for the skill stage at which a text is introduced and **d** stand for the first high-competence appearance of a word in the vocabulary curriculum. At a given skill stage, a word is considered:

Known if **d** is less than **t**

- *Targeted* if **d** is equal to **t**
- *Unknown if* **d** is greater than **t**

Punctuation, number and non-Arabic word tokens were identified and excluded from further analyses. Additional details regarding the extensive data processing and sensitivity analyses performed are described elsewhere (El Mezouar, 2013).

Results of the Study

The major results of the study concern trends in *known*, *targeted*, and *unknown* words in texts and correlation of other common readability variables with skill stages. Progress has also been made towards the construction of a predictive model to be able to assign texts to specific skill stages, but the results obtained, though encouraging, are still preliminary, so this aspect requires further investigation.

Trends in of Known, Targeted, and Unknown Words in Texts

Figure 1 below shows the evolution in proportions of *known*, *targeted*, and *unknown* words in texts across the 25 skill stages. The figure shows the trends for tokens, but very similar patterns were obtained for types (unique word occurrences). While there is a certain amount of oscillation in all word types, there are also clear trends: the percentage of *known* words in texts steadily increases and the percentage of *unknown* words clearly decreases. Moreover, with few exceptions, the number of unknown word types is quite small, never going above 46 and usually below 20. These results confirm general expectations.

Past the first 3 skill stages, where a significant percentage of new vocabulary is used in short texts, *targeted* words hold steady within a range, 2-19% for tokens and 2-20% for types, with an average of 8% of targeted words for both tokens and types. It is an open question whether these values reflect an intentional choice by the textbook authors, who created or selected the texts or whether they represent an upper bound on how many new words can reasonably be targeted in a single text.

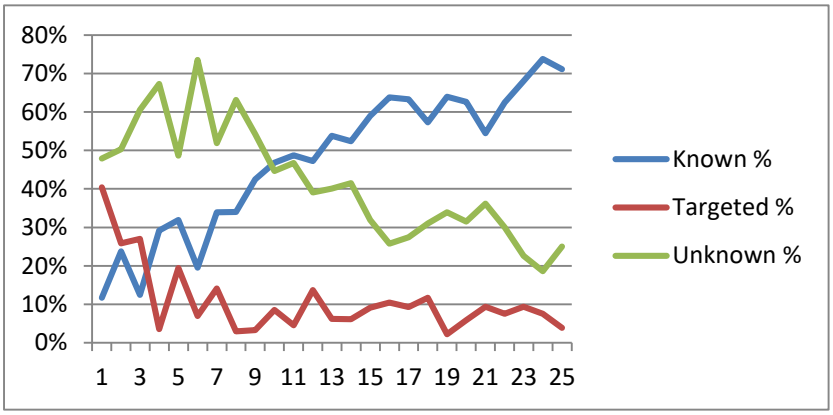


Figure 1 : Evolution of known, targeted and unknown variables for tokens

In addition, we investigated the vocabulary load across different skills stages and remarked there was a definite fluctuation, from low to high rates of new vocabulary introduction, with a periodicity varying between 2 and 4 skill stages. This indicates that there are points in the curriculum in which there is more emphasis on new vocabulary and others where the vocabulary is used and reinforced through practice.

Correlation of Common Readability Variables with Skill Stages

We considered some of the other measurements of text complexity that appear in common readability indices for Arabic and other languages to determine if any could be valuable indicators of the appropriateness of a text to a given skill stage. Table 1 summarizes some of the results obtained.

Particularly interesting is the negative result obtained for average length of words, a variable that appears, in some guise or other, in readability formulas for other languages, where it may measure number of characters or syllables. We note that word formation processes for Arabic, with 'word' defined as a series of characters bounded by spaces or punctuation symbols, differ significantly from those of most of the languages for which readability measures we reviewed were defined. Specifically, the Arabic word becomes longer on account of three processes:

Text feature	**How computed**	**Good determinant?**
Familiarity with vocabulary	Percentages of targeted, known and unknown words in a text	Yes
Open-class words	Percentage of open class words in a text	Yes
Closed-class words	Percentage of closed class words in a text	No
Lexical diversity	Ratio of unique words over total number of words in a text	No
Length of texts	Number of tokens in a text	Yes
Average length of words	Average number of characters per word in a text	No
Average complexity of words	Average number of clitics per word in a text	Yes
Average length of sentences	Average number of words per sentence in a text	Yes

Table 1: Summary of features determining appropriateness of text to level

1. **Derivational Processes**: intersection of the root with longer patterns containing more fixed letters. The different patterns are associated with different meanings and there are a fixed and small number of patterns and character additions, so the reader quickly develops skill in identifying these components of the word. Examples include the triliteral verb measures V and X, which add a 't' (ت) and an 'ist' (است) prefix, respectively. However the amount of word lengthening due to this process is rather limited.
2. **Inflectional Processes**: the addition of inflectional prefixes and suffixes, such as the markers for person and number in verbs and for gender, dual and plural in nouns and adjectives. Again, the amount of word lengthening due to this process is rather limited in both amount and variation.
3. **Cliticization**: The prefixation or suffixation of morphemes, including conjunctions, prepositions, pronouns and other particles, depending on the position. There are up to 4 proclitics (prefixed) and 1 enclitic (suffixed).

In light of the above word-formation processes, it is not surprising that, for Arabic, counting clitics gives more meaningful results than counting letters.

While we are encouraged by the positive results (features that are good determinants), we are also aware of the fact that some of the negative results may be due to the limits of the processing we performed, as described below.

Limitations of the Study

The initial motivation for this work came from a project targeted at developing Arabic reading enhancement tools (Maamouri *et al.*, 2012). That project was itself a response to some of the special difficulties involved in reading in MSA, given that certain aspects of its morphology and its writing system conspire to complicate significantly the recognition of a word in context. Firstly, as noted earlier, Arabic morphology allows the adjoining of several prefixes and suffixes to the basic word stem, so that identifying the stem involves stripping off affixed material, a process that suffers from some ambiguity. Secondly, the stem itself may be difficult to identify because of internal or boundary changes, such as those occurring in broken plurals or in words whose roots contain weak consonants. Thirdly, traditional Arabic dictionaries are organized by root, so looking up a word requires identifying the root letters and the pattern that give rise to the stem, a process that is also complicated by the presence of weak consonant radicals and assimilation processes. Finally, the absence of most diacritic signs in all texts, other than religious texts or texts used in early school years, further enhances the ambiguity of words and contributes to the difficulty in their identification. All these factors contribute to making word recognition in context quite difficult for the learner of MSA, since it requires extensive application of different sources of linguistic knowledge—lexical, morphological, syntactic, and semantic—in addition to general common sense knowledge and topic knowledge.

While the data we worked with contained some morphosyntactic information we could have used, our focus for this study was limited to words and, more particularly, to word senses (as captured by lemma-ID), treating words that share the same sense as equal, independently of the form they took on in context. In so doing, we ignored morphological processes giving rise to different inflectional and derivational variants. Similarly, we did not consider those closed-class words that appeared as clitics attached to an open-class word. We also labeled as unknown words whose meanings were not explicitly presented in the vocabulary but should or could have been guessed by the learner using acquired knowledge of morphological processes covered in the grammar. Investigating the effect of these and other omissions is left for future work.

Conclusion and Implications for Amazigh

We have described an exploratory study that investigated certain aspects of texts assumed to be relevant in determining the appropriateness of a text to a learner at a certain skill stage. We analyzed a text from two perspectives: 1) its fit to the lexical knowledge the learner might (or might not) have and the vocabulary targeted for

learning at a given stage in a curriculum; and 2) specific characteristics of the text, independent of the learner or of the instructional curriculum. The latter characteristics were chosen among the ones used in common readability measures for different languages. Most of the analysis focused on the lexical content of texts, and specifically the words that a learner was supposed to have acquired before reading a text and those contained in the text to aid the learner in practicing new vocabulary. We were able to find definite trends in these measures. We were also able to verify that some of the variables used to measure text complexity in well-known readability indices were informative as determinants of appropriateness to a learner's skill level, while others did not appear to be. However, it was also remarked that some of the negative results for some promising variables—for example, lexical diversity and proportion of closed-class words—may be due to the limited processing we performed and the morphosyntactic information it ignored. A fuller analysis that takes these characteristics into consideration could yield different answers. Indeed, other simple measures of morphological and syntactic complexity, such as average complexity of words—measured in number of clitics—and sentences—measured in average words per sentence—suggest that this information might be more useful than immediately apparent.

The ultimate goal of the analysis reported herein was to build a model that would allow us to predict whether a given text could be appropriate for a given skill stage, that is, would be accessible to a perfect learner who had learned all of the vocabulary presented up to that stage and would be adequate for practicing the use of lexical items introduced at that stage. The predictive model is still under development but preliminary results and feedback from instructors provide grounds for optimism. Further refinements and investigation are required to make the model truly useful and will be the focus of future work.

In general, it is a worthwhile goal to build such a model and use it as part of a language learning tool to provide criteria for (semi-)automatically selecting and/or modifying texts for learners, in order to dynamically enrich a database of learning materials. The general approach followed for Arabic in this study can be applied to other languages as well. Nonetheless, just as it has been shown that standard readability indices do not transfer directly between languages, we do not expect our analysis and the specific results we obtained to be applicable to all languages. With respect to Amazigh, we can expect that familiarity with vocabulary will be relevant to choosing learner-appropriate texts, as will open-class words, length of texts and average length of sentences. On the other hand, we have to withhold judgment on features such as average length of words and average complexity of words, since Amazigh morphology has some similarities to Semitic morphology but also many differences. Specifically, it is significantly less "agglutinating" than Semitic language morphology, even though Semitic languages are not themselves considered agglutinating languages (as opposed to, for example, Turkish). Concerning the significance of other features, such as the proportion of closed-class words or lexical diversity, more detailed analyses need to be performed even for Arabic, and it may well be that Amazigh will turn out to behave similarly to other languages for which these features do play a role in readability of texts. One aspect of morphology that we have not yet investigated may turn out to be important for text difficulty in both MSA and Amazigh: the change in context of

certain radical letters, e.g. the 'w/u' and 'y/i' sounds as well as other assimilation processes that are known to occur in both languages in the presence of adjacent consonants.

As a concluding remark, we point out that a study of the complexity and evolution of learners' texts along a curriculum, such as described above for MSA, is really enabled by the presence of digital text corpora (annotated or not), computational lexicons, part-of-speech taggers and morphological analyzers, among others. Therefore, the advancement of language instruction for both first and second language learners of Amazigh provides yet more motivation for investing in the development of such resources for this language.

Acknowledgements

This research was inspired and made possible by a previous project, housed at the Linguistic Data Consortium (LDC), and supported by the U.S. Department of Education under International Research Study (IRS) Grant No. P017A050040-07-05. We also thankfully acknowledge the support and generosity of the Linguistic Data Consortium for allowing us to use some of their linguistic resources. For further information regarding that project, contact Dr. Mohamed Maamouri (PI) at maamouri@ldc.upenn.edu.

References

Al-Ajlan, A. A., Al-Khalifa, H. S., and Al-Salman, A. S. (2008), « Towards the Development of an Automatic Readability Measurements for Arabic Language », *in Proceedings of the Third International Conference on Digital Information Management*, November 13-16, p. 506-511.

Al-Khalifa, H. S., and Al-Ajlan, A. A. (2010), « Automatic Readability Measurements of the Arabic Text: An Exploratory Study », *The Arabian Journal for Science and Engineering*, Volume 35, Number 2C, p. 103-124.

Al-Tamimi, A-K., Jaradat, M., Aljarrah N., and Ghanem, S. (2013), « AARI: Automatic Arabic Readability Index », *The International Arab Journal of Information Technology*, [Accepted March 12, 2013].

http://www.ccis2k.org/iajit/PDF/vol.11,no.4/5200.pdf, August 2013.

Ben-Simon, A. and Cohen, Y. (2011), *The Hebrew Language Project: Automated Essay Scoring & Readability Analysis*, International Atomic Energy Agency.

http://www.iaea.info/documents/paper_4e1237ae.pdf, August 2013.

Brown, J., and Eskenazi, M. (2004), « Retrieval of authentic documents for reader-specific lexical practice », *in Proceedings of InSTIL/ICALL Symposium*, June 17-19, Venice.

Brustad, K., Al-Batal, M., and Al-Tonsi A. (2004), *Al-Kitaab fii Ta^{c}allum al-^{c}Arabiyya, A Textbook for Beginning Arabic: Part One*, Second Edition, Washington D.C., Georgetown University Press.

Brustad, K., Al-Batal, M., and Al-Tonsi A. (2007), *Al-Kitaab fii Taᶜallum al-ᶜArabiyya, A Textbook for Beginning Arabic: Part Two*, Second Edition, Washington D.C., Georgetown University Press.

Daud, N. M., Hassan H., and Abdul Aziz, N. (2013), « A Corpus-Based Readability Formula for Estimate of Arabic Texts Reading Difficulty », *World Applied Sciences Journal*, Volume 21, p. 168-173.

DuBay, W. H. (2004). *The Principles of Readability*. Institute of Education Sciences, http://eric.ed.gov/?id=ED490073, August 2013.

Dawood, B. A-K. (1977). *The Relationship between Readability and Selected Language Variables*. Thesis submitted to the College of Education Board in the Universit of Baghdad in partial fulfillment of the requirements for the degree of Master of Arts in Education and Psychology.

http://dspace.ju.edu.jo/xmlui/bitstream/handle/123456789/12718/JUA0305740.pdf, August 2013.

El Mezouar, M. (2013), *Appropriateness of a Text for Learners of Arabic as a Foreign Language: A Word-Based Perspective*, Master Thesis Report submitted in partial fulfillment of the requirements for a Master of Science in Software Engineering at the School of Science and Engineering of Al Akhawayn University in Ifrane.

Habash, N., Rambow, O., and Roth, R. (2009), « MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization », *in Proceedings of the 2nd International Conference on Arabic Language Resources and Tool*s (MEDAR), April 22-23, Cairo, 2009.

Maamouri, M. (2005), « Arabic Literacy», *in Encyclopedia of Arabic Language and Linguistics*, Lemma 11,16, Volume 2, Brill.

Maamouri, M., Zaghouani, W., Cavalli-Sforza, V., Graff, D., and Ciul, M. (2012), « Developing ARET: An NLP-based Educational Tool Set for Arabic Reading Enhancement », *in Proceedings of The 7th Workshop on Innovative Use of NLP for Building Educational Application*s (NAACL-HLT-2012), June 7, Montreal, p. 127-135.

Mc Laughlin, H. G. (1969), « SMOG Grading - a New Readability Formula », *Journal of Reading*, Volume 12, Number 8, p. 639-646.

Pang, L. T. (2006), *Chinese Readability Analysis and its Applications on the Internet*, Thesis submitted in partial fullfilment of the requirements for the degree of Master of Philosophy in Computer Science and Engineering, The Chinese University of Hong Kong.

Syrian (2013), « المناهج الجديدة », http://me.syrianeducation.org.sy/ebook/classes.html, March 2013.

Tateisi, Y., Ono, Y., and Yamada, H. (1988), « A Computer Readability Formula of Japanese Texts for Machine Scoring », *in Proceedings of COLING 1988*, Volume 2, August 22-27, Budapest, p. 649-654.

Asinag-Asinag, 9, 2014, p. 119-132

Reconnaissance automatique de caractères et de textes amazighes : état des lieux et perspectives

Ali Rachidi

ENCG et Laboratoire IRF-SIC, Université Ibn Zohr, Agadir, Maroc

The design and implementation of systems OCR Amazigh character is very crucial for the promotion and development of the Amazigh language. To date, there is the lack of this type of system. Therefore, the automatic character recognition and text Amazigh has experienced in recent years a very significant interest in research work. Indeed, some systems have been developed to improve this situation. In this paper, we describe the different systems and approaches that were developed and tested in our laboratory to automatically recognize the Amazigh writing, showing the characteristics and results of each. This description will allow us to conduct a comprehensive summary of the various approaches and proposed systems that will help us to launch the outlook for future work.

1. Introduction

Dans le domaine de la reconnaissance automatique des caractères, plusieurs recherches scientifiques ont été effectuées sur le caractère latin, arabe, et autre. Ceci a permis le développement de plusieurs approches de reconnaissance automatique de ces caractères et, par conséquent, des lecteurs optiques pour scanner et reconnaître automatiquement les documents correspondants. Par contre, le caractère amazighe, appelé tifinaghe, est très peu traité. Des approches ont été proposées pour la reconnaissance de ce caractère. Ces approches sont regroupées généralement en grandes classes telles que les approches statistiques (Oulamara, 1988), (Djematen et *al*., 1997), les approches basées sur les réseaux de neurones (Ait Ouguengay, 2008), (El Yachi et *al*., 2009), (Bouikhalene et *al*., 2009) (Es Saady et *al*., 2011), l'approche syntaxique (Es-Saady et *al*., 2010), les Modèles de Markov cachés (Amrouch et *al*., 2010), (Amrouch et *al*., 2012) et l'approche basée sur la programmation dynamique (El Yachi et *al*., 2010). Dans cet article, on s'intéresse à présenter un état des lieux et une synthèse présentative et comparative des travaux de recherche scientifique effectués et publiés dans le domaine de la reconnaissance automatique de caractères amazighes imprimés et manuscrits, soit au niveau de notre laboratoire soit ailleurs.

Nous présentons en première partie les principales bases de données de caractères amazighes développées pour pouvoir tester et valider des approches. La seconde

partie est consacrée à la description des différents travaux effectués dans le domaine de la reconnaissance automatique de l'écriture amazighe. Nous présentons dans la troisième partie nos contributions dans ce domaine. Enfin, nous donnons une synthèse détaillée et comparative de ces systèmes tout en précisant les avantages et les limites de chaque système et en lançant un certain nombre de perspectives à développer dans les travaux futurs..

2. Principales bases de caractères amazighs développées

Pour pouvoir expérimenter et valider les différentes approches et systèmes développés, il est primordial de créer des bases de caractères amazighes. Des bases de données d'images de caractères amazighes annotées sont inexistantes. Ce domaine souffre ainsi l'absence d'une base de données de référence qui permet des comparaisons objectives entre les différents systèmes de reconnaissance. Tous les travaux publiés dans ce domaine ont été expérimentés sur des bases de données locales, qui contiennent un nombre restreint de l'alphabet amazighe.

Dans les deux sous-sections suivantes, nous décrivons les deux bases de caractères existantes.

2.1. Base des patterns de la graphie amazighe

C'est une base de patterns de différentes fontes amazighes et de tailles variées proposée par Ait Ouguengay (2009). Elle contient au total 12 polices de caractères et les tailles de 10 points à 28 points pour chaque modèle. Les patterns sont fournis sous forme d'images bitonales de tailles variables. La taille maximale est de 102×129 pixels, tandis que la taille minimale est de 19×2 pixels. Une telle disparité s'explique par le fait que le caractère ya (a) est un petit cercle, et est donc beaucoup plus petit que les autres caractères. Outre le cas particulier du caractère ya (a), la base est constituée des patterns de différentes fontes amazighes et de tailles variées qui ne sont pas normalisées. Le Tableau 1 donne une liste de quelques patterns dans cette base.

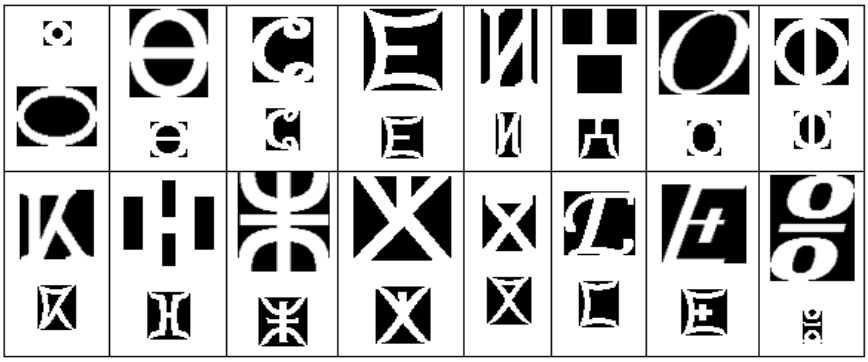


Tableau 1 : Exemples de quelques caractères dans la base des patterns de la graphie amazighe

Dans cette base, la manière dont sont créées les images des patterns ne permet pas la possibilité de renormaliser leur taille en une taille moyenne fixe. En effet, ceci peut être gênant en particulier à cause de la ressemblance des caractères ya (a) et yar (r) qui ne se différencient que par la taille : le caractère ya (a) est un petit cercle, tandis que le caractère yar (r) est un grand cercle. Dans certains cas, on aura une confusion réelle entre des images de ces deux classes.

2.2. Base de caractères manuscrits

C'est une base de données de caractères amazighes manuscrits (A Database for Amazigh Handwritten Character Recognition Research : AMHCD) que nous avons créée et développée au sien de notre Laboratoire IRF-SIC de l'Université Ibn Zohr d'Agadir. La base contient 25740 images de caractères amazighes manuscrits isolés et étiquetés, produites par 60 scripteurs de sexe, d'âge et de fonction différents. Le lecteur peut trouver une description complète et détaillée sur cette base dans (Es Saady et *al.*, 2011).

Jusqu'à présent, la base AMHCD est peu utilisée et explorée pour l'évaluation des systèmes de reconnaissance de l'écriture amazigheD. (2010(Amrouch et *al*., 2010) (Amrouch et *al*., 2012a) (Amrouch et *al*., 2012b). En revanche, elle s'est imposée comme la seule et la première base dans sa catégorie (graphie manuscrite amazighe), grâce à sa taille importante et à sa disponibilité pour les recherches académiques.

Le Tableau 2 présente des exemples des caractères amazighes manuscrits. Chaque caractère est donné sous forme de deux variantes qui correspondent aux deux scripteurs différents.

Caractères amazighs imprimés	Scripteur 1	Scripteur 2	Caractères amazighs Imprimés	Scripteur 1	Scripteur 2

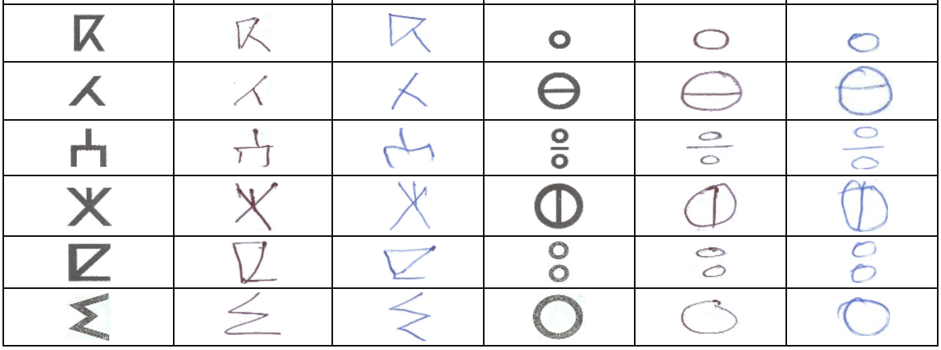


Tableau 2 : Exemples de caractères amazighs manuscrits issus de la base AMHCD (Essaady et al., 2011)

3. Approches de reconnaissance automatique existantes

En comparant au latin, à l'arabe ou au chinois, les recherches sur la reconnaissance automatique d'écriture amazighe n'ont pas atteint la perfection. Autant que nous le sachions, peu de tentatives ont été menées sur la reconnaissance d'écriture amazighe. Dans cette section, nous citons des travaux publiés qui traitent la reconnaissance de l'écriture amazighe.

Parmi les anciennes études qui portent sur la reconnaissance de caractères tifinaghs, on cite, en premier lieu, les travaux d'Oulamara J. (1988). La méthode proposée dans cette référence est une méthode statistique basée sur l'extraction de segments de droite par la transformée de Hough. L'analyse du caractère dans l'espace paramétrique, obtenu par la transformation de Hough, permet d'extraire les caractéristiques spécifiques en association avec un modèle de référence générateur de l'ensemble des caractères de l'alphabet. Un codage original est déduit puis utilisé comme base de construction de la matrice de lecture représentant une forme codée de l'alphabet. L'auteur a obtenu des résultats qui semblent intéressants J. (1988) sur les caractères amazighes imprimés d'une base de caractère locale.

(Djematen et *al*, 1998) considèrent que la méthode publiée par (Oulamara et *al*., 1988) n'est pas une technique appropriée pour les caractères amazighes manuscrits puisqu'elle produit des segmentations incorrectes. Pour surmonter la difficulté des caractères présentant des traits inclinés, ces auteurs (Djematen et *al*, 1997, 1998) proposent une méthode statistique de reconnaissance de caractères berbères manuscrits basée sur la position des points caractéristiques dans le rectangle-enveloppe de l'image du caractère. Après des prétraitements (normalisation bidirectionnelle, le lissage, l'extraction des composantes connexes) sur le caractère,

des primitives sont extraites sur chaque squelette, comme les extrémités, les points, les sommets (points de changements de direction) et les nœuds à 3 et 4 branches. Enfin, la représentation du caractère fournit une description sous forme de lettres utilisant un codage prédéfini. Cette description code les positions des points caractéristiques du caractère dans le rectangle-enveloppe. La reconnaissance consiste à mesurer le degré de ressemblance entre le code élaboré et les codes de référence en utilisant la distance métrique. Les résultats obtenus sont plus ou moins encourageants sur une base de caractères localement définie malgré quelques erreurs qui viennent du module de prétraitement.

Ait Ouguengay (2009) a proposé un réseau de neurones artificiels (RNA) pour la reconnaissance de caractères amazighs. Le réseau de neurones utilisé est un perceptron multicouche à une seule couche cachée. Ce dernier a été entrainé sur une base de données qui contient des patterns de la graphie amazighe de différentes fontes et tailles, créé localement. La simulation du réseau de neurones a été réalisée par le logiciel libre JavaNNS (java neural networks simulator). Les caractéristiques géométriques utilisées sont : les projections horizontales et verticales, les centres de gravité en x et en y, le périmètre, l'aire, la compacité et les moments centraux d'ordre 2. D'après l'auteur, cette approche a donné de bons résultats sur l'ensemble des patterns d'entraînement. Cependant, les résultats de test sont encore loin d'être satisfaisants à cause de la base de test qui est très faible par rapport aux poids de RNA à déterminer.

Pour sa part, El Ayachi et *al*. (2010) propose un système de reconnaissance de l'écriture tifinagh basé sur les moments invariants et la transformée de Walsh utilisant la programmation dynamique. Le système proposé contient trois parties principales : les prétraitements, l'extraction de caractéristiques et la reconnaissance. Dans le processus de prétraitement, l'image du document numérisé est nettoyée puis elle est segmentée en caractères isolés à l'aide des techniques de l'histogramme. Dans le processus d'extraction de caractéristiques, les moments invariants et les coefficients de Walsh sont calculés sur les caractères segmentés. La programmation dynamique est adoptée dans l'étape de reconnaissance. Les tests ont été faits sur plusieurs images d'écriture amazighe. D'après les auteurs, les résultats expérimentaux montrent que la méthode de la reconnaissance utilisant des moments invariants donne de meilleurs résultats par rapport à la méthode fondée sur la transformée de Walsh en termes de taux de reconnaissance, de taux d'erreur et de temps de calcul. Plus récemment, les mêmes auteurs (2011) ont proposé dans un réseau de neurones multicouches avec les mêmes caractéristiques utilisées précédemment. Les résultats trouvés avec un réseau de neurones d'une seule couche cachée sont meilleurs que ceux obtenus avec la programmation dynamique. De plus, le taux de reconnaissance obtenu en utilisant une seule couche cachée est plus élevé que celui obtenu avec deux ou trois couches cachées.

Dans la section suivante, nous présentons nos contributions en terme de conception et de développement de nouvelles approches de reconnaissance automatique de caractères amazighs imprimés ou manuscrits en les testant sur la base de caractères volumineuse et riche (A Database for Amazigh Handwritten Character Recognition Research :AMHCD) (Es Saady et *al*., 2011).

4. Nos systèmes de reconnaissance automatique proposés

4.1. Approches Markoviennes

Dans (Amrouch et *al*., 2010, 2012a), nous avons proposé une approche basée sur les modèles de Markov cachés de type modèle discriminant DM-HMM qui s'intéresse aux problèmes de caractères isolés. Ce type de modélisation est largement utilisé dans le domaine de la reconnaissance de la parole. Cette modélisation est aussi efficace pour reconnaître une forme entachée d'incertitude et d'aspect de dynamisme comme le caractère amazigh. Un processus Markovien a mis en oeuvre des modèles probabilistes spécifiques dans le but de gérer l'incertitude et le manque d'informations qui entachent les formes à reconnaitre. Après des prétraitements sur l'image du caractère amazigh, le système fait recourt à des primitives directionnelles dans la génération des séquences d'observations. Ces observations sont obtenues à l'aide de la technique des fenêtres glissantes opérant sur la transformée standard de Hough des images de caractères. Les séquences d'observations obtenues sont utilisées pour entraîner les modèles HMMs initiaux des caractères lors de la phase d'apprentissage ; chaque modèle utilise les échantillons de sa classe. Par la suite, nous avons utilisé le classifieur Forword pour reconnaître le caractère. En effet, cette approche consiste à associer un ou plusieurs modèles par classe. De ce fait, la reconnaissance s'effectue en estimant les probabilités d'émission de la suite d'observations O de la forme à reconnaître par les différents modèles préalablement construits. La forme à reconnaître est affectée à la classe dont le modèle qui maximise la probabilité. Cette approche est pratiquement utilisée dans le cas où le nombre de classes à reconnaître est relativement limité, c'est-à-dire au vocabulaire limité comme l'alphabet amazigh. Toutefois, elle devient coûteuse en temps de calcul et espace mémoire quand ce nombre dépasse le millier, puisque chaque classe possède au moins un modèle qui lui est propre.

Nous avons évalué les performances de notre système sur la base AMHCD des caractères amazighes manuscrits isolées (Es Saady et *al*., 2011), avec deux variantes. La première adopte la modélisation discrète des probabilités d'émission, quant à la seconde, elle utilise les HMMs continus. Le tableau 3 ci-dessous présente les différents taux de reconnaissance en utilisant le modèle discret et continu selon le nombre d'états de modèle et de la taille de la base de caractère.

Topologie de modèle utilisée	HMMs Discrets		HMMs Continus	
	Taille de la base	Taux de Recon	Taille de la base	Taux de Recon
Modèle à 14 états	2220 caractères	**90.04 %**	-	-
Modèle mono et 2 gaussiens à 6 états	-	-	25740 caractères	**96,21%**
Modèle mono et 2 gaussiens à 10 états	-	-	25740 caractères	**96 , 88%**
Modèle mono et 2 gaussiens à 14 états	-	-	25740 caractères	**97 , 89%**

Tableau 3 : Résultats de reconnaissance du système dans le cas discret et continu

Nous estimons que les erreurs de la variante discrète proviennent essentiellement de: (1) la modélisation discrète utilisée, à ce niveau, on a recours au risque de la perte des informations ; (2) à la taille des données utilisées pour créer la base des modèles de référence.

Pour valider cette hypothèse et remédier à cette défaillance, nous avons augmenté la taille des données utilisées et remplacé les modélisations discrètes par les HMMs continus.

L'augmentation de la taille des données d'apprentissage et la modélisation des densités des probabilités par les gaussiennes ont contribué à diminuer le taux d'erreur commis par notre système. Nous avons passé d'un taux d'erreur global de 9,6% (première expérience avec HMM discrets) à un taux de 2,11% c'est-à-dire un gain d'un facteur de 7,49%. De ce fait, les résultats obtenus dans le cas continu sont meilleurs que ceux obtenus dans le cas discret.

Dans (Amrouch et *al*., 2012b, 2012c), nous avons proposé un autre système pour la reconnaissance de caractères Tifinaghs imprimés, basé sur une nouvelle approche qui exploite les caractéristiques et les spécificités morphologiques de la langue amazighe. La solution apportée adopte une modélisation markovienne de type chemin discriminant (DP-HMM), optimisée par des algorithmes fondés sur la programmation dynamique S.K. (1996 ; Casey, E. (1996). L'approche s'appuie sur la proposition d'une nouvelle liste des segments, qui se compose d'un ensemble de traits fondamentaux constituant les caractères amazighs. Ceci permet de mieux exploiter la redondance de ces traits dans les tracés des lettres amazighes. La description de la structure des caractères repose sur ces éléments. En effet, les caractéristiques exploitées sont extraites des tracés des caractères par une technique de localisation implicite des segments qui le composent. Pour ce faire, nous avons utilisé les points d'intérêts des squelettes. Dans la phase de l'apprentissage, un seul modèle HMM global construit et entraîné sur les éléments du vocabulaire proposé par des primitives structurelles et géométriques. Chaque chemin au long de ce treillis représente une séquence de segments, qui constitue un caractère de

l'alphabet tifinagh. La reconnaissance s'effectue en décodant dynamiquement le chemin optimal suivant le critère de maximum de vraisemblance.

Pour valider le système proposé, nous avons effectué des expérimentations significatives sur la base de données de patterns de la graphie amazighe (Essaady et *al*., 2011). Plusieurs tests ont été effectués pour évaluer le taux de reconnaissance du système en fonction du : nombre d'états et du nombre de mélanges de gaussiennes. Le Tableau 4 ci-dessous présente les résultats obtenus de ces tests sur cette base.

Nombre d'états	3	5
Nombre de gaussienne	1-2-3	1-2-3
Taux de reconnaissance	**98 , 41%**	**98 , 76%**

Tableau 4 : Taux de reconnaissance sur BD1

Ces résultats montrent un taux d'erreur de 1,24% avec un modèle de topologie de 5 états. Nous estimons que les erreurs de reconnaissance sont attribuées, d'une part, aux méthodes utilisées pour la pré-classification et à la détection des points d'intérêts et, d'autre part, à l'insuffisance des caractéristiques utilisées pour mieux décrire chaque segment. En outre, cette faible erreur provient aussi de la déformation de certains caractères dans certaines fontes, notamment dans les fonts « Tassafut » et « Taromeit ». Nous constatons que le nombre de gaussiennes utilisées n'influence pas les résultats alors que son augmentation implique un nombre important de paramètres à calculer. Cependant, le choix de la topologie influence directement les résultats. Par conséquent, l'augmentation de nombre d'états augmente le taux de reconnaissance de système.

4.2. Approches syntaxiques

Dans (Essaady et *al*., 2010), nous avons présenté un système automatique de reconnaissance de caractères amazighs imprimés, basé sur une approche syntaxique utilisant les automates finis. Après des prétraitements sur l'image du caractère, des algorithmes appropriés sur le squelette de caractère permettent de construire la chaîne représentative du caractère à partir du codage de Freeman. La chaîne reconstruite est utilisée à l'entrée d'un automate fini qui reconnaît les caractèrs amazighes. Cet automate global a été construit à partir des automates de reconnaissance spécifique à chaque caractère amazigh. Nous avons testé notre application sur une base de caractères amazighes imprimés que nous avons créée. Nous avons obtenu des résultats encourageants. En effet, sur les 630 caractères lus, 589 ont été reconnus, soit un taux de reconnaissance de 93,49%. Le Tableau 5 ci-dessous présente les taux des mauvaises affectations et mauvais rejets. Ces erreurs proviennent de la forme de certains caractères non reconnus, dont le squelette comporte plus de segments non orthogonaux. En effet, la méthode de reconnaissance est basée sur une vectorisation du squelette du caractère à reconnaître. Donc, une erreur de vectorisation va forcément entraîner une erreur dans la description du caractère. En fait, le principal inconvénient de cette méthode est la sensibilité du squelette au bruit.

Taux d'affectations à tort	2,28 %
Taux de rejets à tort	4,23 %

Tableau 5 : Pourcentages des erreurs de reconnaissance

4.3. Approches neuronales

El Ayachi et *al*. (2010) proposent un système de reconnaissance de l'écriture tifinagh basé sur les réseaux de neurones multi couches avec les mêmes caractéristiques utilisées précédemment. Les résultats trouvés avec un réseau de neurones d'une seule couche cachée sont meilleurs que ceux obtenus avec la programmation dynamique. De plus, le taux de reconnaissance obtenu en utilisant une seule couche cachée est plus élevé que celui obtenu avec deux ou trois couches cachées.

Dans (Essaady et *al*., 2011b), et pour améliorer les résultats du précédent système et d'avoir un système complet qui reconnaît la totalité des caractères amazighs imprimés et manuscrits, nous avons présenté un système de reconnaissance automatique du texte amazigh, basé sur les réseaux de neurones multicouches. L'approche proposée a recours à des primitives statistiques en se basant sur la position des lignes centrales du caractère et les techniques de fenêtres glissantes.

Dans ce système et dans un premier temps, nous avons utilisé la ligne centrale horizontale du caractère pour extraire un ensemble de caractéristiques de densité basées sur cette ligne. Dans un second temps, nous avons proposé une amélioration de ce système de reconnaissance en ajoutant d'autres caractéristiques de densité basées sur la ligne centrale verticale du caractère afin d'exploiter la similarité de plusieurs caractères amazighs par rapport à la ligne centrale verticale du caractère. Les différentes variantes ont été testées et évaluées sur deux bases : la base des patterns de la graphie amazighe (Ait ougangay, 2009) et la base AMHCD des caractères amazighs manuscrits (Essaady et *al*., 2011a). En effet, nous avons montré les résultats de reconnaissance obtenus en fonction de l'intégration des caractéristiques dépendantes et indépendantes à la ligne centrale horizontale du caractère. Enfin, nous avons présenté les résultats obtenus par la version améliorée et leur comparaison avec ceux obtenus par la première variante du système.

Nous avons aussi utilisé des techniques de validation croisée pour l'évaluation des résultats de reconnaissance. La validation croisée est une méthode d'estimation de la fiabilité des résultats, fondée sur une technique d'échantillonnage R. (1995).

Le Tableau 6, ci-dessous, présente les résultats du système proposé en utilisant la validation croisée 10 fois sur la base des patterns de la graphie amazighe et sur la base de caractères manuscrits.

Les caractéristiques intégrées	**Base des patterns de la graphie amazighe**		**Base de caractères amazighs manuscrits**	
	Taille de la base	*Taux de Recon*	*Taille de la base*	*Taux de Recon*
Caractéristiques indépendantes de la ligne centrale	19437 caractères	**88.68 %**	20150 caractères	**84.49 %**
Caractéristiques dépendantes et indépendantes de la ligne centrale horizontale	19437 caractères	**98.49 %**	20150 caractères	**92.23 %**
Caractéristiques dépendantes et indépendantes de la ligne centrale horizontale et verticale	19437 caractères	**99.28 %**	20150 caractères	**96.32 %**

Tableau 6 : Résultats de reconnaissance du système amélioré en fonction des caractéristiques intégrées en utilisant la validation croisée 10 fois

Pour la base des patterns de la graphie amazighe, le taux de reconnaissance est 88,68% lors de l'utisation seulement des caractéristiques indépendantes de la ligne centrale horizontale et augmente à 98,49% lors de l'ajout des caractéristiques basées sur la position de la ligne centrale horizontale. Ce dernier taux s'élève aussi à 99,28% lors de l'ajout des caractéristiques basées sur la position de la ligne centrale verticale.

Pour la base de caractères amazighs manuscrits, le taux augmente de 84.49% à 92,23% lors de l'ajout des caractéristiques basées sur la position de la ligne centrale horizontale et monte à 96,32% lors de l'ajout des caractéristiques basées sur la position de la ligne centrale verticale.

Les erreurs sont principalement dues à une grande similarité morphologique entre certains caractères amazighs et, parfois, sur des fontes différentes.

En comparant les différents taux de reconnaissance obtenus par le système en fonction de caractéristiques intégrées, nous constatons une amélioration due à l'intégration des caractéristiques basées sur la position des deux lignes centrales (verticale et horizontale). Cela confirme que ces caractéristiques offrent une amélioration significative à la performance de reconnaissance.

5. Synthèse comparative

Le Tableau 7 récapitule les principales caractéristiques, les techniques et les résultats obtenus par des approches et les systèmes de reconnaissance de l'écriture amazighe que nous avons proposés.

Auteurs	Primitives	Modélisation	Résultats
(Amrouch et *al*., 2010, 2012a)	- primitives directionnelles (Transformée de Hough)	- HMM continu et discret 14 -états - topologies de type gauche droit. - apprentissage le critère (MLE) - Classification FORWARD	- 90.4 % en cas discret. - 97, 89% en cas continu. - Base (AMHCD). - Base d'apprentissage (2/3). - Base de test (1/3).
(Amrouch et *al*., 2012b, 2012c)	- primitives structurelles (diamètre, excentricité, étendue, Centre de masse, orientation, longueur d'axe minimal et principal, moment d'ordre 1 et 2)	- HMM continu. - topologies de type linéaire et ergodique - Apprentissage Baum-welsh - Classification Viterbi	- 98 ,76% la base de données de patterns de la graphie amazighe. Base d'apprentissage (2/3) Base de test (1/3)
(Essaady et *al*., 2010)	- Codage de Freeman (suivi de squelette)	- réseaux des automates. - (Grammaire formelle)	- 93,49% - selon les données utilisées.
(Essaady et *al*., 2011c)	- primitives statistiques - caractéristiques dépendantes et indépendantes de la ligne centrale horizontale et verticale - fenêtres glissantes horizontales et verticales	- Réseaux de neurones multicouches. - couche d'entrée 95 neurones - couche de sortie 31 neurones - couche cache (nb d'entrées + nb de sorties)/2	- 99.28 % Base des patterns de la graphie amazighe (19437 caractères) - 96.32 % Base AMHCD (20150 caractères)

		- Apprentissage de Heb (η=0,2, taux=0,3, itéra=1000)	
(Elaychi et *al*., 2010)	- les moments invariants et la transformée de Walsh utilisant la programmation dynamique.	- La programmation dynamique	- Taux de reconnaissance très intéressant
(Elaychi et *al*., 2011)	- un réseau de neurones multi couches avec les mêmes caractéristiques utilisées précédemment.	- Programmation dynamique	- Résultats très intéressants
(Elaychi et *al*., 2011b)	- Les caractéristiques géométriques : les projections horizontales et verticales, les centres de gravité en x et en y, le périmètre, l'aire, la compacité et les moments centraux d'ordre 2.	- réseau de neurones artificiels (RNA)	- Résultats très intéressants
(Djematen et *al*., 1998)	- les extrémités, les points, les sommets (points de changements de direction) et les nœuds à 3 et 4 branches.	- méthode statistique : mesurer les degrés de vraisemblance	- Résultats plus ou mois intéressants

Tableau 7 : synthèse des approches et systèmes proposés

Après avoir étudié et comparé les différentes approches, nous constatons que la meilleure et la performante approche, quant au temps de calcul, espace mémoire et taux de reconnaissance, est l'approche basée sur les primitives statistiques et les fenêtres glissantes dans la phase de prétraitement et les réseaux de neurones dans la phase de classification (Essaady et *al*., 2011c).

6. Conclusion et perspectives

Nous avons présenté dans ce papier des travaux traitant la reconnaissance automatique des caractères amazighs. L'objectif est de faire une synthèse des travaux effectués sur ce sujet. Ces approches présentent un certain nombre de limites qui proviennent à la fois du module de prétraitement et des caractéristiques prises dans la phase d'apprentissage. En plus, les bases de caractères utilisées dans les tests restent parfois faibles et parfois non standards. Par conséquent, des travaux de recherche futurs doivent apporter des améliorations, d'un côté, sur ces approches et, d'un autre côté, développer d'autres systèmes qui répondent aux attentes. Et parmi nos travaux futurs, nous allons proposer des systèmes hybrides qui utilisent des primitives de nature différente en combinant ces approches dans le processus de traitement. Ce qui permettra de profiter *a priori* des avantages de chacune des approches tout en évitant les principaux inconvénients.

7. Références bibliographiques

Ait Ouguengay Y., Taalabi M. (2009), Elaboration d'un réseau de neurones artificiels pour la reconnaissance optique de la graphie amazighe: Phase d'apprentissage, *Systèmes intelligents-Théories et applications, Paris : Europia, cop. (impr. au Maroc)*, ISBN-102909285553, Avril 2009.

Amrouch M., Rachidi A., El Yassa M., Mammass D. (2010), "Handwritten Amazigh Character Recognition Based On Hidden Markov Models", *ICGST-GVIP Journal*, vol.10, Issue 5, pp.11-18, December 2010.

Amrouch M., Es Saady Y., Rachidi A., El Yassa M., Mammass D. (2012a), "Handwritten Amazigh Character Recognition System Based on Continuous HMMs and Directional Features*", IJMER journal*, Vol.2, Issue 2, pp.436-441, Mar.-Apr. 2012.

Amrouch M., Es Saady Y., Rachidi A., El Yassa M., Mammass D. (2012b), "A New Approach Based on Strokes for Printed Tifinagh Character Recognition Using Discriminating Path-HMM ", *IRECOS journal*, Vol.7, N°.2, Mars 2012.

Amrouch M., Es Saady Y., Rachidi A., El Yassa M., Mammass D. (2012c), "A Novel Feature Set for Recognition of Printed Amazigh Text Using Maximum Deviation and HMM", *IJCA journal*, Vol.44, N°.12, pp.23-30, April 2012.

Casey R.G. and Lecolinet E. (1996), A survey of methods and strategies in character segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol.18,N°.7, pp.690–706, Jul 1996.

Djematen A., Taconet B. and Zahour A. (1997), "A Geometrical Method for Printing and Handwritten Berber Character Recognition", *ICDAR'97*, pp.564, 1997.

Djematen A., Taconet B. and Zahour A. (1998), « Une méthode statistique pour la reconnaissance de caractères berbères manuscrits », *CIFED'98*, pp.170-178, 1998.

El Ayachi R., Moro K., Fakir M. and Bouikhalene B. (2010), "On the Recognition of Tifinaghe Scripts", *Journal of Theoretical and Applied Information Technology*, vol.20 (2), pp.61-66, 2010.

El Ayachi R., Moro K., Fakir M. and Bouikhalene B. (2011), "Recognition of Tifinaghe Characters Using a Multilayer Neural Network", *International Journal Of Image Processing (IJIP)*, vol. 5, Issue 2, 2011.

Es Saady Y., Rachidi A., El Yassa M., Mammass D. (2011), AMHCD: A Database for Amazigh Handwritten Character Recognition Research. *International Journal of Computer Applications* , Vol.27, N°.4, pp:44-48, August 2011. Published by Foundation of Computer Science, New York, USA.

Es Saady Y., Rachidi A., El Yassa M., Mammass D. (2010), "Printed Amazigh Character Recognition by a Syntactic Approach using Finite Automata*"*, *ICGST-GVIP Journal*, vol.10, Issue 2, pp.1-8, 2010.

Es Saady Y., Rachidi A., El Yassa M., Mammass D. (2011a), "AMHCD: A Database for Amazigh Handwritten Character Recognition Research", *International Journal of Computer Applications*, vol.27 (4), pp.44-48, published by Foundation of Computer Science, New York, August 2011.

Es Saady Y., Rachidi A., El Yassa M., Mammass D. (2011b), "Amazigh Handwritten Character Recognition based on Horizontal and Vertical Centerline of Character", *International Journal of Advanced Science and Technology*, vol.33, pp.33-50, August, 2011.

Kim H.J., Jung J.W. and Kim S.K. (1996), On-line Chinese character recognition using ARTbased stroke classification. *Pattern Recognition Letters*, Vol.17, N°.12, pp.1311–1322, 1996.

Kohavi R. (1995), "A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", *International Joint Conference on Artificial Intelligence (IJCAI)*, 1995.

Ly Van B. (2005), "Réalisation d'un Système de Vérification de Signature Manuscrite En-ligne Indépendant de la Plateforme d'Acquisition", *Thèse de doctorat de l'Institut National des Télécommunications,* Décembre 2005.

Oulamara A. and Duvernoy J. (1988), "An application of the Hough transform to automatic recognition of Berber characters", *Signal Processing*, vol.14, pp.79-90, 1988.

Asinag-Asinag, 9, 2014, p. 133-145

Vers un système de reconnaissance automatique de la parole en amazighe basé sur les transformations orthogonales paramétrables

Abenaou Abdenbi (1), Ataa Allah Fadoua (2) & Nsiri Benayad (3)

(1) ENSA, Université Ibn Zohr,
(2) CEISIC, IRCAM,
(3) FSAC, Université Hassan II

In this paper, we propose a new approach for Amazigh isolated word recognition, based on relevant speech signal parameters' extraction algorithm. In general, the approach consists on the application of adaptive orthogonal transforms that are characterized by a linear operator constituted of configurable functions, which allows the transform adaptation to the initial data and the reduction of feature vector dimension, that improve the isolated word recognition rate.

1. Introduction

Durant cette dernière décennie, l'évolution permanente des technologies de l'information et de la communication a été marquée par des progrès majeurs dans le déploiement du traitement du langage humain, notamment la reconnaissance automatique de la parole, pour la promotion et le développement des langues peu dotées.

De nos jours, en effet, la reconnaissance automatique de la parole est introduite dans de nombreuses applications ; à savoir les systèmes d'apprentissage des langues pour améliorer la prononciation des apprenants (Bahi, 2008), les applications téléphoniques du type serveur vocal pour l'accès aux services (Barnard et *al.*, 2009) ou l'accès à l'information à travers la recherche dans des bases de données vocales particulièrement pour les personnes à besoins spécifiques et les analphabètes surtout dans les régions rurales (Barnard et *al.*, 2010 ; Patel et *al.*, 2010 ; Kumar et *al.*, 2011), ainsi que les applications de transcription automatique des documents radio et télédiffusés.

Cependant, les technologies de la parole ne sont pas suffisamment exploitées pour la langue amazighe. Afin de profiter des avantages de ces technologies, nous avons consacré cette étude à la réalisation d'un premier système de reconnaissance de mots isolés amazighes à la base des transformations orthogonales paramétrables.

Généralement, en traitement du signal vocal, la résolution des problèmes de reconnaissance passe nécessairement par une étape d'extraction des caractéristiques informatives des signaux avant d'entamer la phase d'analyse. Parmi les travaux de recherche traitant l'extraction des caractéristiques à partir des mots isolés, nous distinguons deux principaux types d'approches : les méthodes à base des théories statistiques (Bourlard et Morgan, 1993 ; Doddington, 1985 ; Cappe, 1995) et les méthodes déterministes à base des transformations orthogonales classiques (Walsh, Haar, Fourier, …) (Kekre et *al.*, 2010 ; Ahmed et Rao, 1975). Néanmoins, les méthodes statistiques, tel que le modèle de Markov caché, ont atteint leurs limites dans l'amélioration des systèmes de la reconnaissance automatique des signaux vocaux, malgré la disposition de corpus suffisamment représentatifs. Tandis que les méthodes spectrales ont émergé dans plusieurs applications du traitement du signal vocal grâce à la richesse de leurs propriétés et la rapidité du calcul de leur algorithme (Bello et *al.*, 2004 ; Doets et Lagendijk, 2004).

Le principe fondamental de ces méthodes, particulièrement celles liées à un système de fonction de base orthogonale (non paramétrable) comme la transformée de Fourier ou la transformée en ondelettes, est d'obtenir le vecteur spectral des caractéristiques informatives.

Cependant, le spectre obtenu par ces méthodes est généralement trop large, vu que le signal vocal est un processus non stationnaire. Ce qui complique souvent la procédure de reconnaissance des signaux et conduit, dans certains cas, à des résultats insatisfaisants. D'où la nécessité d'une méthode de détermination des caractéristiques informatives du signal vocal dont le coût de calcul est optimal.

Dans cet article, nous proposons une solution au problème en utilisant les transformations orthogonales adaptables pour l'extraction des caractéristiques informatives du signal vocal, tout en visant la réalisation d'un système de reconnaissance de la parole amazighe dédié à l'apprentissage de la prononciation. L'utilisation de ces transformations (Abenaou et Sadik, 2011a, 2011b, 2011c) est favorisée par la possibilité d'adaptation de la forme de leurs fonctions de base en fonction du caractère du vecteur étalon. Ce dernier est formé par les différents signaux vocaux de chaque mot. Autrement dit, à chaque classe de mots est associé un système de fonctions de base paramétrables pour la projection des signaux. En outre, ces fonctions répondent au critère de la complétude du système, qui assure les transformations des signaux sans perte de leur contenu informatif. Le système de fonctions de base formé s'exprime sous forme d'un opérateur matriciel orthonormé factorisable, ce qui permet une transformation à la base d'un algorithme à calcul rapide.

2. Méthode et algorithme de synthèse de l'opérateur de la transformée orthogonale adaptable

En traitement numérique, la transformée linéaire orthogonale d'un signal X peut être représentée par l'équation matricielle (1):

$$Y = \frac{1}{N} HX \qquad (1)$$

où:

- $\boldsymbol{X} = [x_1, x_2, \ldots, x_N]^T$ est le signal initial à transformer, dont la taille $\boldsymbol{N} = 2^n$;
- $\boldsymbol{Y} = [y_1, y_2, \ldots, y_N]^T$ est le vecteur des coefficients spectraux, calculé par l'opérateur spectral orthogonal $\boldsymbol{H}$ de dimension $\boldsymbol{N}$ x $\boldsymbol{N}$.

La factorisation de Good (Good, 1960) a montré la possibilité de représenter l'opérateur matriciel $\boldsymbol{H}$ sous forme de produit de matrices creuses $\boldsymbol{G_i}$ (2) avec une proportion plus élevée des zéros ce qui permet la construction des algorithmes de transformation rapide de Walsh, de Haar et de Fourier. Les matrices $\boldsymbol{G_i}$ ($i = 1,\ldots, n$) sont construites par des blocs de matrices $\boldsymbol{V_{i,j}}$ de dimension minimale qui s'appellent noyaux spectraux (Abenaou et Sadik, 2011a).

$$G_i = \begin{bmatrix} \begin{bmatrix} \alpha_{i1} & 0 \\ \beta_{i1} & 0 \end{bmatrix} & 0 & \cdots & 0 & \begin{bmatrix} \gamma_{i1} & 0 \\ \delta_{i1} & 0 \end{bmatrix} & 0 & \cdots & 0 \\ 0 & \begin{bmatrix} \alpha_{i2} & 0 \\ \beta_{i2} & 0 \end{bmatrix} & & 0 & 0 & \begin{bmatrix} \gamma_{i2} & 0 \\ \delta_{i2} & 0 \end{bmatrix} & & 0 \\ 0 & & \ddots & 0 & 0 & & \ddots & 0 \\ 0 & \cdots & 0 & \begin{bmatrix} \alpha_{iN/2} & 0 \\ \beta_{iN/2} & 0 \end{bmatrix} & 0 & \cdots & 0 & \begin{bmatrix} \gamma_{iN/2} & 0 \\ \delta_{iN/2} & 0 \end{bmatrix} \end{bmatrix} \qquad (2)$$

où :

$$V_{i,j} = \begin{bmatrix} \alpha_{ij} & \cdots & \gamma_{ij} \\ \beta_{ij} & \cdots & \delta_{ij} \end{bmatrix} = \begin{bmatrix} \cos(\varphi_{i,j}) & \cdots & w_{i,j} \sin(\varphi_{i,j}) \\ \sin(\varphi_{i,j}) & \cdots & -w_{i,j} \cos(\varphi_{i,j}) \end{bmatrix},$$

$$w_{i,j} = \exp(j\theta_{i,j}), \quad \varphi \in [0, 2\pi], \quad \theta \in [0, 2\pi].$$

D'où la relation (1) peut s'écrire comme suit :

$$Y = \frac{1}{N} HX = \frac{1}{N} G_1 G_2 \ldots G_n X = \frac{1}{N} \prod_{i=1}^{n} G_i X \qquad (3)$$

En définissant les paramètres angulaires $\boldsymbol{\varphi}_{i,j}$ et $\boldsymbol{\theta}_{i,j,}$ les opérateurs de transformations orthogonales $\boldsymbol{H}$ peuvent être formés avec des fonctions de base complexes, ou avec des fonctions réelles lorsque $\boldsymbol{\theta}_{i,j} = 0$. Le calcul des paramètres $\boldsymbol{\varphi}_{i,j}$ dépend du choix des structures des noyaux spectraux $\boldsymbol{V_{i,j}}$ **(**Abenaou et Sadik, 2011c). Ce qui permet de générer un système de fonctions de base adaptable à une classe de signaux donnée.

Or, dans la perspective d'assurer un calcul rapide, dans ce travail, les noyaux spectraux dans les matrices $\boldsymbol{G_i}$ sont constitués de telle sorte qu'ils contiennent une proportion plus importante de zéros, tel qu'il est expliqué ci-dessous.

L'adaptation de l'opérateur ***H*** (1) est assurée par la condition :

$$\frac{1}{N} H_a Z_{et} = Y_c = [y_{c,1}, 0, 0, \ldots 0]^T, \quad y_{c,1} \neq 0 \qquad (4)$$

où :

- $\boldsymbol{Y}_c$ est le vecteur cible qui construit le critère d'adaptation de l'opérateur $\boldsymbol{H_a}$;
- $\boldsymbol{Z}_{et}$ représente le vecteur étalon d'une classe calculé par la moyenne des estimations statistiques des enregistrements de plusieurs signaux vocaux, d'un même mot, prononcés par divers locuteurs ;
- $\boldsymbol{H_a}$ est l'opérateur adaptable à synthétiser.

La synthèse de l'opérateur adaptable $\boldsymbol{H_a}$ à l'étalon $\boldsymbol{Z}_{et}$, pour une classe donnée, consiste à calculer les paramètres angulaires $\boldsymbol{\varphi}_{i,j}$ des matrices $\boldsymbol{G_i}$ selon la condition (4). La procédure du calcul des paramètres est illustrée par la figure 1 dont le principe est basé sur l'algorithme itératif introduit par la figure 2, qui permet le calcul du vecteur cible $\boldsymbol{Y_c}$ selon la relation :

$$Y_i = G_i Y_{i-1}$$

Le calcul du vecteur $\boldsymbol{Y_c}$ permet l'obtention de l'opérateur adapté $\boldsymbol{H_a}$. Pour la reconnaissance des signaux, nous devons disposer de deux ensembles d'enregistrements de signaux vocaux pour chaque mot. Le premier sert à calculer l'étalon $\boldsymbol{Z_{et,i}}$ du mot *i* (classe *i*) et permet de générer la synthèse de l'opérateur. Tandis que le deuxième ensemble sert à former l'étalon spectral $\boldsymbol{Y_{et,i}}$ du mot *i*, qui est obtenu par la projection des enregistrements du deuxième ensemble dans les bases adaptables $\boldsymbol{H_a}$.

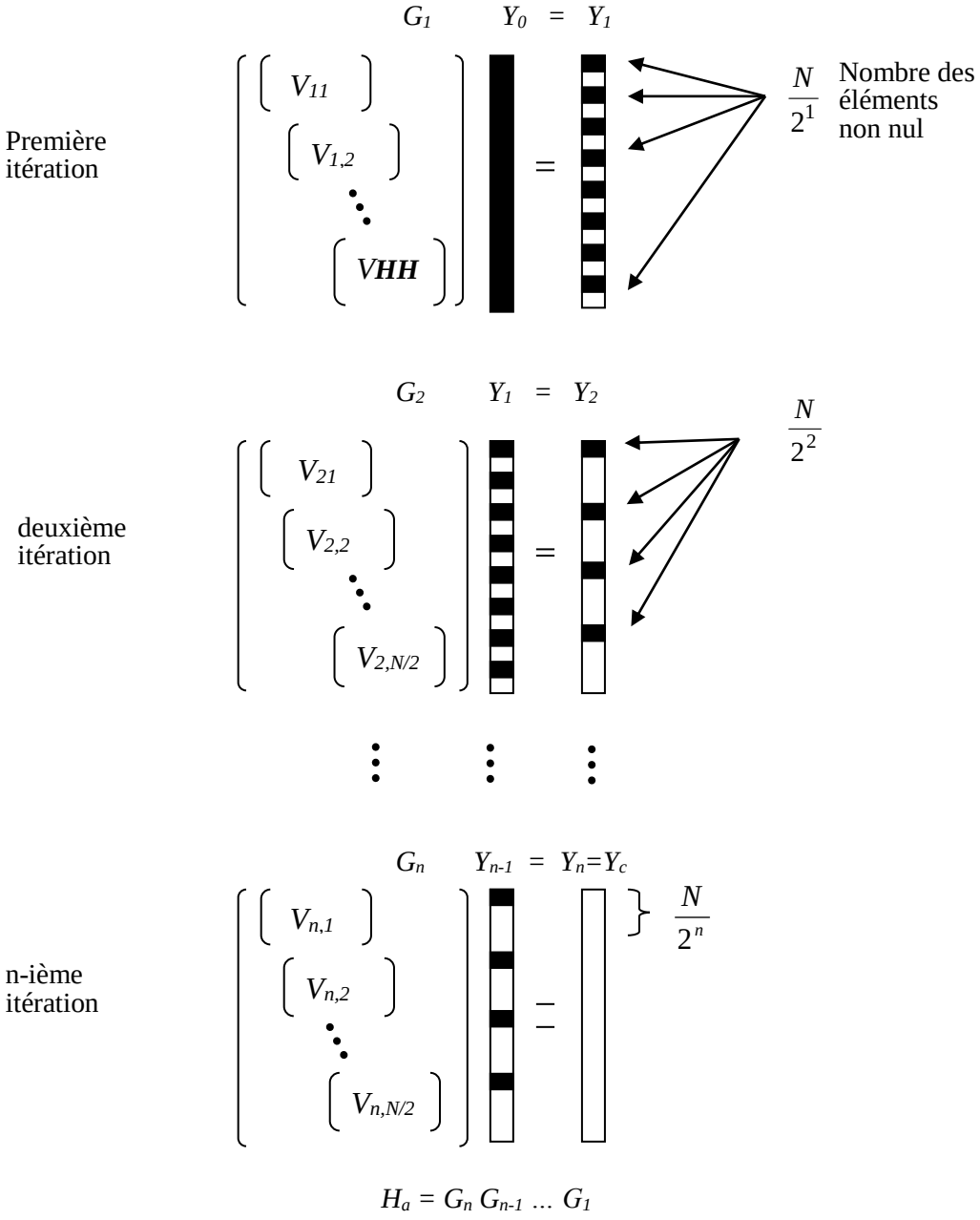


Figure 1 : Schéma illustratif de la procédure de synthèse de l'opérateur de la transformée adaptable

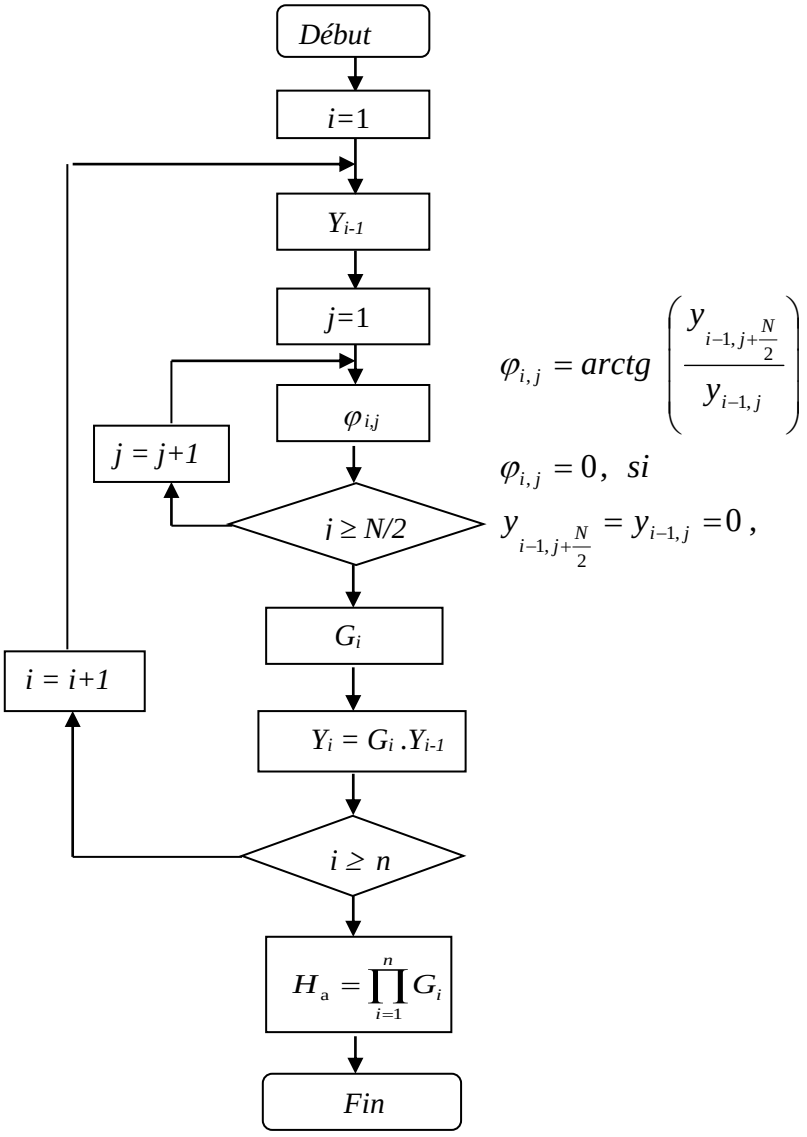


Figure 2 : Schéma de l'algorithme de synthèse de l'opérateur de la transformée adaptable

La reconnaissance d'un vecteur ***Z*** consiste à calculer son spectre Y_i dans chaque base $H_{a,i}$. Pour définir le mot correspondant au vecteur Y_i des caractéristiques informatives, nous nous appuyons sur une règle de décision formée par une combinaison de deux critères :

- la distance euclidienne $\delta_i = \| \boldsymbol{Y_i} - \boldsymbol{Y_{et,i}} \|$ et
- la différence de l'énergie concentrée dans leurs premiers coefficients de la décomposition $\varepsilon_i = \left| y_{1,i}^2 - y_{1,et,i}^2 \right|$.

Ainsi, le vecteur $\boldsymbol{Y_i}$ correspondra au mot *i* si $\delta_i = \min(\delta_{k=1..M})$ et $\varepsilon_i = \min(\varepsilon_{k=1..M})$, avec *M* est le nombre de classes. Cette procédure de reconnaissance est illustrée par la figure 3.

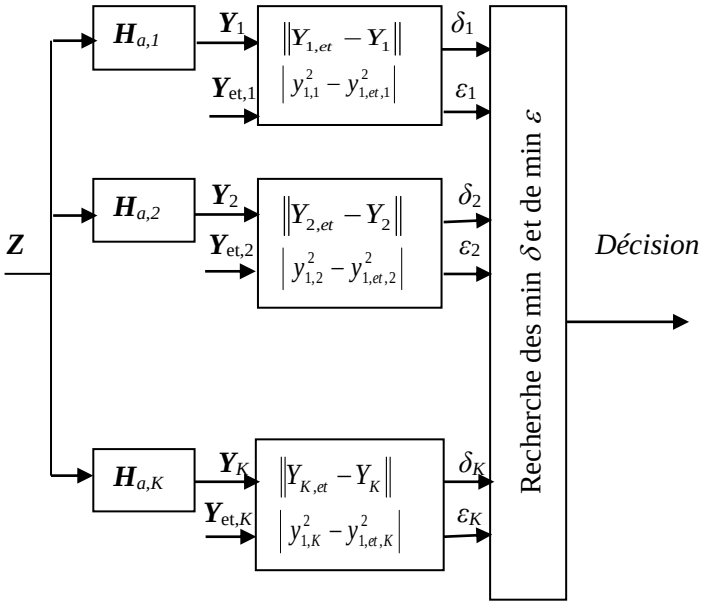


Figure 3 : Procédure de reconnaissance

3. La reconnaissance de la parole amazighe

Malgré l'avènement des technologies de la reconnaissance de la parole pour l'anglais, le français et l'arabe, des recherches approfondies au profit de la langue amazighe semblent insuffisantes et la mise en œuvre de ses applications est presque inexistante. D'où l'intérêt de la réalisation d'un système de reconnaissance de la parole en amazighe, en particulier un système qui pourra être dédié à l'apprentissage de la prononciation. Néanmoins, dans la perspective d'atteindre cet objectif, nous avons recours à un corpus qui caractérise la langue parlée.

3.1. Corpus

Vu la rareté et la non disponibilité des ressources électroniques en langue amazighe, particulièrement les corpus audio, nous avons recueilli, pour une première paramétrisation de notre système de reconnaissance de mots amazighes isolés, un

corpus de données vocalees multi-locuteurs. Ce dernier est constitué de 140 enregistrements des chiffres de 1 à 10, réalisés par trois locuteurs de différentes variétés régionales (Tarifit, Tamazight, Tachelhit).

En outre, ce corpus est regroupé en trois ensembles : le premier servira à calculer l'étalon de chaque mot pour générer la synthèse de l'opérateur ; le deuxième, à former les étalons spectraux des mots tandis que le troisième sera utilisé pour évaluer et analyser les performances de l'approche proposée.

3.2. Mesures d'évaluation

Afin d'évaluer la performance de notre système de reconnaissance, nous utilisons la mesure de taux d'exactitude par mot (Word Accuracy, WA), définies par la formule suivante (Sopheap, 2010) :

$$\textit{Taux d'exactitude} = j/h * 100 \quad (5),$$

où j correspond au nombre de mots justes et h est le nombre total de mots.

3.3. Résultats expérimentaux

Pendant l'expérience, nous avons utilisé des enregistrements de signaux vocaux des mots amazighes isolés prononcés par différents locuteurs de diverses régions, ce qui a induit à un chevauchement assez considérable entre les classes des mots. Pour évaluer l'efficacité du système proposé, un test a été effectué pour la reconnaissance d'un même mot amazighe prononcé par divers locuteurs de diverses régions. La figure 4 illustre la projection du signal vocal du mot « □□□ (*sin*) » (deux) dans les bases classiques (Walsh, Haar et Fourier). D'après cette figure, nous constatons que les spectres calculés du mot sont trop larges. Cependant, en utilisant la méthode proposée, nous remarquons une convergence rapide du spectre obtenu à l'aide des fonctions de base paramétrables.

Par ailleurs, grâce à l'application des transformations orthogonales paramétrables aux bases synthétisées, nous constatons que :

- l'énergie de la projection du signal dans la base adéquate est concentrée dans les premiers composants du spectre (figure 4) ; et
- la projection du signal d'un mot donné, qui caractérise une classe, dans d'autres classes (représentant d'autres mots amazighes) permet l'obtention de spectres assez larges dont l'énergie est dispersée sur plusieurs coefficients (figure 5).

Ce qui nous permet de reconnaître le mot prononcé avec une grande certitude.

En effet, les résultats de l'étude expérimentale de la méthode élaborée pour la reconnaissance des mots amazighes isolés indiquent, selon les courbes de la figure 6 qui présente les taux de certitude de la reconnaissance des signaux, une efficacité considérable par rapport aux autres méthodes qui sont basées sur l'application des transformations spectrales dans les bases traditionnelles (Walsh, Haar et Fourier).

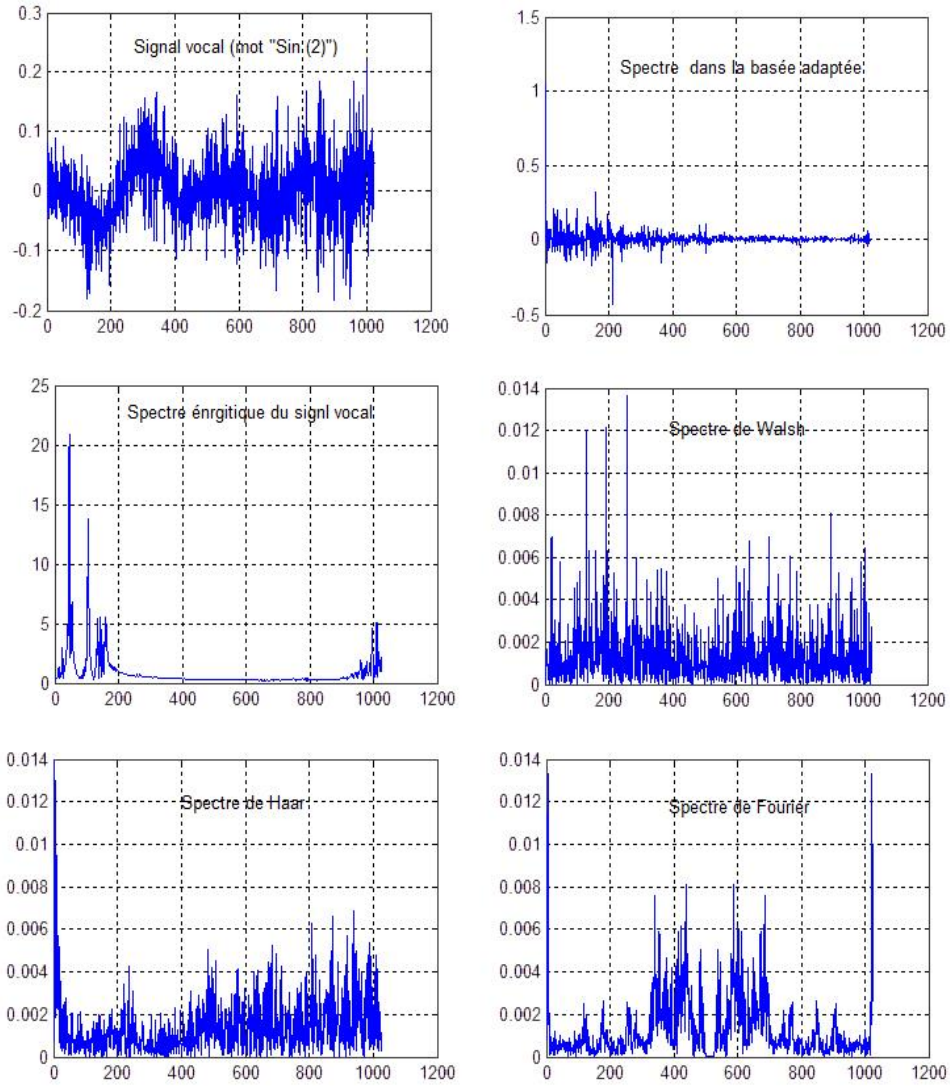


Figure 4 : Projection du signal vocal (fragment du mot □□/ « Sin » - 2)

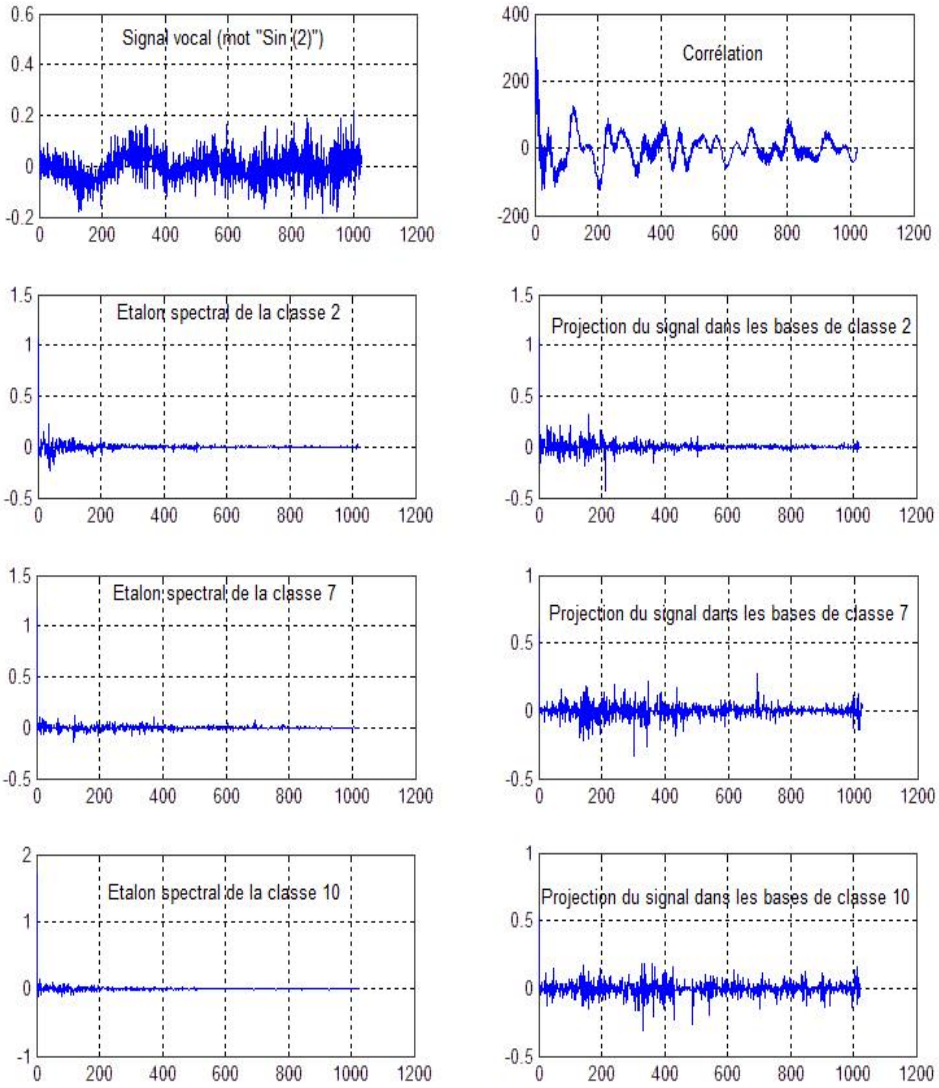


Figure 5 : Calcul du spectre du signal vocal à l'aide des fonctions de bases paramétrables

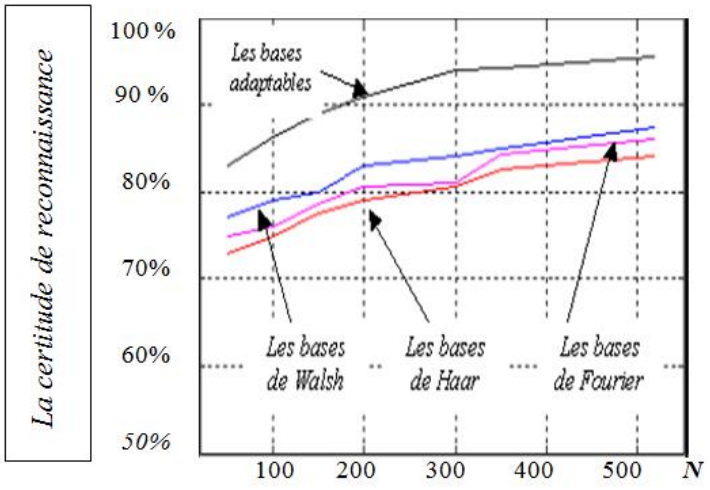


Figure 6 : Résultat de la reconnaissance lors de l'application de divers systèmes de fonctions de base

A partir de ces courbes, nous pouvons constater que dans le cas de l'utilisation des bases traditionnelles, la certitude de reconnaissance des signaux des mots amazighes ne dépasse pas 87%. Tandis que dans le cas où nous utilisons les fonctions de bases adaptables, le taux de reconnaissance des signaux s'élève à 96% lorsque la taille de l'intervalle de l'analyse est égale à 512.

Ce qui peut être expliqué par le fait que :

- la méthode proposée est basée sur l'utilisation des fonctions de bases adaptables selon le caractère du signal vocal du mot prononcé par les divers locuteurs des différentes régions ; et
- la propriété de sélectivité des fonctions de base synthétisées simplifie la distinction des signaux dans l'espace des caractéristiques informatives.

Conclusion

Dans la présente contribution, nous proposons un système de reconnaissance automatique des mots isolés de la langue amazighe basée sur les transformations orthogonales paramétrables. Ce dernier est composé de deux sous-systèmes : un sous-système d'apprentissage et un sous-système de reconnaissance. Le sous-système d'apprentissage est conçu à la base d'un corpus multi-locuteurs de la parole amazighe de différentes régions afin de prendre en considération la diversité de la prononciation d'un même mot et de contribuer à la stabilité des caractéristiques statistiques dans le calcul de l'étalon de chaque mot. En outre, ce sous-système nous offre la possibilité de synthétiser des fonctions de base adaptables de chaque mot avec une propriété de sélectivité plus importante, ce qui

nous a permis d'extraire les caractéristiques les plus informatives de chaque mot prononcé indépendamment du locuteur.

Suite à l'étude comparative réalisée sur le système à base des transformations orthogonales paramétrables et sur les transformations orthogonales de Walsh, Haar et Fourier, nous avons pu atteindre un taux de reconnaissance plus élevé qui tend vers les 96%. Cependant, nous considérons que ce travail est une première initiative pour la réalisation d'un système de reconnaissance de la parole amazighe assurant l'apprentissage de la prononciation, qui suscite l'intérêt de recueillir un corpus oral riche et varié composé de mots amazighes dédiés à l'apprentissage de la langue.

Références bibliographiques

Abenaou A. et Sadik M. (2011), « Elaboration d'une méthode de compression des signaux aléatoires à base d'une transformation orthogonale paramétrable avec algorithme rapide », *The Fourth Workshop on Information Technologies and Communication (WOTIC'11), ENSEM*, Casablanca.

Abenaou A. et Sadik M. (2011), « Méthode et algorithme de formation d'un système de fonctions de base adaptables pour le diagnostic des signaux biologiques », *Colloque International des Telecom'2011 & 7èmes Journées Franco-Maghrébines des Micro-ondes et leurs Applications*, Tanger.

Abenaou A. et Sadik M. (2012), « Méthode et algorithme d'identification des signaux vocaux à base des transformations orthogonales adaptables », *Network Security and Systems*, p. 41-45.

Ahmed N. et Rao K.R., (1975), *Orthogonal transforms for digital signal processing*, Springer Ber.

Bello J. et *al.* (2004), "*On the use of phase and energy for musical onset detection in the complex domain", IEEE Signal Procesing letters*, 11(6) : 553–556.

Bahi H. (2008), "Hybrid ASR system for teaching pronunciation". ICL Conference, 24 -26 Septembre, Villach, Austria.

Barnard E. et *al.* (2009), "Asr corpus design for resource-scarce languages". In *Interspeech*.

Barnard E. et *al.* (2010). Voice search for development. In Interspeech.

Bourlard H. et Morgan N. (1993), "Continuous speech recognition by connectionist statistical methods", *IEEE Transaction on Neural Networks*, Vol. 4, n° 6, pp. 893-909.

Cappe O. (1995), « Etat actuel de la recherché en reconnaissance du locuteur et des application en criminalistique », Rapport interne, Ecole Nationale des Télécommunications, Département du Signal, Paris.

Doddington G.R. (1985), "Speaker reconnaissance-identification people by their voices", Proc. IEEE; vol 73; no.11; p. 1651

Doets P. et Lagendijk R. (2004), "Theoretical modeling of a robust audio fingerprinting system", In *IEEE Benelux Signal Processing Symposium*.

Good I.J. (1960), The interaction algorithm and practical Fourier analysis, J. Roy. Statist. Soc. Ser. B, B-20, 361-372, 1958, B-22, 372-375.

Kekre H. B. et *al*., "Performance Comparison of Speaker Identification Using DCT, Walsh, Haar On Full And Row Mean Of Spectrogram", *International Journal of Computer Applications*, August 2010 Edition, in press.

Kumar A. et *al.* (2011), Rethinking speech recognition on mobile devices. In IUI4DR. ACM.

Patel N. et *al.* (2010), "Avaaj otalo: a field study of an interactive voice forum for small farmers in rural India", In *CHI*, pages 733–742, ACM.

Sopheap S. (2010), *Vers une modélisation statistique multi-niveau du langage, application aux langues peu dotées*. Thèse de doctorat, Université de Grenoble..

Asinag-Asinag, 9, 2014, p. 147-170

Conception et peuplement d'une ontologie modélisant la notion de contexte enrichie par les fonctions lexicales pour la détection du sens dans le texte

Parler du Maroc central

Hammou Fadili (1) & Malika Chakiri (2)

(1) Laboratoire CEDRIC du CNAM de Paris
(2) Paris-Descartes-Sorbonne

This work focuses on the design and the development of a context ontology model, based on a domain ontology enriched, by lexico-semantic relations defining the lexical functions introduced by Mel'cuk in the Meaning-Text Theory, and by the concept of the context, in order to improve the analysis and the detection of meaning in text. This is motivated, by the fact that, unstructured data constitute the majority of produced contents, requiring for their exploitation, the development of tools and technologies allowing their integration into knowledge-based and reasoning-based systems. Existing technologies in the analysis and the extraction of semantic information from text can have a lot of imperfections; indeed, important elements such as concept of the context and use of all possible relations between terms, are not fully and formally supported. It is why we propose an extended model of a domain ontology as a context (in our case an ontology of fauna and flora), enriched by aspects that can help to address the cited problems. Such ontologies constitute the basis of an advanced approach for the detection and the extraction of contextual and semantic information from text, published in a parallel article (H.Fadili ACS/IEEE, 2013).

Introduction

Ce travail porte sur l'élaboration d'un modèle d'ontologie de contexte à partir d'une ontologie de domaine enrichie par des relations lexico-sémantiques définissant les fonctions lexicales de la Théorie Sens-Texte (Mel'cuk, 1988) et par la notion du contexte, dans un but d'améliorer l'analyse et la détection sémantique des données dans les documents de type textes. Cette démarche, s'intéressant au traitement sémantique des données non structurées, est motivée par le fait que, ce type de contenus constitue la majorité des données produites aujourd'hui, nécessitant pour leur exploitation la mise en place d'outils et des technologies spécifiques permettant leur intégration dans les systèmes à base de connaissances

et à base de raisonnements. Dans ce domaine, les technologies existantes relatives à l'analyse et à l'extraction sémantiques d'information peuvent représenter beaucoup de lacunes du fait que des concepts importants, comme la notion de contexte et l'exploitation de toutes les relations possibles entre les termes, n'y sont pas totalement pris en charge.

C'est dans ce contexte que nous avons proposé une nouvelle approche permettant d'améliorer certains processus de gestion contextuelle de la sémantique (Fadili, 2013), basée sur les ontologies de domaine que nous améliorons suivant un « modèle » présenté dans cet article. Le but est de détecter et de relever tous les « traits sémantiques » relatifs à un contexte donné et d'augmenter ainsi l'efficacité de la formalisation du sens d'un texte afin qu'il soit compris et interprété par la « machine ». La modélisation de cette « relation » entre le texte et le contexte consiste dans la formalisation du sens en extrayant des mots et des relations permettant l'obtention d'un réseau sémantique « contextualisé » reflétant fidèlement le sens des contenus étudiés. Concrètement, elle permet de faire « la projection » du texte représenté par l'arbre conceptuel (AC) (F.AMARDEILH, 2009) issu des différents processus du TAL « sur » le contexte représenté par l'ontologie du contexte (OC) en utilisant les « relations lexico-sémantiques » pour faire le « mapping » entre les concepts, mots, relations, instances, attributs, etc. des deux graphes.

Nous avons tenu à mettre l'accent sur l'utilisation de ce type de relations parce que nous considérons que c'est le moyen le plus sûr permettant d'éviter la déperdition du sens et de relever toutes les nuances dans un contenu en rapport avec le contexte. En effet, nous considérons que si un mot est important et qu'il doit être retenu dans un contexte, il en sera de même pour les mots qui lui sont liés, tels que ses synonymes, ses antonymes, ses converses, ses génériques, ses spécifiques, etc. La spécificité de chaque type de relations lexico-sémantiques est gérée dans les différents algorithmes de la projection. Outre les mots simples, notre ontologie contient des mots composés. Ils sont codés ou indexés pour que la machine puisse les extraire facilement et leur réserver un traitement spécial. Cette démarche permet d'éviter toute ambiguïté au niveau de l'interprétation. De par sa nature, cette approche peut être utilisée dans beaucoup de domaines liés à la notion du contexte comme l'indexation « sémantique » des données, la recherche linguistique ou sémantique d'information, l'extraction contextuelle d'information, la gestion de corpus multilingues, la génération automatique de textes, etc.

Le nombre des relations lexico-sémantiques étant important, une soixantaine environ, nous essayons, à travers cette analyse, de ne prendre en compte (dans un premier temps) que certaines relations pour valider notre approche. Bien évidemment, l'extension à d'autres relations est possible suivant le même principe.

Dans cette recherche, nous expérimentons le cas de la langue amazighe. La raison pour laquelle nous avons développé et peuplé, un modèle de contexte basé sur une ontologie du domaine de la faune et de la flore, enrichie par des relations lexico-sémantiques. Dans ce qui suit, nous présentons une analyse des motivations des

choix de conception du modèle, ainsi qu'un aperçu de l'implémentation de l'ontologie[1].

Ontologies

Définition

D'une manière générale, les ontologies reposent sur des outils de modélisation et de représentation des connaissances permettant à des communautés d'experts humains et logiciels d'un domaine donné de partager, d'une manière consensuelle, un vocabulaire et de parler un même langage, c'est-à-dire communiquer et se comprendre. Définir une ontologie d'une manière précise est une tâche très difficile. Plusieurs définitions existent suivant le contexte d'utilisation. Celle qui fait autorité aujourd'hui, au sein de la communauté scientifique, est celle de Gruber qui l'a définie comme suit :

« Une ontologie est la spécification d'une conceptualisation d'un domaine de connaissance ». En d'autres termes, une ontologie est un modèle de représentation de la formalisation d'une conception d'un domaine. La modélisation d'un domaine repose essentiellement sur deux aspects importants, la représentation des connaissances et le raisonnement qui peut leur être associé. La partie description permet de décrire et de formaliser le domaine en utilisant les notions de : classes, instances, attributs, relations, fonctions (ensembles de relations : simplification), restrictions (conditions sur certains éléments), etc. La partie raisonnement, quant à elle, décrit les aspects dynamiques liés à une ontologie, à travers des règles (règles mettant en valeur les éléments de l'ontologie), et des événements (événements modifiant certains attributs ou relations), etc.

Quelques éléments constitutifs

Nous présentons dans ce qui suit des éléments constitutifs de l'ontologie étudiée.

1. Concepts et instances

Les concepts représentent des objets ou des idées. Ils sont organisés en taxonomie au sein d'un réseau de concepts et structurés hiérarchiquement. Chaque concept est caractérisé par un ensemble de propriétés :

[1] Concernant la notation des entrées lexicales, nous utiliserons le protocole suivant :

- **Voyelles** : *a, i, u* et *e* pour noter le schwa.

- **Semi-voyelles** : *w, y*.

- **Consonnes** : *p, b, t, d, k, g, l, m, n, s, z, š, ž, ḥ/ϵ* notent les fricatives pharyngales sourde et sonore, *x/ġ* les fricatives vélaires sourde et sonore, *h* la spirante, *q* l'occlusive dorso-uvulaire, *r* la vibrante apicale. Le point sous la lettre indique l'emphase, le *w* en exposant note la labiovélarisation, le trait sous la lettre note la spirantisation, le dédoublement de la consonne indique la gémination.

- un concept est *générique* s'il exclut toute extension. Dans nos corpus, aucun concept ne relève de cette catégorie.

- un concept portant une propriété d'*identité* permet de différencier deux instances de ce concept.

- un concept est *rigide* s'il ne peut pas être une instance d'autres concepts. Exemple : *amuder* « être vivant » :

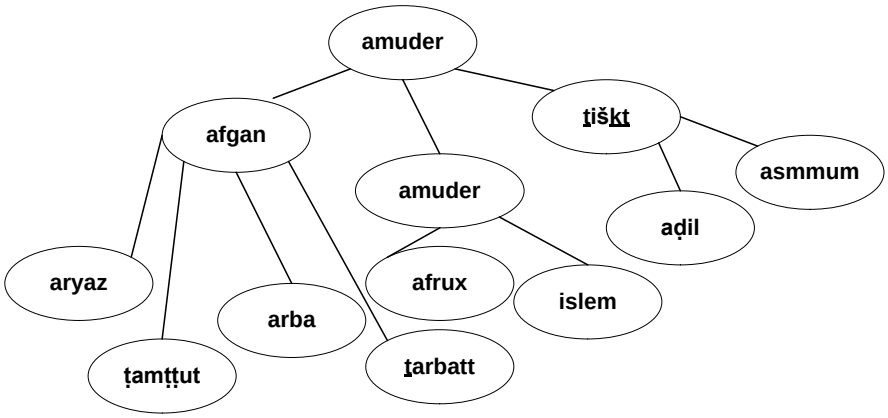


Figure 1 : Concept rigide

- un concept est *anti-rigide* lorsqu'il peut être une instance pour d'autres concepts. Dans l'exemple ci-dessus « afgan», « amuder » et « ṯišḵt » sont des concepts anti-rigides.

Alors que le concept désigne l'intention du concept, l'instance renvoie à l'élément de l'ensemble constituant l'extension de concept.

2. Relations entre concepts

Au sein d'une ontologie, les concepts sont liés entre eux par des relations qui sont définies comme les liens entre des entités. Ces liens sont classés en deux catégories :

(1) Les *relations hiérarchiques* lient des éléments supérieurs dits « hyperonymes » et des éléments inférieurs dits « hyponymes ». De ce fait, l'hyperonyme englobe l'hyponyme. Exemple :

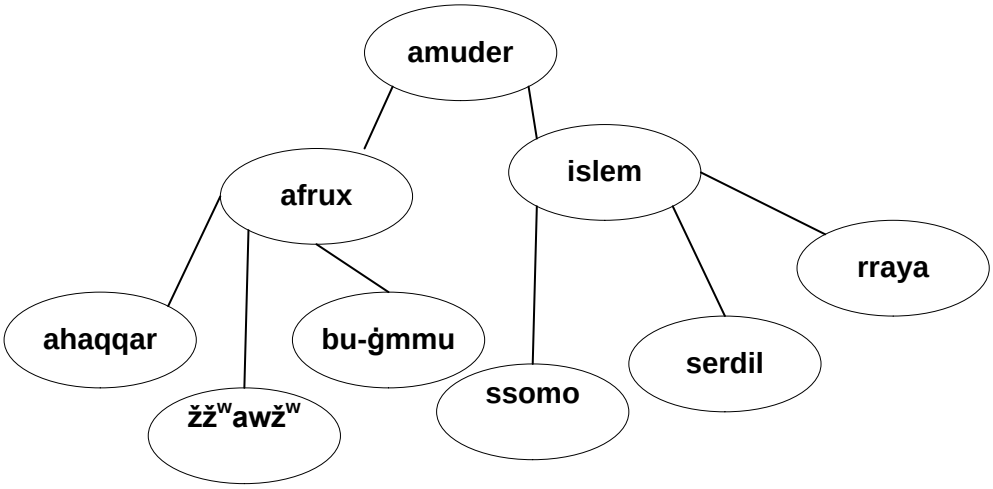


Figure 2 : Relations entre concepts

Ici, *amuder* est l'hyperonyme de *afrux* et *islem* ; *afrux* et *islem* sont les hyponymes de *amuder*.

afrux et *islem*, par rapport aux autres entités, peuvent, à leur tour, être considérés comme leurs hyperonymes.

A ne pas oublier de mentionner que *serdil* « sardine » dans certaines régions comme la région des Ait Wirra à Elksiba (Maroc), joue le rôle de l'hyperonyme (générique), dans ce sens qu'il renvoie à tout type de poisson.

(2) Les *relations sémantiques* correspondent à la structuration d'holonymie-méronymie. Elles peuvent être également considérées comme une relation hiérarchique liant un couple de concepts dont l'un dénote une partie de l'autre et l'autre comme un tout. Ce genre de relation est différent de la catégorie (1) dans la mesure où l'holonyme dispose des propriétés qui ne sont pas obligatoirement transmises à ses parties. Exemple : *aqbu* (tronc) est une partie de *tišk̲t* « arbre » mais n'est pas une sorte de *tišk̲t*. On peut représenter cette relation comme suit (1) faune, (2) flore :

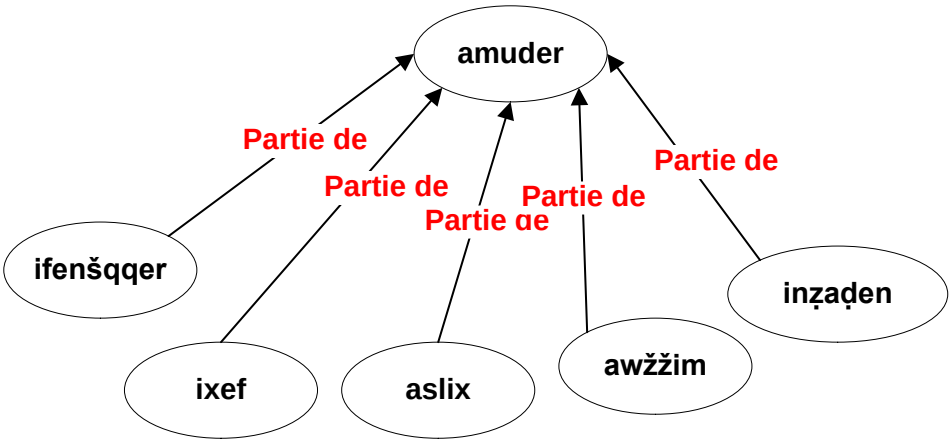


Figure 3 : Relations d'holonymie-méronymie

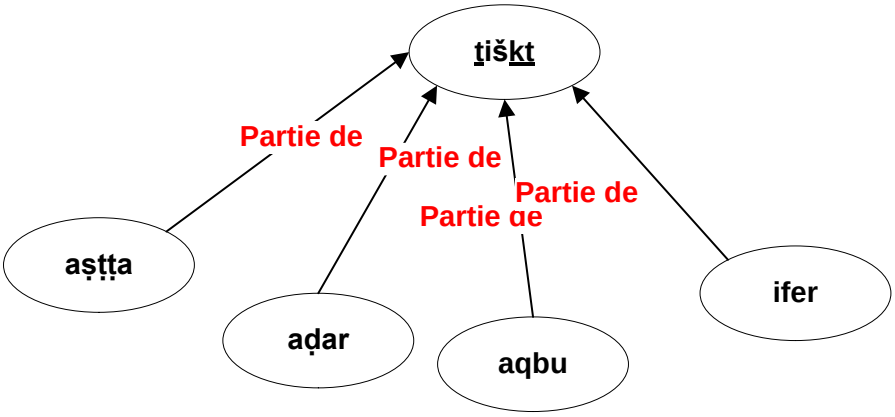


Figure 4 : Relations d'holonymie-méronymie

Outre ces relations sémantiques, les relations synonymiques et antonymiques, qui relèvent des liens horizontaux, seront également abordées dans l'élaboration de notre ontologie. Exemple : *iydi, buzaher, anmmuṯer, amḥḍaw* « chien » ; *iddew, abaġus* « singe ».

3. Propriétés

Ce sont des informations rattachées à chaque nœud du réseau sémantique. Exemple :

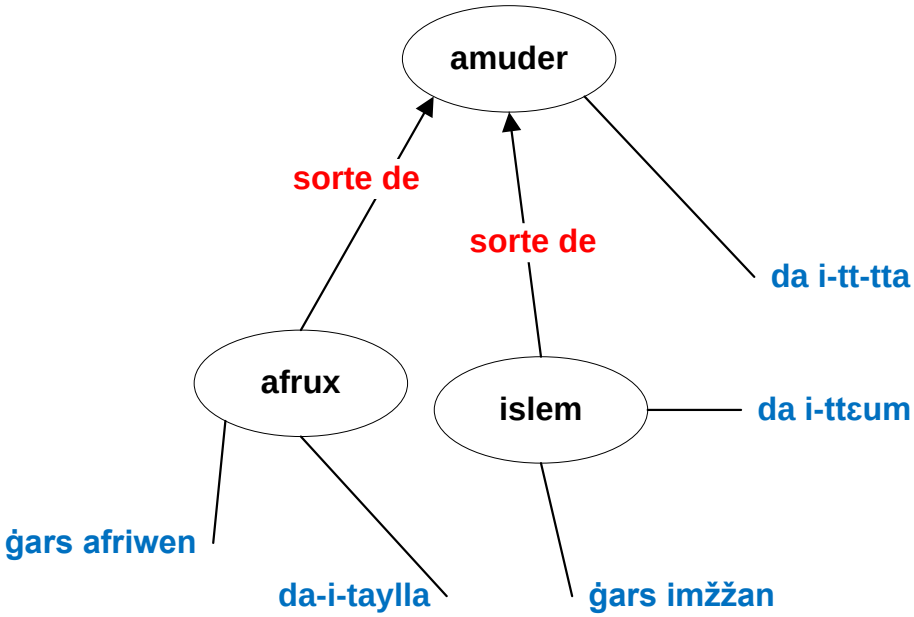


Figure 5 : propriétés

4. Attributs

Ce sont des relations qui relient un nœud concept à une valeur ou propriété. Ils définissent la structure de données. Exemple :

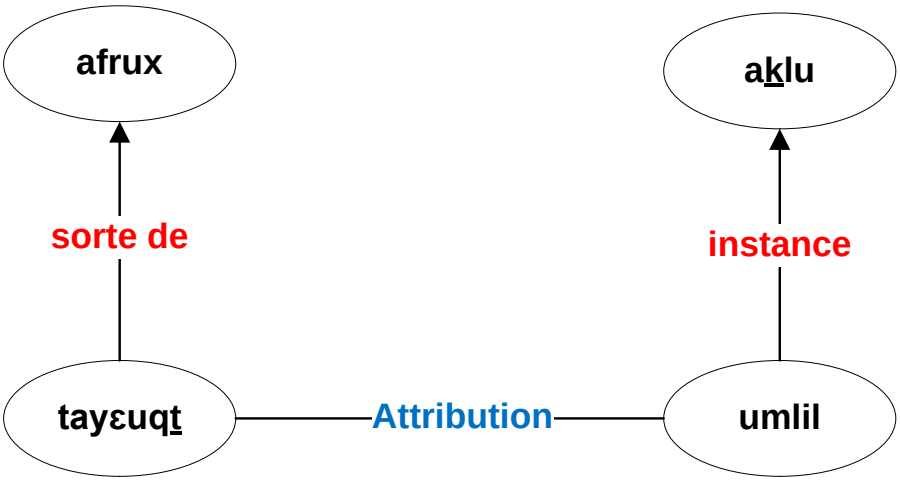


Figure 6 : Attributs

5. L'héritage

Il repose sur des liens de type « sorte de » ou « est un » qui relient un concept à un autre concept plus élevé. Ces liens sont étroitement liés à l'expression linguistique de la copule « est ». Ici, ils expriment soit l'appartenance d'un objet à une classe, soit l'inclusion. Exemple :

« *tayɛuqt* » est une sorte de « *afrux* ».

« *tayɛuqt* » est un « *afrux* ».

6. Types

En se basant sur certains travaux effectués dans le domaine de la classification d'ontologie (GómezPérez, 1999), (Guarino, 1997b), (Mizoguchi, 1998), (MizoguchiIkeda, 1996), (VanHeijstAl, 1997), (VanwelkenhuysenAl, 1994), (VanwelkenhuysenAl, 1995), (WieliingaSchreiber, 1993), etc., on peut en déduire la classification (dépendante des besoins d'utilisation) suivante :

- L'ontologie de haut niveau définit et représente les concepts de haut niveau qui sont des concepts de l'univers de modélisation commun aux ontologies des niveaux inférieurs, comme la notion du temps, de l'espace, etc. Son rôle est de réduire les ambiguïtés des termes entre les différentes ontologies et utilisations.
- L'ontologie générique se situe d'un point de vue abstraction entre l'ontologie de haut niveau et l'ontologie du domaine. Ses concepts sont

moins abstraits que l'ontologie de haut niveau et plus génériques par rapport à ceux de plusieurs ontologies de domaines, réutilisables par plusieurs domaines.

- L'ontologie de domaine représente et décrit les concepts se rapportant à un domaine donné.
- L'ontologie de tâches représente et décrit le vocabulaire propre à une activité décrivant une structure de résolution d'un problème.
- L'ontologie d'application représente et décrit les concepts propres à une application. Dans la plupart des cas, ces concepts proviennent de l'ontologie du domaine et de l'ontologie des tâches.
- L'ontologie de représentation représente et décrit les primitives des langages de représentation et de formalisation de connaissances : RDF, OWL, réseau sémantique, LD, LPO, etc.

7. Les langages

Les langages constituent une partie très importante des systèmes à base d'ontologies ; ils permettent la gestion de la base de connaissances et de raisonnements qui peuvent leur être associés. Les langages présentés dans ce paragraphe sont des recommandations du W3C, basés sur des normes, standards, etc. ayant pour but de permettre des traitements automatiques sur des contenus par des applications différentes. Nous verrons, un peu loin, que ceci fait de la notion d'ontologie un principe fondamental pour le partage des connaissances, la communication et la coopération au sein des systèmes distribués. Ci-après, nous reproduisons quelques éléments de description de ces langages, extraits de la pyramide du Web sémantique de Berners-Lee.

URI/IRI : les URIs (Uniform Resource Identifier) permettent d'identifier d'une manière unique les ressources sur le WEB et les IRIs (Internationalized Resource Identifiers) permettent aux personnes d'identifier des ressources Web dans leur propre langue.

XML : Extended Markup Language (XML) est un langage de description et d'échanges de documents et des données ne permettant pas leur présentation, appelé aussi format d'échanges standardisé. Il permet aux applications reconnaissant ce format d'échanger tous les types de données décrits dans ce langage, utilisé souvent pour assurer la compatibilité des données entre les applications hétérogènes. Exemple, on peut décrire la voiture 123 de la marque Renault et de couleur rouge.

RDF : Resource Description Framework (RDF) est un métalangage qui sert à décrire les ressources, leurs propriétés et les valeurs des propriétés sous forme d'un graphe (ressource, propriété, valeur). Il est considéré comme un modèle standardisé de description des métadonnées qui peut être défini et associé à des documents ou à des contenus. Ces annotations et métadonnées permettent d'associer du sens à des contenus qui peuvent être traitées d'une manière automatique par les agents logiciels. Pour la gestion sémantique des données, les métadonnées RDF peuvent être des informations sémantiques associées à des mots du texte. Ces éléments peuvent être ensuite analysés et interprétés pour l'extraction du sens global. Il est à

noter que RDF peut être exprimé dans plusieurs langages, mais c'est XML qui est souvent utilisé, une version XML de RDF appelée RDF/XML a été même créée.

RDFS : RDF peut également être utilisé pour décrire des situations et des utilisations particulières avec un vocabulaire bien précis en utilisant la notion de RDF Schéma (RDFS). RDFS consiste à adapter RDF à des domaines modélisés particuliers décrivant des utilisations particulières au sein d'une communauté. La définition d'un schéma RDF consiste en une activité de typage et de classification des ressources, des propriétés et des relations sous forme de classes définies dans des « espaces de noms » servant principalement à désambigüiser les mêmes éléments d'un vocabulaire définis dans des utilisations ou espaces de noms différents.

OWL : c'est une extension de RDF enrichie avec des propriétés sémantiques, de contraintes, de comparaisons, de cardinalités, etc. pour décrire et manipuler les ontologies. C'est un langage recommandé par le W3C basé, comme RDF, sur XML qui permet à des moteurs d'inférences d'agents l'interprétation et le raisonnement automatiques sur les ontologies. En effet, OWL est basé sur les logiques de description utilisées dans les systèmes de représentation de connaissances et offrant de fortes possibilités de manipulation de prédicats de classes, de rôles et d'individus, donc d'ontologies, contrairement aux logiques de premier ordre classiques ne manipulant que des objets de même type.

Il existe trois versions d'OWL :

OWL Full : c'est la version la plus complète, elle combine toutes les primitives d'OWL avec RDF/S sans respecter aucune contrainte si ce n'est celle de RDF, cette version n'est pas décidable.

OWL DL : c'est un sous-ensemble d'OWL Full basé sur la logique de description n'autorisant que des constructions garantissant la décidabilité des inférences.

OWL Lite : c'est la version la plus simple, elle sert principalement à la création de hiérarchies de classes. Elle est dotée seulement de quelques propriétés de classes comme la comparaison, la restriction, la cardinalité (0 ou1).

SPARQL (Protocol and RDF Query Language) : langage d'interrogation des ontologies représentées sous forme de graphes RDF/S. Il est pour les bases de connaissances RDF ce que SQL est pour les bases de données relationnelles. *Exemple : SELECT ?x, ?y, ?z FROM URI/IRI WHERE {Conditions1, condition2, etc.}*

RIF (Rule Interchange Format) : format d'échange de règles standardisé. Il permet de faciliter les échanges de règles utilisables par des systèmes distribués sur le WEB en assurant l'interopérabilité et la portabilité entre divers langages et moteurs de règles.

SWRL (Semantic Web Rule Language) : langage de raisonnement sémantique à base de règles sur les ontologies. C'est une combinaison d'OWL-DL et RuleML language créé principalement pour le Web sémantique pour pouvoir développer des règles sémantiques au niveau des agents. Il permet, contrairement à OWL, de manipuler les instances par des variables, de définir des fonctions mathématiques, des types de données, etc. *Exemple : avec OWL, la relation oncle ne peut être*

définie que comme suivant : intersectionOf(SubClassOf(Homme), estfrereDe(Pere), avec SWRL on peut la définir au niveau des instances représentées par des variables x,y,z comme suivant : Personne(?x) ^ Personne(?y) ^ Personne(?z) ^ pere(?x,?y) ^ frere(?x,?z) => oncle(?z,?y). Cela permet de définir qui est l'oncle de qui, impossible à définir avec OWL. SWRL se différencie, en plus, d'OWL par le fait qu'il ne peut créer ni de nouveaux concepts ni de nouvelles relations, excepté ceux créés par la manipulation des variables et par la satisfaction des règles d'inférences.

Le choix des fonctions lexicales et des relations lexico-sémantiques

Les relations sémantiques et conceptuelles

Ci-après, nous présentons un bref aperçu sur les relations sémantiques et conceptuelles qui nous ont permis d'introduire les relations lexico-sémantiques utilisées dans notre démarche.

Les relations sémantiques sont les liens entre les éléments du lexique dans le texte. Elles représentent les liens de sens que peuvent entretenir deux ou plusieurs mots par rapport à leurs significations, comme les relations de type synonymie, antonymie, hyperonymie, etc. Les relations conceptuelles sont des relations ou associations utilisées dans la modélisation informatique. Elles permettent de faire des liens entre les classes et les instances. Dans le cas de la modélisation ontologique, nous pouvons bien sûr utiliser ce genre de relation, mais ce sont les relations de types hiérarchie et classification : héritage, classification comme « is-a, sort-of, part-of, etc. » qui sont privilégiées et souvent utilisées.

Dans la littérature, une distinction est faite entre les relations sémantiques Ahmad et Fulford (1992) et les relations conceptuelles Sager (1990) et Condamines et Rebeyrolle (1998). Dans L'Homme (2004), on explique la différence entre les relations conceptuelles et les relations sémantiques. Dans d'autres travaux, on explique aussi que certaines relations sémantiques sont très proches des relations de représentation conceptuelle : c'est le cas, notamment, de celles qui interviennent dans les taxinomies et les méronymies. Ce type de relation concerne, le plus souvent, des termes qui renvoient à des entités.

On peut conclure qu'il est extrêmement difficile de faire une réelle différence entre ce qui relève du conceptuel et ce qui est lié au sémantique. Les relations lexico-sémantiques définies dans le cadre des fonctions lexicales de Mel'čuk est un modèle de relations complet intégrant au moins les deux types de relations définies précédemment. C'est une des raisons qui nous a permis de les adopter dans notre démarche.

Les relations lexico-sémantiques

C'est suite à ces différentes analyses et expériences menées dans ce domaine que le choix de l'utilisation des relations lexico-sémantiques définies par les fonctions lexicales (Mel'čuk 1988), (Mel'čuk et *al*., 1995) et (Polguère, 2003) s'est imposé. Car elles permettent de représenter tous les types de relations dont nous avons

besoin : sémantiques, ontologiques, lexicales, etc. Nous considérons que cela permet une modélisation fine de tous les traits sémantiques améliorant le processus d'analyse des textes et de détection de la sémantique à travers une ontologie de contexte. En effet, l'utilisation de ces relations est importante. Elle permet, dans le cas de la recherche ou de l'extraction d'information, par exemple, d'éviter d'ignorer des mots ou des phrases bien qu'ils soient pertinents par rapport à un contexte (silence). Elle permet également d'éviter d'extraire des mots ou des phrases non pertinentes (bruit). On considère que si un mot est pertinent dans un contexte, alors tous les mots que l'on peut atteindre *via* les relations lexico-sémantiques sont aussi importants que le mot lui-même.

Suivant le nombre des relations utilisées, la probabilité qu'un mot et ceux qui lui sont sémantiquement liés soient considérés est supérieure à la probabilité ne tenant compte que du mot seul. L'utilisation de ces relations peut donc avoir un impact très positif sur l'amélioration des performances des applications où elles peuvent être utilisées.

Fonctions lexicales

Comme il a été mentionné ci-dessus, le présent travail porte sur l'élaboration d'un modèle d'ontologies de domaine enrichie par des liens lexico-sémantiques associés aux fonctions lexicales de la Théorie Sens-Texte (Mel'cuk, 1997) et la notion du contexte, dans un but d'améliorer l'analyse et la détection sémantique des données dans les documents de type textes. La Théorie Sens-Texte est un cadre théorique linguistique, développée à Moscou par Aleksandr Žolkovskij et Igor Mel'čuk, pour la construction de modèles du langage naturel Mel'čuk (1981) et (1988). C'est une des théories majeures pour le Traitement Automatique des Langues fournissant des bases solides pour beaucoup de domaines d'application liés au traitement automatique de la sémantique, comme la traduction automatique, la recherche sémantique ou linguistique, l'extraction d'information, etc. Elle est basée sur une formalisation globale de la langue utilisant des fonctions mathématiques définissant les fonctions lexicales, comme suivant :

Une fonction lexicale (FL) permet de définir la description et la modélisation des relations lexicales, de collocations et de dérivation sémantique entre les unités lexicales (UL) d'une langue. Elle se présente sous forme d'une fonction au sens mathématique *: **f** (L) = {L1, L2…Ln} où f = le nom de la FL, L = lexie qui est l'argument de la FL et {L1, L2…Ln} = ensemble des lexies qui constituent la valeur de la FL auprès de L*, cf. Žolkovskij & Mel'čuk 1967, Mel'čuk1974, 1996, 1998, 2003, 2007, et Wanner (éd.) 1996.

Voici un extrait de la liste des fonctions lexicales :

Syn, Anti, $Conv_{ij}$, Contr, Epit, Gener, Figur, S_0, A_0, V_0, Adv_0, S_i, S_{instr}, S_{med}, S_{mod}, S_{loc}, S_{res}, $Able_i$, $Qual_i$, Sing, Mult, Cap, Equip, A_i, Adv_i, Imper, Result, Centr, Magn, Plus, Minus, Ver, Bon, Pos_i, Loc_{in}, Loc_{ab}, Loc_{ad}, Instr, Propt, Copul, Pred, $Oper_i$, $Func_i$, $Labor_{ij}$, Incep, Cont, Fin, Caus, Perm, Liqu, $Real_i$, $Fact_i$, $Labreal_{ij}$, Involv, Manif, Prox, $Prepar_i$, Degrad, Son, $Obstr_i$, $Stop_i$, $Excess_i$, $Sympt_{ijk}$.

Parmi toutes les fonctions lexicales, deux groupes sont à distinguer : les FL PARADIGMATIQUES et les FL SYNTAGMATIQUES.

Dans ce qui suit, on trouvera un bref aperçu sur quelques fonctions, les principales sur lesquelles des tests ont été effectués dans le cadre de ce travail. Une étude détaillée de toutes les fonctions ainsi que leur intégration dans le processus d'analyse et de détection est prévue dans un travail ultérieur.

Les fonctions lexicales paradigmatiques permettent de définir les liens entre les lexies et la manière dont l'une peut être atteinte à partir de l'autre. Elles correspondent à des relations à la fois sémantiques et formelles entre deux mots. Exemples :

- **Syn**, qui met en relation une entité avec ses parasynonymes : *Syn(iydi)=buzaher, amħdar.*

- **Anti**, qui met en relation une entité avec ses antonymes : *Anti(amsksum)=amstuyya,*

- **Conv**, qui prend en considération les relations de conversions entre deux entités : *conv(isġa)=izzenza.*

La notion de contexte et l'ontologie de contexte

Le contexte

La notion de contexte constitue un élément fondamental dans plusieurs domaines, notamment dans le domaine du Traitement Automatique du Langage Naturel (TALN), l'analyse sémantique de données, la gestion de connaissances (Semantic Data Mining & Knowledge Management : SDMKM), etc. où l'on utilise plus particulièrement au niveau de la désambiguïsation et de l'interprétation automatiques de la sémantique des données. On peut le définir comme un ensemble de situations, où chaque situation est composée d'un domaine, d'un ensemble de contraintes d'utilisation et d'un profil utilisateur. Il peut être représenté par un ensemble de propriétés décrivant des situations d'utilisations particulières et un vocabulaire associé au domaine modélisé. Ces éléments peuvent être partagés au sein d'une même communauté et représentés sous forme d'une ontologie échangeable entre les différents agents humains et logiciels. Ce qui leur permet de parler un même langage et d'interpréter les éléments de la même manière.

Concrètement, le contexte est composé :

- D'un domaine défini par un schéma représentant une activité de typage et de classification des ressources, des propriétés et des relations sous forme de classes dans des « espaces de noms » servant principalement à désambigüiser les mêmes éléments d'un vocabulaire par rapport à des utilisations ou espaces de noms différents.
- Contraintes définies par des règles représentant les contraintes d'une utilisation particulière, comme par exemple inclure et exclure des éléments qui seraient respectivement exclus et y inclus du domaine, cela correspond à un réglage fin supplémentaire de l'ontologie du domaine. Des éléments

issus de la théorie du discours peuvent être intégrés dans cette partie. On peut citer dans ce cas-là et à titre d'exemple, l'interprétation de certains éléments suivant leur auteur, le message qu'on souhaite extraire du texte, etc.

- Profil utilisateur qui est un élément important de la notion de contexte. Cela peut correspondre dans le cas de l'analyse du texte à ce qui est perçu par le lecteur suivant ses compétences, expériences, etc.

Une vue synthétique du méta-modèle de la notion de contexte est représentée dans le graphe suivant.

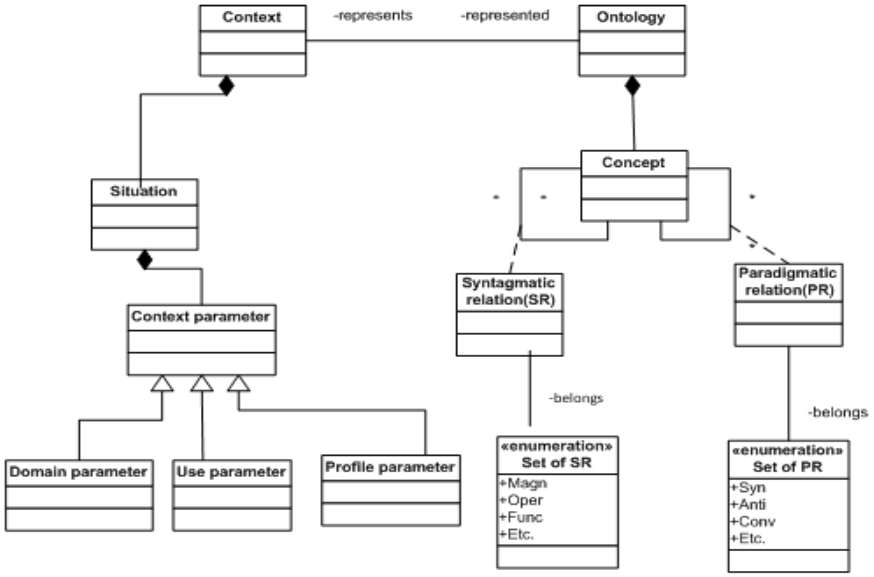


Figure 7 : Méta-modèle du contexte

L'ontologie de contexte

Dans cette étude, nous avons conçu et développé une ontologie du contexte intégrable dans le processus général de l'analyse sémantique et contextuelle du texte, proposé dans un travail parallèle (H.Fadili ACS/IEEE, 2013). C'est une ontologie du domaine que nous avons enrichie par la notion de contexte et les relations lexico-sémantiques décrites précédemment. En effet, étant donné que l'ontologie du domaine est seulement une ontologie conceptuelle du domaine étudié, nous avons estimé qu'il est indispensable, pour l'analyse sémantique à proprement dit, d'enrichir l'ontologie du domaine classique par des relations lexico-sémantiques permettant de prendre en compte, outre les relations conceptuelles entre les concepts, les relations lexico-sémantiques. Cette démarche rend possible une analyse linguistique de bas niveau entre les mots dans le contexte. Elle peut être considérée comme une ontologie du domaine fusionnée avec une ontologie linguistique, définie par deux types de liens : liens verticaux (contextualisés) qui sont des liens de catégorisation, et les liens horizontaux (décontextualisés) sont des liens linguistiques (relations lexico-sémantiques, actions, verbes, etc.). Tous ces éléments pris en compte, sont enrichis par des contraintes d'utilisation.

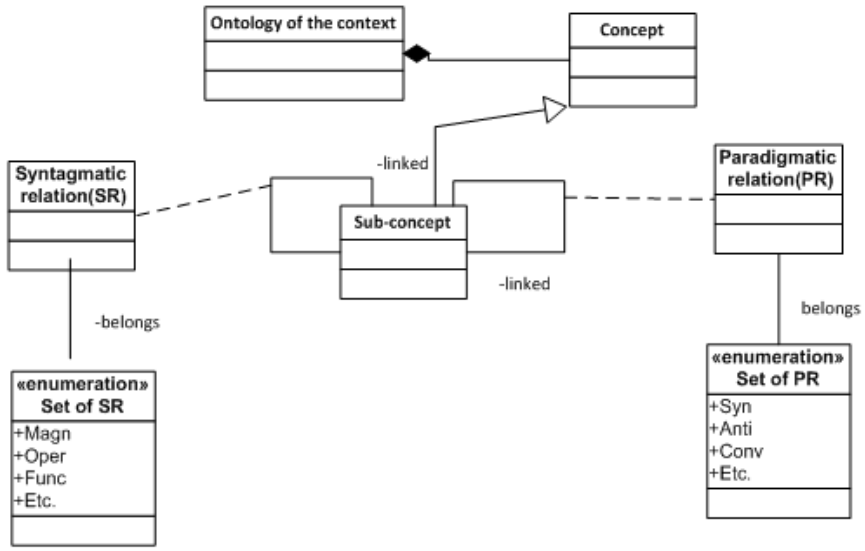


Figure 8 : Méta-modèle de l'ontologie du contexte

Constitution du corpus

D'une manière générale, la constitution d'une ontologie est une tâche très difficile ; cette tâche est encore plus difficile dans le cas de l'amazighe. Dans cette partie, l'objectif est de relever les exigences et les problèmes à résoudre pour la création

d'une telle ontologie en amazighe. Parmi lesquelles, on cite, les problèmes classiques liés à la création et à la gestion du corpus, d'une manière générale, puis ceux liés à la création et à la gestion de corpus amazighs, en particulier. A mentionner que la définition des termes scientifiques dans l'ontologie qui consiste dans la recherche, la création ou la réutilisation de termes scientifiques ou savants par rapport à un domaine ou à une discipline donnée n'est pas sans poser problème.

Dans ce qui suit, nous présentons les difficultés rencontrées, lors de la mise en place de l'ontologie scientifique faune-flore, en amazighe :

1. Nous nous sommes rendus compte que des données sources, que ce soit des données brutes y compris papier, ou des données électroniques, sont très rares :

Absence de sources de données, seulement quelques publications.

2. D'autres problèmes sont liés à la création de nouveaux contenus :

- Mots, concepts, etc.,
- Relations,
- Lexique scientifique,
- Problèmes de traduction,
- Traduction des mots scientifiques,
- Création de liens et des correspondances avec d'autres langues.

3. Applications avancées supportant le Tifinagh

- Absence d'applications totalement multilingues, problèmes rencontrés avec les éditeurs d'ontologies dans le Cloud2 par exemple.

4. Copyright

- Manque de données ouvertes et libres d'accès,
- Non institutionnalisation de la langue, ne permet pas la création de nouvelles données publiques.

Quelques solutions de contournements

Au niveau du contenu

Concernant la définition des concepts, nous nous sommes basés sur notre connaissance de la langue amazighe, sur des dictionnaires spécialisés et sur d'autres sources de données existantes.

Pour ce qui est de la définition des termes, il a été procédé ainsi :

1. Etant donné que les termes traités sont des entités du monde réel, nous avons défini chaque mot connu par ses relations et ses propriétés avant de l'insérer dans l'ontologie.

[2] Le cloud ou le cloud computing (le nuage en Français) est une technologie permettant la mutualisation des services et ressources informatiques ouverts en libre service.

2. Un mot non connu est traduit à partir d'autres langues. Dans certains cas, on a procédé par l'emprunt des termes les plus utilisés dans le langage courant : darija, arabe standard ou français, etc.

Au niveau technologique

Afin de rester conforme aux normes internationales, nous avons utilisé trois catégories de relations :

- les versions anglaises des relations de hiérarchisation OWL,
- les relations lexico-sémantiques dans leur version normalisée,
- puis les relations en amazighe pour les autres.

Ces règles ont été utilisées pour les propriétés (les propriétés hors la norme OWL).

Pour la graphie, faute de compatibilité avec le Tifinagh, nous avons opté, quant à cette première version de l'ontologie, pour la transcription phonétique. Une version en Tifinagh est prévue dans les versions futures.

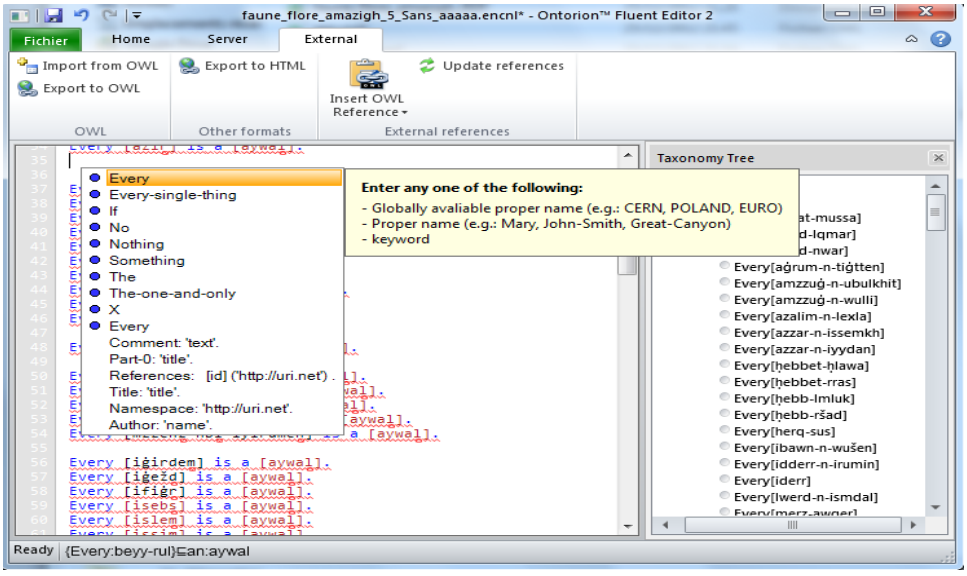


Figure 9 : Editeur facile

Afin d'accélérer la saisie, nous avons contourné, dans un premier temps, l'utilisation directe de l'interface Knoodl qui peut paraître difficile, nous avons utilisé l'application et la notation « *Fluent éditeur* » basé sur l'anglais contrôlé pour saisir facilement les données de l'ontologie, avant de les convertir en OWL et les importer dans la plateforme finale de travail (knoodl).

La saisie de fait en respectant la grammaire du CNL pour l'anglais comme suivant :

(1) Every *abaġus* is an *amuder*.
Every *abrriḍ* is an *amuder*.
Every *abrriḍ-n-taydwin* is an *amuder*.
Every *abulxir* is an *amuder*.
Every *aḍil-n-wušen* is an *amuder*.
Every *afrux* is an *amuder*.
Every *afullus* is an *amuder*.
Every *aġyul* is an *amuder*.
Every *aġʷilas* is an *amuder*.
Every *aġerḍay-n-lxla* is an *amuder*.
Every *aḥuliy* is an *amuder*.
Every *ašiṯar* is an *amuder*.
Every *ayis-n-lbḥer* is an *amuder*.
(2) Every *azmmur* is an *tiškt*.
Every *ṯasaft* is an *tiškt*.
Every *ṯasmmumt* is an *tiškt*.
Every *ṯašṯa* is an *tiškt*.
Every *ṯaylilut* is an *tiškt.*
Every *ṯiqqi* is an *tiškt.*

Après cela, on procède à un export en RDF, facilement intégrable de dans la *plateforme finale.*

```
- <Class rdf:about="http://ontorion.com/unknown.owl/a3sat-mussa">
      <rdfs:subClassOf rdf:resource="http://ontorion.com/unknown.owl/ašklu"/>
  </Class>
      <!-- http://ontorion.com/unknown.owl/a3ud-lqmar -->
- <Class rdf:about="http://ontorion.com/unknown.owl/a3ud-lqmar">
      <rdfs:subClassOf rdf:resource="http://ontorion.com/unknown.owl/ašklu"/>
  </Class>
      <!-- http://ontorion.com/unknown.owl/a3ud-nwar -->
- <Class rdf:about="http://ontorion.com/unknown.owl/a3ud-nwar">
      <rdfs:subClassOf rdf:resource="http://ontorion.com/unknown.owl/ašklu"/>
  </Class>
      <!-- http://ontorion.com/unknown.owl/abaġus -->
- <Class rdf:about="http://ontorion.com/unknown.owl/abaġus">
      <rdfs:subClassOf rdf:resource="http://ontorion.com/unknown.owl/aywal"/>
  </Class>
      <!-- http://ontorion.com/unknown.owl/abrriḍ -->
- <Class rdf:about="http://ontorion.com/unknown.owl/abrriḍ">
      <rdfs:subClassOf rdf:resource="http://ontorion.com/unknown.owl/aywal"/>
  </Class>
      <!-- http://ontorion.com/unknown.owl/abrriḍ-n-taydwin -->
- <Class rdf:about="http://ontorion.com/unknown.owl/abrriḍ-n-taydwin">
      <rdfs:subClassOf rdf:resource="http://ontorion.com/unknown.owl/aywal"/>
  </Class>
      <!-- http://ontorion.com/unknown.owl/abulxir -->
- <Class rdf:about="http://ontorion.com/unknown.owl/abulxir">
      <rdfs:subClassOf rdf:resource="http://ontorion.com/unknown.owl/aywal"/>
  </Class>
      <!-- http://ontorion.com/unknown.owl/adar-n-ufullu -->
- <Class rdf:about="http://ontorion.com/unknown.owl/adar-n-ufullu">
      <rdfs:subClassOf rdf:resource="http://ontorion.com/unknown.owl/ašklu"/>
  </Class>
```

Figure 10 : Version RDF

Eléments sur l'implémentation

Eléments de l'ontologie

Ci-dessous une présentation illustrative de cette méthode. Dans un premier temps, nous présentons des relations simples des êtres et dans un deuxième temps des relations complexes :

1. La hiérarchie simple des êtres. On parle ici des relations qui « spécifient les relations verticales de l'ontologie, et se résument à des formes de subsomption, de l'élément par la classe (relation is a) ou de la partie par le tout (relation has a). Ces relations sont déterminantes, car toute ontologie a une fonction unificatrice, l'Etre étant défini par son unité à soi » (Rastier 2004).

Exemple :

1. Relations de base

Every *izem* is an *amuder*.

Pour aller plus loin.

Every *aserdun* is an *amuder*.
Every *amuder amsksum* is an *amuder*.
Every *amuder amstuyya* is an *amuder*.
Every *tafunast* is an *amuder amstuyya*

On traite également la flore.

Every *azmmur* is a *tiškt*
Every *tasmmumt* is a *tuyya*
Every *tazdayt* is a *tiškt*
Every *bu-ħmmu* is a *tuyya*
Every *alili* is a *tiškt*
Every *tamemt n waman* is a *tiškt*
Every *tišiṭ* is a *tiškt*

2. attributs et specifications

Every tiškt ittuyatšan is a tiškt

3. Disjointes

Every *amuder* is not a *tiškt*.
Every *tixsi* is not a *aserdun*.
Every *tiškt* is not a *amuder*.

4. Les relations simples

Every *izem da-i-tt-tta* a *tamlalt*.
Every *imiššəw da-i-tt-tta* an *aġərḍay*.

Every *tafunast da-tt-tta a tuyya.*
Every *iydi da-i-tt-tta an aksum.*

5. Les relations complexes

Every *lfṣṣa* is something that *i-tt-tta* an *amuder* and *da-i-tt-tta timzin* and *da-t-tt-tta* a *agʷlas*.
If X *i-yya zi* something that *i-ttu-yyan zi* Y then X is *-i-yya zi* Y.

6. Equivalence

Something is a *amuder amsksum* if-and-only-if-it *da-i-tt-tta* nothing-but *amuder* and-or *da-i-tt-tta* nothing-but thing that *i-yyan zi* an *amuder*.
Something is an *anarmu* if-and-only-if-da-i-tt-tta an *amuder* and *i-tt-tta* an *tiškt* and *i-tt-tta* a thing that *i-yyan zi* an *amuder* and-or *i-yyan-zi* a *tiškt*.
Something is a *amstuyya* if-and-only-if-it *i-tt-tta* nothing-but *tiškt* and-or *i-tt-ttan* nothing-but thing that *i-yyan zi tiškt*.

7. Disjointnes

Anything either is an *anarmu*, is an *amstuyya* or is an *amsksum* or-something-else.

8. Relations entre les entités

X *i-yyanzi* Y if-and-only-if Y *i-ttuyya zi* X.
X *da-i-tt-tta* Y if-and-only-if Y *da-i-ttuyatša zi* X.

Spécifications techniques

Pour la mise en place de l'ontologie, nous avons utilisé la plateforme knoodl. C'est une solution de gestion d'ontologies dans le Cloud. Elle est dotée de plusieurs services permettant une gestion collaborative d'ontologies, de gestion de communautés, des wiki, des liens avec d'autres ontologies et données, etc. Ce choix a été motivé, entre autres, par le fait que cette technologie permet de profiter d'une part de la technologie du Cloud ne nécessitant aucune installation ni maintenance de solutions logicielles et matérielles, dans un esprit communautaire pour la création et la maintenance de ce type de contenus. Concrètement, knoodl permet notamment la création, la gestion et l'analyse des données de type RDF/OWL. Cette solution intègre en plus des fonctionnalités avancées comme : l'import/export de contenus de type RDF/OWL, l'interrogation SPARQL, visualisation des sources de données, des vues d'analyse, des visualisations graphiques, etc.

C'est une solution gratuite hébergée dans le cloud d'Amazon EC2.

Aperçu sur l'ontologie (la faune et la flore en amazighe)

L'ontologie a été conçue en se basant sur les relations conceptuelles et lexico-sémantiques, comme représentée dans le méta-modèle précédent.

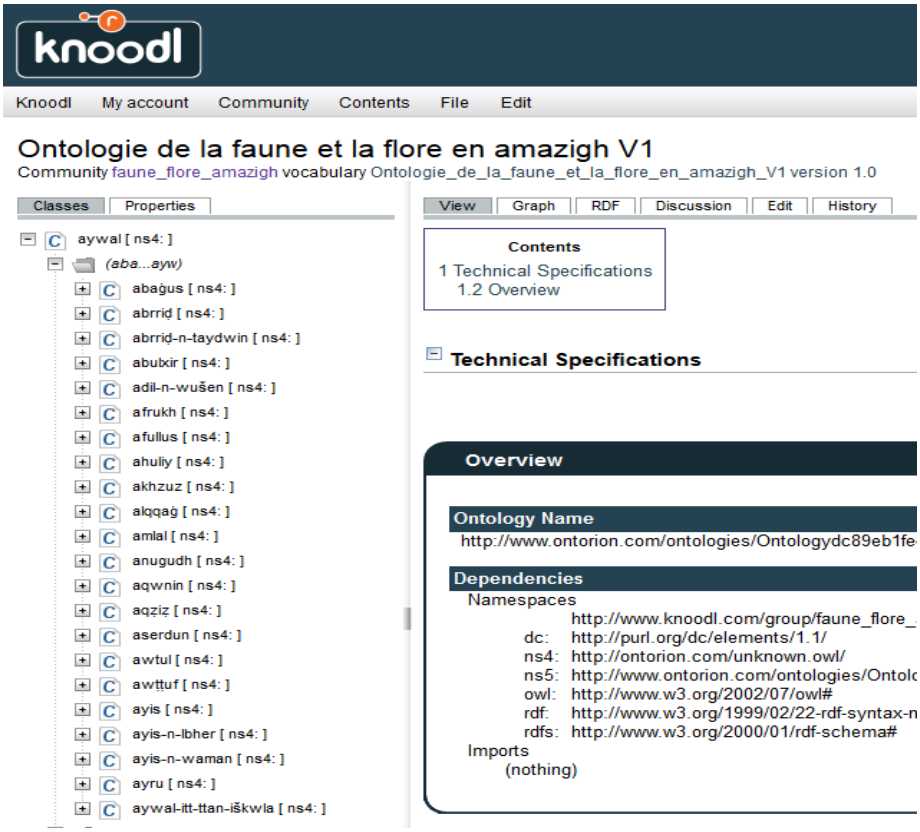


Figure 11: Extrait de l'ontologie à partir du Cloud (Knoodl)

C'est une ontologie du domaine sur la faune et la flore en amazighe enrichie par des éléments linguistiques, intégrable dans notre démarche d'analyse et d'extraction sémantiques des données à partir du texte.

Quelques éléments statistiques sur le contenu de l'ontologie

A l'état actuel du développement de l'ontologie, nous avons intégré au format RDF :

- Plus que 1500 concepts,
- Plus que 3000 relations (conceptuelles, ontologiques, sémantiques, etc.),

- Quelques contraintes types.

Cette ontologie est prête à être utilisée dans notre processus d'analyse sémantique des textes. A la fin de son développement, nous envisageons de l'intégrer dans le Cloud du Linked Open Data (LOD).

Conclusion et perspectives

Bien qu'il existe des systèmes capables d'analyser et de traiter des contenus d'un point de vue sémantique, la relation qui lie le contenu à son utilisation est généralement peu ou pas du tout prise en compte. En effet, cette relation peut être d'une extrême complexité qui nécessite des approches et systèmes intelligents difficiles à mettre en œuvre et capable de s'adapter en fonction du contexte d'utilisation. Dans cet article, nous avons présenté la conception d'une ontologie de contexte à partir d'une ontologie de domaine enrichie par la notion de contexte et les relations lexico-sémantiques, intégrable dans une approche de gestion de la sémantique dans son contexte. Ceci, afin d'améliorer les performances de certains domaines d'application comme l'indexation, la recherche ou encore l'extraction d'information. On a pu, à travers quelques exemples de relations et de définition particulière du contexte, répondre aux soucis relatifs à des énoncés ou à des mots auxquels peuvent correspondre plusieurs et différentes structures sémantiques en analysant fidèlement les relations sémantiques que peuvent entretenir les mots simples, voire les expressions figées au sein d'un même texte et qui relève d'un même domaine, ainsi que les règles de raisonnement qui leur sont applicables. Une extension de cette étude à toutes les relations, une étude approfondie des spécificités de chaque relation ainsi que son utilisation constituent les perspectives du présent travail. On appliquera les résultats de ces améliorations pour affiner « la pertinence » et « l'optimisation » dans des domaines et des utilisations particuliers.

Références

Ahmad K., Fulford H. (1992), "Knowledge Processing: Semantic Relations And Their Use In Elaborating Terminology", *Computing Science Report,* University of Surrey.

Amardeilh, F et *al*., (2005), « Annotation documentaire et peuplement d'ontologie à partir d'extractions linguistiques », *Actes de la Conférence Ingénierie des Connaissances (IC'05)*, Nice.

Bachimont, B., (2001), « Modélisation linguistique et modélisation logique des ontologies : l'apport de l'ontologie formelle », *Actes des journées francophones d'Ingénierie des Connaissances (IC'2001)*, Grenoble.

Chakiri, M. (2011), **«** La locution nominale entre opacité et transparence (parler des Aït Wirra, Moyen Atlas, Maroc », *Etudes et Documents Berbères,* p. 97-108.

Chakiri, M. (2010), **«** Analyse stylistique des locutions nominales en amazighe », *Revue Asinag*, IRCAM, Maroc, p. 201-210.

Condamines A., Rebeyrolle J. (1998), "Ctkb : A Corpus-Based Approach For Terminological Knowledge Base ». in *Proceedings Of The First Workshop On Computational Terminology (COMPUTERM'98), Workshop of Coling'98*, Montréal.

Cunningham, H. et *al*., (2002), « Framework and Graphical Development Environment for Robust NLP Tools and Applications». *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'2002)*, Philadelphia.

Fadili, H., (2013), « Towards a new approach of an automatic and contextual detection of meaning in text », AICCSA'2013, Fès/Ifrane.

Gómez-Pérez A. (1999), *Ontological Engineering: A State Of The Art*, Expert Update.

Guarino N. (1997), "Understanding, Building And Using Ontologies". *International j. Human-computer studies*.

Guerin, É., (2003), « Un exemple d'article dans les actes d'une conférence », *Actes de la conférence CORESA'03*, Lyon.

Hernandez N., (2005), *Ontologies de domaine pour la modélisation du contexte en Recherche d'information*, Thèse de doctorat, Université Paul Sabatier de Toulouse.

L'homme Marie-Claude (2004), *La terminologie : principes et techniques*, Montréal, les presses de l'Université de Montréal.

Mel'čuk, I. (1981), "Meaning-Text Models: A Recent Trend In Soviet Linguistics", *Annual Review Of Anthropology*.

Mel'čuk, I. (1995), *The Russian Language In The Meaning-Text Perspective*, Wiener Slawistischer Almanach/Škola "Jazyki Russkoj Kul'tury": Vienna/Moscow.

Mel'čuk, I. (1997), *Vers une linguistique sens-texte. leçon inaugurale*, Paris: Collège de France.

Mel'čuk, I. (1998), "Dependency Syntax: Theory And Practice", Albany, N.Y.: The SUNY Press.

Michael, B., (1998). *Fractals everywhere*. Academic Press.

Mizoguchi R. And Ikeda M. (1996), *Towards Ontological Engineering (Ai-Tr-96-1,)*. Osaka: Isir, Osaka University.

Mizoguchi R (1998), "A Step Towards Ontological Engineering", paper presented at the 12th National *Conference On Ai Of JSAI*.

Polguere, A., (2003), *Lexicologie et sémantique lexicale. notions fondamentales (paramètres)*, Montréal, PUM.

Rastier, F. (2003), *De la signification au sens. Pour une sémiotique sans ontologie*. [En ligne]. Disponible sur : http://www.revue-texto.net/Inedits/Rastier/Rastier_Semiotique-ontologie.html.

Rastier, F. (2004), « Ontologie(s). Revue des sciences et technologies de l'information », Série : *revue d'intelligence artificielle*, 4.

Sager, J. (1990), *A Practical Course In Terminology Processing*, Philadelphia : John Benjamins publishing company, Amsterdam.

Taifi, M. (1991), *Dictionnaire tamazight-français* (parler du Maroc central), Paris, L'Harmattan-Awal.

Van Heijst G., Schreiber A. And Wielinga B. J. (1997), "Using Explicit Ontologies In Kbs Development". *International Journal Of Human And Computer Studies /Knowledge Acquisition*.

Vanwelkenhuysen J. And Mizoguchi R. (1995), "Workplace-Adapted Behaviors: Lessons Learned For Knowledge Reuse", *Paper presented at the KB&KS '95*.

Vanwelkenhuysen J. And Mizoguchi R.(1994), "Maintaining The Workplace Context In A Knowledge Level Analysis", *Paper presented at the Proc. of JKAW'94*, Hatoyama, Japan.

Wielinga B. And Schreiber A. (1993), "Reusable and Sharable Knowledge Bases: A European Perspective", *Paper presented at the KB & KS'93*, Tokyo.

The GATE platform: http://gate.ac.uk/, mars 2013.

Asinag-Asinag, 9, 2014, p. 171-184

Jeu d'étiquettes morphosyntaxiques de la langue amazighe*

Fadoua Ataa Allah (1), Siham Boulaknadel (1) et Hamid Souifi (2)

(1) CEISIC, (2) CAL - IRCAM

يعتبر الوسم النحوي علما لتصنيف وتدليل كلمات النص أو المتن النصي إلى نوع محدد من أجزاء الكلام بناء على تعريف الكلمة وسياقها. ويعد حجر الزاوية للعديد من تطبيقات المعالجة الآلية للغات الطبيعية وخاصة التصحيح النحوي، التحليل الدلالي والترجمة الآلية. ويعتمد في ذلك على وضع وتحديد لائحة أجزاء الكلام المناسبة للخصوصيات اللغوية للغة المدروسة.

في هذا المقال، نقترح مجموعة *من العلامات النحوية* تضم تصنيفا مفصلا لأجزاء كلام اللغة الأمازيغية، آخذين بعين الاعتبار الخصائص الصرفية والتركيبية للغة الأمازيغية ومستندين في بناء هذه الأقسام على نموذج " EAGLES "، الذي يساهم في الانفتاح على التطبيقات المتعددة اللغات.

This work aims to provide the Amazigh language with a morphosyntactic tagset. In this process, morphology and syntax are considered as an inextricable asset. This tagset will assign to each meaningful unit information concerning the "shape variations of signifiers, their amalgams and their discontinuity" and information about its function in the statement. The proposed tagset is based on EAGLES guidelines in order to ensure the reuse of corpora and language's comparability in natural language processing.

Ce travail se veut une contribution à l'élaboration d'un Jeu d'étiquettes morphosyntaxiques de la langue amazighe. Il s'appuie sur la morphologie et la syntaxe en tant qu'un tout indissociable. Ce jeu permettra d'attribuer à chaque unité significative des informations sur les « variations de forme de signifiants, à leurs amalgames et à leur discontinuité [1]» et sur sa fonction dans l'énoncé.
Le jeu que nous proposons ici se base sur les recommandations EAGLES, visant la réutilisation des corpus et la comparabilité entre les langues dans le domaine du traitement automatique du langage naturel.

* Nous tenons à exprimer nos vifs remerciements à Abdallah Boumalk et Rachid Laabdelaoui (CAL, IRCAM) pour avoir bien voulu relire et commenter ce travail.

[1] Martinet, A. (*Grammaire fonctionnelle du français, cf. 1.8)*

1. Introduction

L'étiquetage morphosyntaxique, appelé aussi *étiquetage grammatical* ou *part-of-speech tagging* (*POS tagging*, en anglais), est un préalable pour de nombreuses applications du traitement automatique des langues, particulièrement la correction grammaticale, l'analyse sémantique, l'analyse pragmatique et la traduction automatique. Toutefois, il nécessite le développement d'un jeu d'étiquettes adéquat aux spécificités linguistiques de la langue étudiée.

Cet étiquetage consiste à attribuer à toute unité significative du texte (occurrence d'un corpus), et à l'aide d'un outil informatique, un symbole correspondant à son comportement syntaxique[2] (nom, verbe, pronom, adverbe, préposition…), et des informations morphologiques qu'il affiche dans un contexte précis du corpus linguistique (masculin, singulier,…). Idéalement, un jeu d'étiquettes doit permettre de[3] :

- représenter la richesse des informations lexicales ;
- représenter l'information nécessaire à la désambiguïsation[4] en contexte des étiquettes morphosyntaxiques ;
- et d'encoder les informations utiles au traitement linguistique pour lequel l'étiquetage morphosyntaxique a été déployé.

Le choix de jeux d'étiquettes est particulièrement délicat. Ainsi, une panoplie de jeux d'étiquettes pour chaque langue a été développée au fil des années et au gré des projets par différents groupes. Face à ces divergences de pratique, qui forment un obstacle à l'échange et à la réutilisation des corpus, basées parfois sur des conceptions différentes de ce que doit être un jeu d'étiquettes de données, d'importants efforts d'uniformisation et d'alignement ont été déployés au sein de projets internationaux tels que EAGLES (EAG, 1996) et MULTEX (MUL, 1996).

Dans un souci de parvenir à une « standardisation » d'étiquettes morphosyntaxiques pour l'amazighe lui assurant la comparabilité entre langues, notamment dans un système multilingue, nous avons élaboré un jeu d'étiquettes en nous inspirant des recommandations fournies par le groupe EAGLES[5] en matière d'annotation morphosyntaxique. Ces dernières sont assez intéressantes, dans la

[2] Bien que les unités significatives (ou mots) appartiennent toutes à une ou plusieurs parties du discours, selon le contexte de leur emploi, des difficultés d'étiquetage peuvent découler de l'ambigüité des rôles grammaticaux de certains mots dans certaines langues, où la mutation catégorielle paraît vivante et dans lesquelles la classification des catégories grammaticales s'avère difficile à établir.

[3] Nous reproduisons les mêmes objectifs visés par *Technolangue.net* cités sur http://www.technolangue.net/article.php3?id_article=296. Site consulté en septembre 2013.

[4] La désambiguïsation d'un mot consiste à faire cesser son ambiguïté en ne retenant qu'un seul de ses sens lié à une seule forme morphosyntaxique, et ce, dans le but de le rendre plus aisé à comprendre.

[5] http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html. Site consulté en septembre 2013.

mesure où elles proposent un cadre, qui peut être aisément suivi, fondé sur une distinction entre *étiquettes obligatoires*, *étiquettes recommandées* et *extension particulière*.

2. Ambiguïté en amazighe

L'amazighe se caractérise par un comportement morphologique riche dans la formation des signifiants ; par flexion, par dérivation ou par **agglutination**. Ce qui le laisse confronté à un problème d'ambiguïté pénalisant tout processus de traitement automatique. Ainsi, l'ambigüité se manifeste sous différentes formes selon les niveaux de traitement lexical, morphologique et syntaxique.

Dans ce contexte, nous essayons, à travers cet article, de décrire l'une des formes d'ambiguïté qui s'imposent pour le lexique amazighe et que seule la syntaxe pourrait désambiguïser comme c'est le cas[6] *dans les exemples qui suivent :*

– *Ambiguïté entre* Nom et Verbe :

Nom : □□□□ [***illi***] (*Litt.* fille de moi) « ma fille »,

Verbe : □□ □□□□[7] [*ur* ***illi***] (*Litt.* ne pas être) « il n'existe pas ».

– *Ambiguïté entre Nom et Adjectif :*

Nom : □□□□ □□□□□□□[8] [*zṛiɣ* ***amqqran***] (*Litt.* voir je amqqran) *« j'ai vu Amqqran »,*

Adjectif : □□□□□ □□□□□□□ [argaz ***amqqran***] (*Litt.* homme âgé) *« l'homme âgé ».*

– *Ambiguïté entre Particule modale du futur (PMF) et Particule démonstrative de proximité :*

PMF : □□ □□□□ [***ad*** *ffɣɣ*] (*Litt.* Particule modale, sortir-1PS, *« je sortirai»,*

Particule démonstrative de proximité : □□□□□ □□ [*argaz* ***ad***] (*Litt.* homme ce) *« cet homme-ci».*

– *Ambiguïté entre Particule modale de l'inaccompli (PMI) et Adverbe de lieu de proximité :* □□.

PMI : □□ □□□□□ [*da ichtta*] (*Litt.* PMI, 3PMS manger (inacc.)) *«Il est en train de manger» ;*

---

[6] L*e problème de la mutation catégorielle est omni présent dans d'autres langues, c'est le cas du Français* (ex. *écailler* (nom « personne qui ouvre et vend des huîtres, des fruits de mer »/verbe « dépouiller un poisson de ses écailles »), *faire* (nom « manière de faire une œuvre »/verbe « construire, fabriquer »), *gabarier*, *baiser*… ou du Vietnamien (ex. *trên* (préposition « sur »/nom « le supérieur »), *trong* (préposition « dans »/nom « l'intérieur »)…voir à ce propos Nguyen *et al.* (2003), « Une étude de cas pour l'étiquetage morphosyntaxique de textes vietnamiens », *cf.*3.2.

[7] Accompli négatif du verbe □□□ *ili* « être ».

[8] Le nom □□□□□□□ [*amqqran]* « *Amqqran » peut être un nom propre.*

Adverbe de lieu : □□□□ □□ □□ [*ssrs it da*] (*Litt.* Poser (V. impératif), masc. sing. le, ici) *«Pose-le ici».*

– *Ambiguïté entre Auxiliaire (ou particule) de prédication, coordonnant et Particule d'orientation de rapprochement :*

Auxiliaire de prédication : □ □□□□□ [***d** argaz*] (*Litt.* c'est homme) *« c'est un homme »,*

Coordonnant : □□□□□ □ □□□□□ [*argaz **d** trbat*] (*Litt.* homme et fille) *« l'homme et la fille »,*

Particule d'orientation spatiale de rapprochement : □□□□ □ □□□□ [*idda **d** urba*] (*Litt.* 3PMS.Venir ici enfant) *« l'enfant est venu (vers ici)».*

3. Pourquoi le modèle EAGLES ?

Le groupe *EAGLES* (*Expert Advisory Group on Language Engineering Standards*), fondé en février 1993, est une initiative de la commission européenne du programme « Ingénierie et Recherche Linguistique » (*Linguistic Research and Engineering, LRE*). Il a pour but l'élaboration de standards des ressources langagières (corpus écrit et oral, lexiques électroniques) et des moyens de structuration et d'exploitations, ainsi que des procédés d'évaluation de ressources et d'outils. Ainsi, plusieurs acteurs industriels, professionnels et universitaires à l'échelon européen ont participé à la réalisation d'une batterie de recommandations, notamment en matière d'annotation morphosyntaxique.

Les spécifications proposées par EAGLES sont le fruit d'une étroite collaboration avec le projet MULTEX (Multilingual Text Tools and Corpora). Elles résultent de l'observation et de l'analyse d'un ensemble de projets de corpus et de lexiques, qui ont permis de relever des traits communs aux différentes langues, et de déterminer un noyau d'informations morphosyntaxiques sur lesquelles un accord assez large peut être établi. Ce noyau se complète par des couches d'informations optionnelles, ou propres à des applications particulières.

Bien que le modèle EAGLES ne soit pas applicable tel quel à l'amazighe, il peut servir de point de départ à de nombreux acteurs pour disposer de ressources linguistiques directement exploitables ainsi que d'outils de traitement réutilisables. Le modèle a été développé d'une manière à systématiser les stratégies d'étiquetage dans un environnement aussi bien monolingue que multilingue, et par conséquent à assurer la comparabilité de l'amazighe avec d'autres langues tout en permettant une grande flexibilité.

4. Définition du « jeu d'étiquettes » pour l'amazighe

Pour l'amazighe, la question de la classification des catégories grammaticales (*cf.* 2) est une tâche difficile et toujours en débat, du fait que cette langue est en voie de standardisation et que ses règles sont en phase de codification. A ce jour, malgré

les quelques études menées[9] dans ce cadre, il n'existe aucun standard reconnu pour les catégories des mots. C'est dans cette perspective que nous proposons un jeu d'étiquettes pour les applications du traitement automatique de l'amazighe.

Sachant que la grammaire se définit comme l'étude systématique des éléments constitutifs d'une langue et la science qui permet de montrer la structure d'une langue et d'expliquer les règles de changement et de combinaison des mots formant un énoncé, elle se subdivise en deux parties : la morphologie, qui étudie les changements des mots, et la syntaxe qui s'intéresse à la combinaison des mots en formes plus étoffées que de simples mots. Néanmoins, la catégorie grammaticale représente l'unité dialectique de la valeur grammaticale et de la forme grammaticale, reflétée par les différentes oppositions existant dans un système syntaxique d'une langue.

Pour bien refléter toutes les relations syntaxiques possibles, généralement, il faut disposer d'un important jeu d'étiquettes. Cependant, plus le nombre des étiquettes est important plus la tâche d'annotation ou d'étiquetage est difficile. D'où la nécessité de gérer ce compromis afin de parvenir à un jeu d'étiquettes assez précis et de taille acceptable. Ainsi, en nous basant, d'une part, sur la définition du mot graphique en amazighe[10], et d'autre part, sur les recommandations formulées par EAGLES, nous avons proposé deux listes d'étiquettes [11] spécifiques aux caractéristiques de la langue amazighe. Une assez précise, elle contient les étiquettes obligatoires qui reflètent toutes les oppositions du système syntaxique (*cf.* Tableau 1). L'autre traduit toutes les relations syntaxiques contenant les étiquettes recommandées, qui fournissent des indications plus précises telles que le nombre, le genre et l'état pour les noms et les adjectifs, le temps, le mode et la personne pour les verbes.

4.1. Etiquettes obligatoires

Les étiquettes obligatoires comprennent les parties du discours d'une langue, représentées par les classes de mots ayant les mêmes propriétés sémantiques et grammaticales. Ces classes sont principalement caractérisées par le sens général catégoriel, puis par la forme grammaticale ainsi que la fonction syntaxique des mots.

[9] Entre autres : (Loikkanen, 2007), (Ataa Allah et Jaa, 2009), (Outahajala *et al.*, 2010), ….

[10] Selon (Ameur *et al.*, 2004 : 34; Boukhris *et al.*, 2008 : 27), un mot graphique est constitué d'une séquence de lettres, et éventuellement d'une seule lettre, délimitée par deux blancs typographiques.

[11] Pour une fiabilité optimale des systèmes d'étiquetage, il est à recommander d'avoir un important jeu d'étiquettes représentant toutes les relations morphosyntaxiques de la langue annotée, plus le corpus est volumineux, plus le résultat est précis. Mais cela ne va pas sans contrainte, car plus le jeu d'étiquette est important, plus la tâche d'annotation est difficile aussi.

Compte tenu des spécifications lexicales de chaque classe en amazighe12, le jeu d'étiquettes que nous proposons (Tableau 1) contient douze (12) catégories principales, à l'encontre du modèle EAGLES qui en propose 13 et dont la catégorie « Article » n'est pas supportée par l'amazighe. En outre, la catégorie « Unique » a été remplacée par la catégorie « Particule » afin de répondre aux spécificités morphosyntaxiques de l'amazighe.

N°	*CATEGORIES*	*N°*	*CATEGORIES*
1	Nom (N)	7	Particule (P)
2	Verbe (V)	8	Conjonction (C)
3	Adjectif (AJ)	9	Interjection (I)
4	Pronom (PR)	10	Résiduel (R)
5	Adverbe (AV)	11	Numéral (NU)
6	Préposition[13] (AP)	12	Ponctuation (PU)

Tableau 1 : Liste des étiquettes obligatoires

4.2. Etiquettes recommandées

Dans cette section, nous envisageons de définir les étiquettes recommandées, où chaque catégorie spécifiée dans la liste obligatoire sera subdivisée à l'aide d'étiquettes plus spécifiques relevant de la catégorie morphologique par le changement de la forme du mot et de la propriété sémantique.

1. Nom (N) :

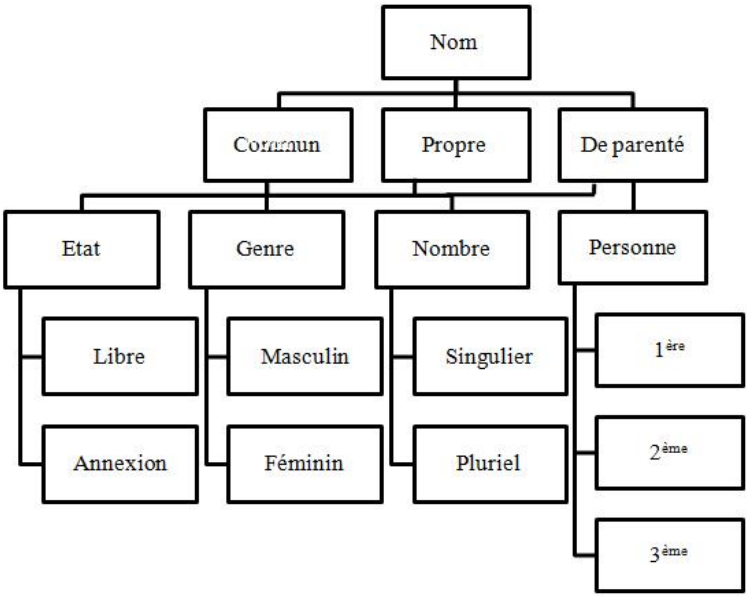
Le nom peut être de type commun, propre ou de parenté[14]. Tous les types varient en genre, en nombre et en état[15]. En outre, les noms de parenté varient aussi en fonction de la personne.

[12] Ces classes se basent sur les traits morphologiques et sur les parties du discours décrites dans *Initiation à l'amazighe* (Ameur *et al.* : 2004) et la *Nouvelle grammaire de l'amazighe* (Boukhris *et al.*, 2008 : 27).

[13] Le modèle EAGLES propose l'étiquette adposition (AP) qui comprend la préposition et la postposition.

[14] Selon les recommandations d'EAGLES, nous avons classé le nom du type numéral dans la catégorie « Numéral ».

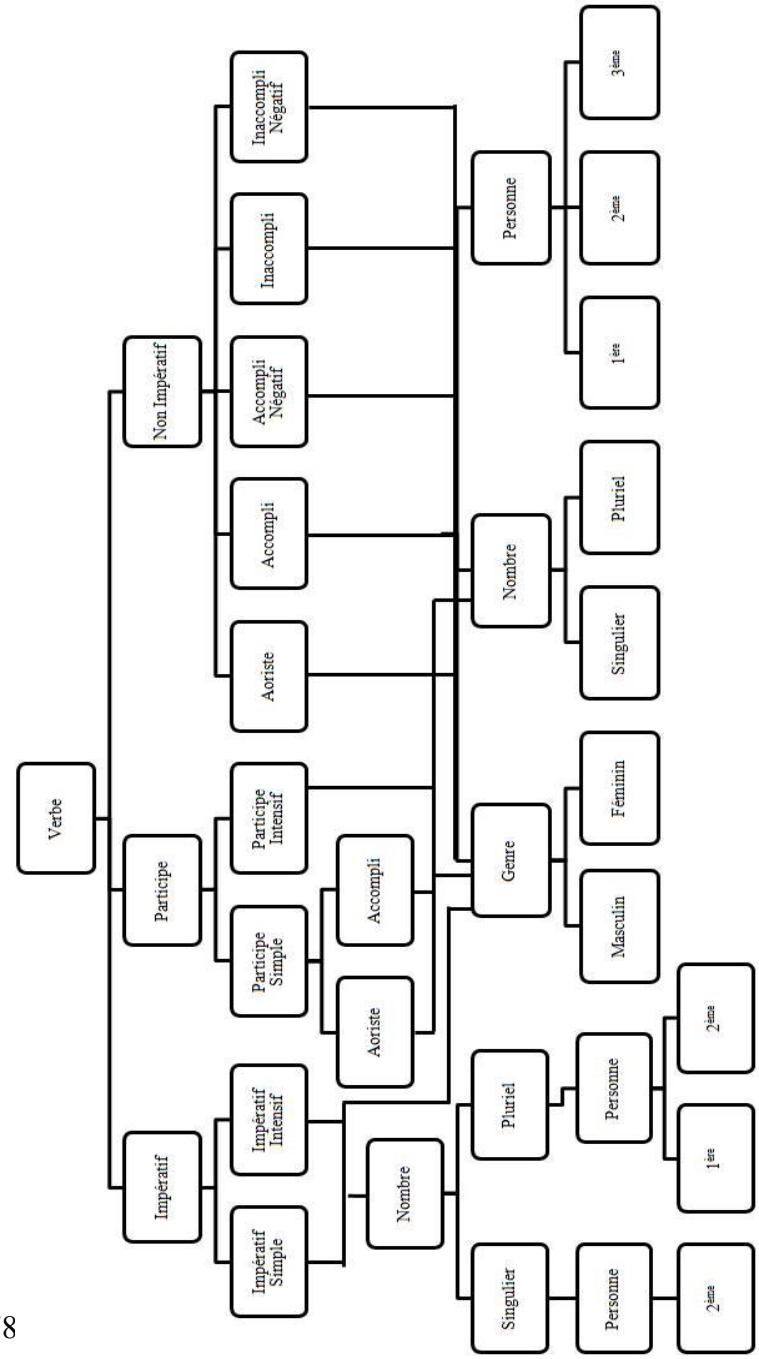
[15] C'est un concept grammatical morphosyntaxique qui affecte l'initiale vocalique des noms masculins par changement de formes : □→□/□□ (*a*→*u*/*wa*), □→□□ (*i*→*yi*) et □→□□ (*u*→*wu*). Pour les noms féminins, ils marquent l'état d'annexion par l'effacement de la voyelle (□ [*a*] / □ [i]) placée après la marque préfixée □ [*t*].



2. Verbe (V) :

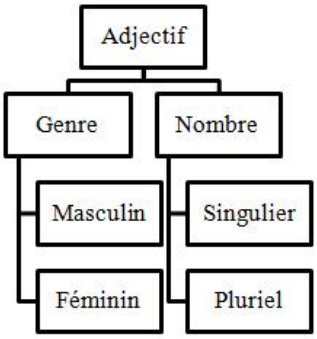
Le verbe se combine avec les modalités aspectuelles des quatre thèmes : l'aoriste, l'accompli, l'inaccompli et l'accompli négatif. Néanmoins, afin de couvrir toutes les variantes amazighes du Maroc, nous introduisons dans le schéma du verbe, en plus des quatre thèmes verbaux communs, la forme de l'inaccompli négatif présente dans les parlers du Rif et de Figuig.
Dans sa conjugaison, le verbe reçoit des désinences verbales qui sont de trois types : (1) non impératives, (2) impératives et (3) participiales.
Lorsque les deux formes simples[16], de l'impératif et du participe, se combinent avec les modalités de l'inaccompli (gémination d'une consonne radicale ou la préfixation de : □□ *tt*), elles sont appelées successivement impératif intensif et participe intensif.

[16] Le participe simple est obtenu par la combinaison des désinences de la forme participiale avec l'aoriste ou l'accompli (Manuel de conjugaison, p. 14).


Verbe
Impératif
Impératif Simple
Impératif Intensif
Participe
Participe Simple
Participe Intensif
Non Impératif
Aoriste
Accompli
Inaccompli
Accompli Négatif
Inaccompli Négatif
Nombre
Singulier
Pluriel
Genre
Masculin
Féminin
Personne
1ère
2ème
3ème

3. Adjectif (AJ) :

L'adjectif est un nom de qualité qui suit directement le nom qualifié et il s'accorde avec lui en genre et en nombre. Il se distingue du complément de nom par sa position syntaxique[17]



4. Pronom (PR) :

On distingue les pronoms autonomes et les pronoms affixes. Le premier type qui concerne ce travail comprend en plus des pronoms personnels autonomes, les démonstratifs, les interrogatifs[18], les possessifs et les pronoms régimes. Tandis que le deuxième type comporte les pronoms affixes aux présentatifs, aux quantifieurs et à la préposition.

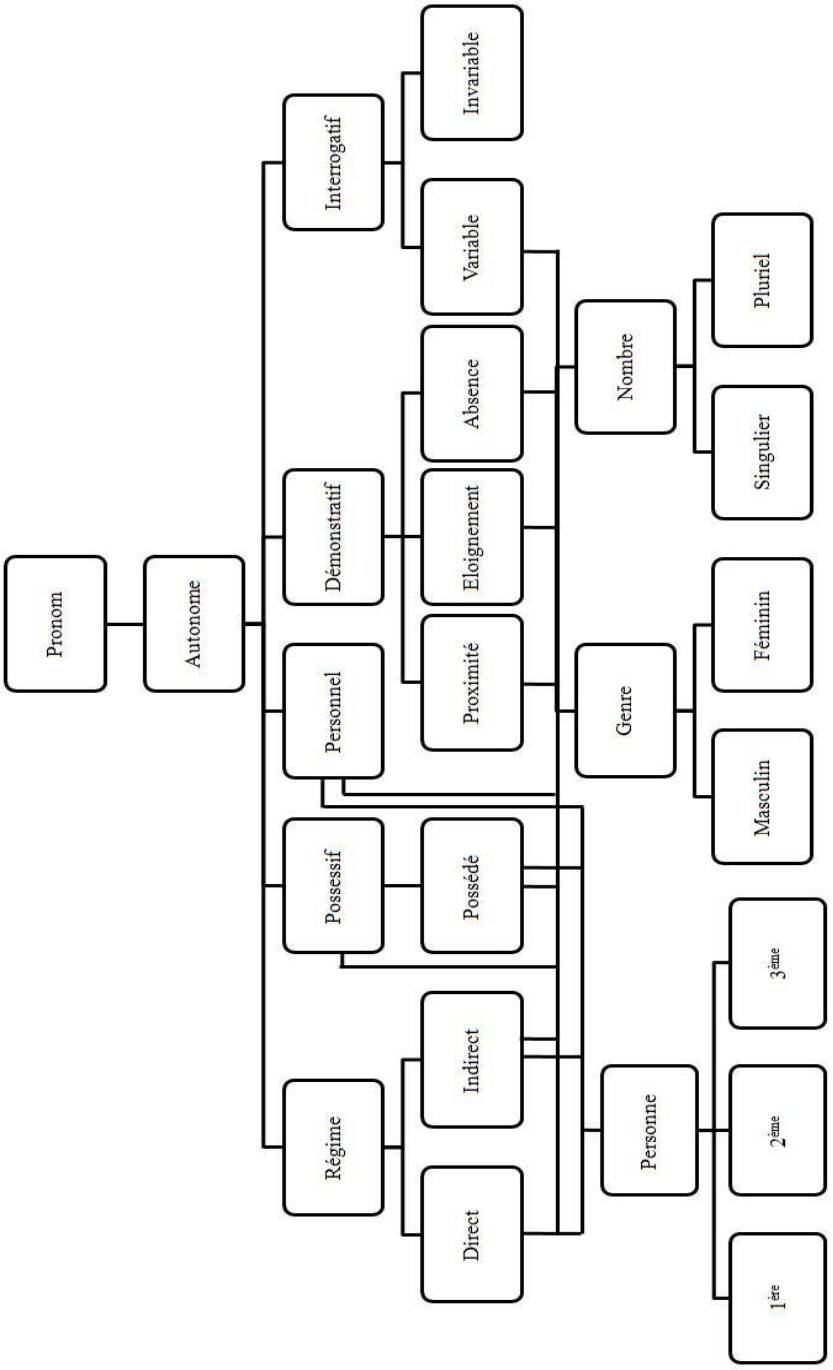
Partons du principe que toute unité lexicale est délimitée par deux blancs typographiques, les présentatifs, les quantifieurs et la préposition forment ainsi un seul mot graphique avec leur complément pronominal, ce qui explique l'exclusion des pronoms affixes du jeu des étiquettes recommandées.

Concernant les pronoms régimes (complément d'objet direct/indirect), beaucoup d'amazighisants les considèrent comme étant des « affixes ». Or, en nous basant, d'une part, sur le rôle syntaxique de chacun des deux pronoms (leur substitution à un nom ou à un syntagme prépositionnel, leur ordre dans l'énoncé verbal…), et d'autre part sur la tâche principale de l'étiquetage morphosyntaxique, nous les considérons comme « pronoms autonomes unifonctionnels » par opposition aux «

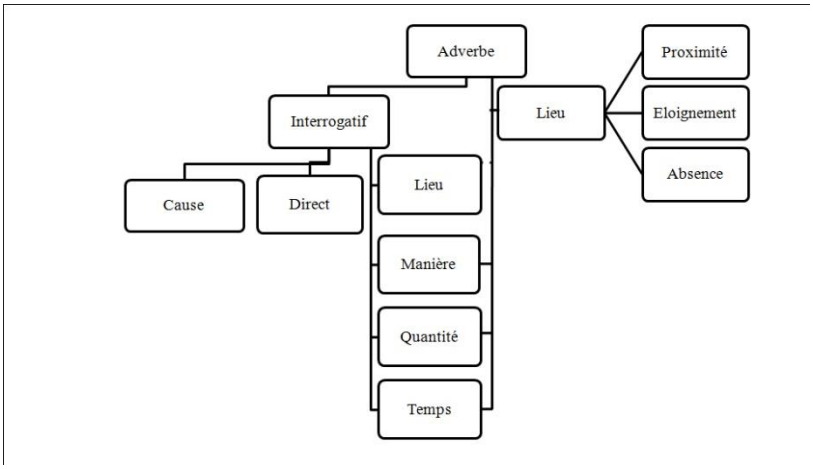
[17] Le complément du nom affiche l'état d'annexion qui résulte de la présence de la préposition [*n*] « de » généralement supprimée pour les locuteurs dans leur langage quotidien. Il se distingue de l'adjectif par sa structure syntaxique étant donné que ce dernier est toujours juxtaposé au nom qu'il détermine, sans le biais d'aucun fonctionnel.

[18] En plus des pronoms interrogatifs associés aux pronoms déictiques (proximité, éloignement, absence) variables en genre et en nombre, il existe des pronoms interrogatifs invariables qui portent sur une partie de la phrase ([ma], [u], …). Voir (*La nouvelle grammaire de l'amazighe*).

pronoms autonomes plurifonctionnels », connus sous le nom « pronoms personnels autonomes ».



Généralement, l'adverbe est un mot invariable, sauf exception liée à des emplois régionaux. Il exprime le lieu, le temps, la qualité et la manière.

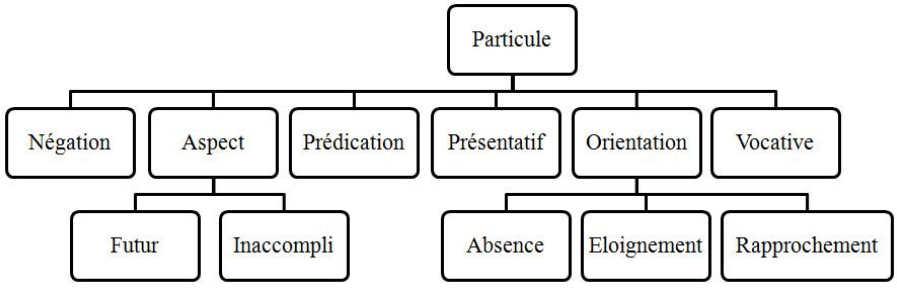


6. Préposition (AP) :

La préposition relève de la catégorie générale des mots de relation. Elle exprime des valeurs sémantiques diverses, notamment, la localisation spatio-temporelle, l'instrument, la direction, la possession, l'appartenance et l'accompagnement.

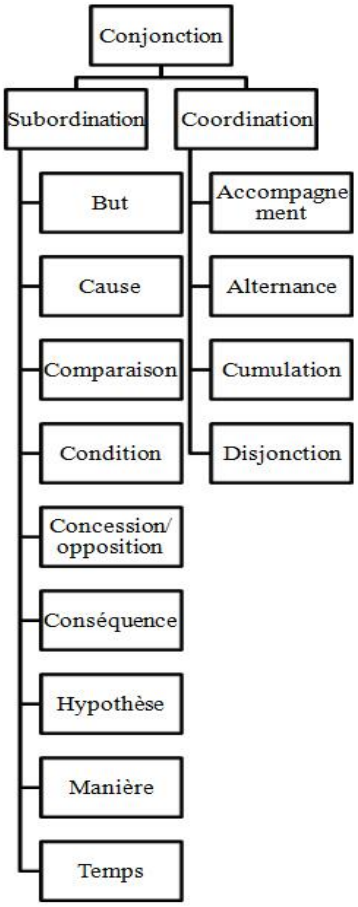
7. Particule (P) :

Les particules sont un ensemble de mots assez courts qui jouent le rôle d'indicateurs grammaticaux au sein d'une phrase.



8. Conjonction (C) :

La conjonction peut marquer la subordination ou la coordination.



9. Interjection (I) :

Il est question ici d'un mot invariable, autonome et n'ayant pas de fonction grammaticale. Il exprime une sensation, une émotion, un agacement, un étonnement …

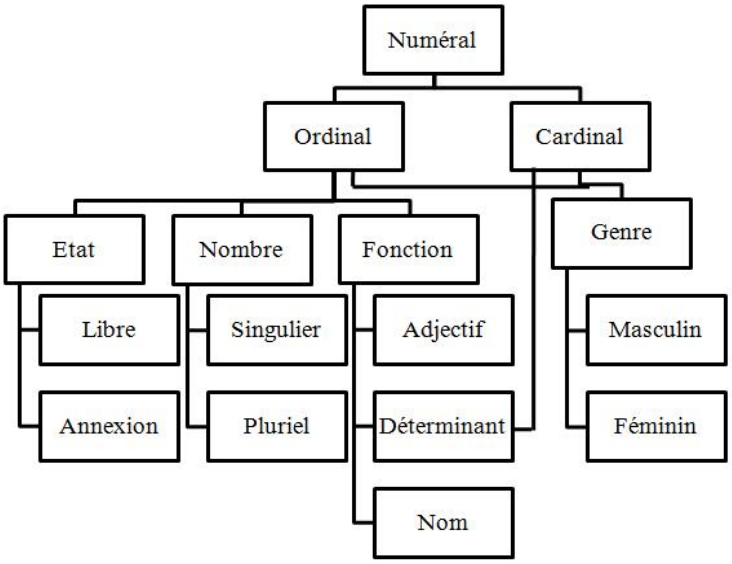
10. Résiduel (R) :

La valeur résiduelle est affectée aux termes qui ne rentrent pas dans les catégories usuelles. Par exemple: les mots étrangers, les formules mathématiques et les symboles.

Même si ces termes ne font pas partie du lexique de la langue traitée, ils se produisent assez fréquemment. Par conséquent, ils ont besoin d'être étiquetés.

11. Numéral (NU) :

La classe du numéral se compose de deux sous catégories : cardinal et ordinal. Cette classe peut se combiner avec les trois modalités démonstratives (déterminants) : de proximité, d'éloignement et d'absence.



12. Ponctuation (PU).

5. Conclusion

Dans la perspective de doter la langue amazighe d'outils nécessaires au traitement automatique, l'élaboration d'un jeu d'étiquettes est une étape préalable dans le processus d'automatisation. Ainsi, nous avons proposé deux jeux d'étiquettes grammaticales, dont le premier correspond à des étiquettes obligatoires et le deuxième à des étiquettes recommandées, où nous avons essayé de prendre en considération les particularités et la richesse grammaticales de l'ensemble des variantes amazighes marocaines.

Références bibliographiques

Ameur, M. et *al.* (2004), *Initiation à la langue amazighe*, Publications de l'Institut Royal de la Culture Amazighe, Série : Manuels-N°1, Rabat.

Ataa Allah, F., Jaa, H. (2009), « Etiquetage morphosyntaxique : Outil d'assistance dédié à la langue amazighe ». Actes du 1er Symposium International sur le Traitement Automatique de la Culture Amazighe. 12-13 décembre 2009, Agadir, Maroc, p. 110-119.

Boukhris, F. et *al.* (2008), La nouvelle grammaire de l'amazighe, Publications de l'Institut Royal de la Culture Amazighe, Série : Manuels-N°2, Imprimerie El Maârif Al Jadida, Rabat.

EAG (1996). EAGLES, Recommendation for the Morphosyntactic Annotation of Corpora. EAGLES Document EAG-TCWG-MAC/R.
http://www/ilc.cnr.it/EAGLES96/home.html [15.1.2007].

Laabdelaoui, R. et *al*. (2012), *Manuel de conjugaison amazighe*, Publications de l'Institut Royal de la Culture Amazighe, Série : Manuels-N°5, Rabat.

Loikkanen, S. (2007), « Étiquetage morpho-syntaxique de textes kabyles ». Actes de la conférence en Traitement Automatique des Langages Naturelles. 5–8 juin 2007, Toulouse, France. p. 193-202.

Martinet, A. (1979), *Grammaire fonctionnelle du français, Paris, Didier-Crédif, XII + 276 p.*

MUL (1996). Multilingual Text Tools and Corpora. http://www.lpl.univ-aix.fr/projects/multext.

Nguyen, T. M. H. et *al.* (2003), « Une étude de cas pour l'étiquetage morpho-syntaxique de textes vietnamiens », Actes de la conférence Traitement Automatique du Langage Naturel, 11-14 juin 2003, Batz-sur-Mer, France.

Outahajala, M. et *al.* (2010), « Tagging Amazigh with AnCoraPipe », Actes de l'atelier Language Resources and Human Language Technologies for Semitic Languages, 17 mai 2010, Valette, Malta. p. 52-56.

Paroubek P., Rajman M. (2000), « Etiquetage morpho-syntaxique », In J.-M. Pierrel, Ed., Ingénierie des langues, Paris : HERMES Science Europe, p. 131-150.

Valli A. et *al.* (1999), « Étiquetage grammatical des corpus de parole : problèmes et perspectives », *Revue française de linguistique appliquée*, Vol. 4, No. 2, p. 113-133.

Sitographie :

http://www.technolangue.net/article.php3?id_article=296.

http://www.ircam.ma/doc/publica/nouvel-gram-amazigh.pdf.

http://www.ircam.ma/doc/publica/initiation-langue-amazighe-1.pdf.

http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html.

http://www.lexilogos.com/vietnamien_dictionnaire.htm.

Asinag-Asinag, 9, 2014, p. 185-196

Eurêka un dépôt d'objets d'apprentissage compatible avec le profil d'application Normetic (LOM)

Robert Bibeau

La normalisation des technologies de l'information destinées à l'apprentissage, à l'éducation et à la formation, a pour objet d'améliorer l'accessibilité des ressources d'enseignement et d'apprentissage produites dans une variété d'établissements d'enseignement et d'organismes privés et publics, de divers pays, utilisant différentes langues, sous divers environnements technologiques. L'édifice d'un environnement d'apprentissage standard pour les langues minoritaires tel l'amazighe trouve dans l'approche normative la meilleure alternative pour organiser les ressources éducatives futures de la langue amazighe. Dans ce travail, nous commencerons par l'examen du profil NORMETIC qui est une variante d'application de la norme IEEE 1484.12.1 (LOM) des métadonnées d'objets d'apprentissage développé par le GTN-Québec. Dans un deuxième lieu, nous détaillerons les caractéristiques majeures de la Banque Eurêka des ressources d'apprentissages à partir du point de vue de sa compatibilité avec le profil NORMETIC.

Introduction

L'introduction des nouvelles technologies a pour effet de transformer les modes d'enseignement et d'apprentissage. Les établissements d'enseignement ainsi que les ministères, les entreprises et les organismes impliqués dans des activités de formation investissent des moyens financiers croissants dans la production de ressources pédagogiques numériques.

Plusieurs de ces acteurs voient l'intérêt de réutiliser, d'échanger, d'exporter ces ressources, de même que l'opportunité de les partager ou de les commercialiser. Ces acteurs s'intéressent à la création de banques d'actifs pédagogiques constituées d'ensembles de ressources d'enseignement et d'apprentissage (REA) qui sont accessibles, durables et réutilisables et qui répondent à des exigences d'interopérabilité. Les institutions se préoccupent aussi des enjeux de l'efficacité pédagogique, des échanges internationaux de ressources didactiques et de la rentabilité des investissements dans les contenus numériques (coûts de la maintenance, coûts de la migration d'un environnement technologique à un autre, coûts de la répétition indue des mêmes développements, etc.).

Pourquoi un dépôt de ressources standard d'enseignement et d'apprentissage ?

Il faut visiter de nombreux sites Web pour trouver des ressources d'enseignement et d'apprentissage utiles à l'élaboration d'une activité éducative, d'une session de formation ou d'une leçon (Bibeau, 2002). C'est pourquoi les enseignants visitent régulièrement leurs sites web préférés en quête de nouveautés. [1] Ce travail de recherche est long et les résultats sont souvent décevants, sans compter que les objets d'apprentissage trouvés sont parfois inutilisables et exigent une adaptation. Le dépôt de ressources d'enseignement et d'apprentissage (REA) ***Eurêka***, de la Vitrine APO, peut leur venir en aide.[2] ***Eurêka*** est une base de données et un catalogue, non pas des REA elle-mêmes, mais des métadonnées décrivant des milliers de ressources d'enseignement et d'apprentissage soigneusement sélectionnées et décrites par des partenaires TIC du réseau scolaire et collégial ainsi que par certains organismes francophones[3].

Le profil D'application Normetic - Version 1.1 - [4]

NORMETIC est un profil d'application du standard IEEE 1484.12.1-2002 (LOM) (*Learning Object Metadata*)[5] portant sur les métadonnées pour la description des ressources d'enseignement et d'apprentissage (REA). Nous définissons ces REA comme toute entité, numérique ou non-numérique, conçue ou pouvant être utilisée pour des fins d'apprentissage, d'éducation et de formation[6]. Le profil d'application NORMETIC est une sélection d'éléments du standard LOM formant un sous-ensemble adapté aux besoins communs clairement définis par divers acteurs du domaine de l'éducation et de la formation. Parmi les 77 éléments de métadonnées de LOM, NORMETIC comprend notamment 19 éléments requis. Tous ces éléments doivent obligatoirement être documentés pour permettre une documentation homogène et uniforme dans la perspective de constituer un patrimoine éducatif. Cette obligation s'applique également à quatre autres éléments dans certaines situations (requis conditionnels).

[1] Carrefour-éducation est certainement l'un des sites Web les plus fréquentés par les enseignants québécois. http://carrefour-education.telequebec.qc.ca/

[2] Pour en savoir plus sur Eurêka http://www.robertbibeau.ca/eureka/Normes-standard.ppt http://eureka.ntic.org

[3] Un profil d'application est une sélection d'éléments d'une norme ou d'un standard formant un sous-ensemble adapté aux besoins des groupes qui l'utilisent.

[4] Texte de Robert Thivierge, Adapté par Robert Bibeau.

[5] http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf

[6] Cette définition a été présentée par la délégation canadienne à l'ISO/IEC JTC1/SC36 en 2005 (GTN-Q, 2005).

Plusieurs consultations ont été tenues auprès de divers experts et usagers du monde de l'éducation et de la formation (auteurs, producteurs et diffuseurs de ressources didactiques, professeurs, bibliothécaires, administrateurs, technologues).

Ces consultations ont permis de dégager un consensus sur les besoins exprimés et sur les propriétés des REA pour qu'elles y répondent :

- **Accessibilité** [7] **:** permettre une recherche facile, l'identification et la livraison de REA d'une façon distribuée.
- **Durabilité :** permettre aux REA d'affronter les changements des environnements technologiques en minimisant la ré-ingénierie ou le re-développement.
- **Interopérabilité :** permettre l'utilisation des REA développées par une organisation dans un environnement technologique donné par d'autres organisations dans d'autres environnements technologiques en assurant l'échange de l'information et son utilisation.
- **Pertinence pédagogique :** pouvoir identifier lors de la recherche les contextes pédagogiques des REA (âge, discipline, milieu, objectif pédagogique, niveau d'interactivité, type de ressources, etc.) et les rendre compréhensible.
- **Partage et collaboration :** favoriser l'échange, la coproduction et l'enrichissement des REA.
- **Reconnaissance de la propriété intellectuelle :** permettre de documenter et de reconnaître la propriété intellectuelle et de respecter les droits d'auteur.
- **Réutilisation et adaptabilité :** permettre la réutilisation des REA à différentes fins, dans différentes applications, dans différents produits et différents contextes, par différents modes d'accès.

Les REA qui ont ces propriétés peuvent être de véritables actifs pédagogiques[8] et faire partie d'un patrimoine éducatif, ou un « bien collectif » durable et universel, qui peut être partagé.

La portée de l'étude NORMETIC s'est limitée à la description normalisée des REA par l'utilisation de métadonnées. Pour répondre pleinement aux besoins identifiés, plusieurs autres dimensions devront être abordées dans des étapes subséquentes. Celles-ci aborderont notamment les protocoles de communication, les interfaces de programmation d'applications (API), les architectures, le séquencement, les mécanismes d'évaluation, les dépôts de REA, les mécanismes de diffusion, les mécanismes de certification, le design pédagogique, etc.

[7] À l'ISO, le terme « Accessibilité » concerne plutôt l'accès des technologies de l'information pour les personnes handicapées ou pour ceux dont les moyens sont limités (enfants, personnes âgées). Des informations additionnelles sont consultables à l'adresse suivante : http://www.iso.org/iso/fr/SiteQueryResult.SiteQueryResult

[8] Nous définissons cette expression de cette façon, Actif pédagogique : un ensemble de REA accessibles, durables et réutilisables.

Le profil d'application NORMETIC

Les travaux ont permis d'établir un consensus sur un ensemble d'éléments de métadonnées qui permettent de décrire les REA. Ces travaux ont répondu aux besoins et aux exigences minimales communes dégagées par les divers acteurs consultés tout en étant conformes à un standard reconnu internationalement.

Le profil d'application NORMETIC a pour objet de fournir une méthode commune pour la description des REA. Il facilite, de cette manière, leur identification, leur référencement et leur repérage et il permet de maximiser l'accessibilité, la diffusion, la réutilisation et la durabilité de ces REA. Dans un contexte où les différents acteurs souhaitent l'émergence d'un espace d'échange ou de partage des REA ainsi que la constitution d'un patrimoine éducatif, le profil d'application Normetic répond à des questions communes évoquées dans la documentation des REA au moment de leur production.

Un profil d'application est une sélection d'éléments d'une norme, d'un standard ou d'une spécification[9] qui forme un sous-ensemble ayant une valeur répondant aux besoins communs des groupes qui l'utilisent et leur fournit un cadre d'opération. Les éléments sont retenus selon qu'ils aient une valeur concordant avec un contexte facilitant l'intégration de normes internationales existantes, souscrivant à des besoins spécifiques formulés par des usagers et soutenant l'émergence de meilleures pratiques.

Les nouveaux profils d'application développés au Royaume–Uni, en Finlande et en France proposent aussi des éléments obligatoires assurant ainsi un minimum d'information pour une ressource donnée. Une revue des pratiques exemplaires dans divers pays a servi à déterminer le degré de compatibilité recherchée entre l'ensemble de métadonnées choisies pour le profil d'application NORMETIC et les autres normes, standards, ou profil d'application en usage dans le domaine de la description normalisée des REA.

Le standard IEEE 1484.12.1-2002 (LOM) comporte 77 éléments de métadonnées regroupés en 9 catégories. Dans une perspective technique, la seule exigence pour être conforme au LOM est que tous les éléments de métadonnées (77) puissent être soutenus bien que chacun de ces éléments soit considéré comme optionnel.

Du point de vue des usagers consultés, force a été de conclure qu'il est essentiel de s'assurer qu'un nombre limité d'éléments de métadonnées soient obligatoirement

[9] Une norme est une entente consensuelle établie par les partenaires d'un organisme international reconnu officiellement (ISO/IEC JTC1/SC36) pour fournir une solution normative à des problèmes communs de description, d'indexation et de classification des informations, des processus ou des services. Un standard se veut une entente consensuelle issue d'une pratique commune qui offre aux partenaires adhérents à des organismes nationaux et internationaux (IEEE LTSC, IMS Global Learning Consortium) des solutions normatives pour décrire, indexer et classifier des informations, des processus ou des services. Une spécification désigne des exigences techniques auxquelles des informations, des processus ou des services doivent se conformer pour qu'ils puissent être décrits, indexés et classifiés. Ces exigences peuvent être indépendantes d'une norme ou d'un standard (CREPUQ-Novasys, 2003).

documentés pour décrire, indexer et classifier les REA selon les propriétés recherchées et conformément aux besoins identifiés.

La version 1.1 de ce profil présente un nombre réduit de dix-neuf (19) éléments de métadonnées LOM dont les champs doivent obligatoirement être documentés. Cette prescription minimale, conforme à ce profil d'application, garantit la disponibilité d'une documentation homogène et uniforme. Le profil d'application NORMETIC v. 1.1 prévoit également l'utilisation de quatre autres (4) éléments de métadonnées requis conditionnellement à l'existence de certaines données relatives à la REA.

Les propriétés recherchées et les éléments LOM correspondant aux statuts requis et *requis conditionnel* du profil d'application NORMETIC **:**

Accessibilité :

- Général (Titre, Langue, *Mot-clé*)
- Technique (Format, Localisation)
- Classification (Objectif, Source, ID, Entrée)

Durabilité :

- Cycle de vie (Version)
- Technique (Format, Localisation)

Interopérabilité :

- Métamétadonnées (Schéma de métadonnées)
- Technique (Format, Localisation)

Pertinence pédagogique :

- Pédagogie (Type de ressource pédagogique, Contexte)

Partage et collaboration :

- Classification (Objectif, Source, ID, Entrée)
- Droits (Coût, Copyright et autres restrictions, Description)

Reconnaissance de la propriété intellectuelle :

- Cycle de vie (Rôle, Entité, Date)
- Droits (Coût, Copyright et autres restrictions, Description)

Réutilisation et adaptabilité :

- Général (Langue, Description, Mot-clé)
- Méta-métadonnées (Schéma de métadonnées)
- Technique (Format)
- Pédagogie (Type de ressource pédagogique)

Ces dix-neuf éléments requis et ces quatre éléments requis conditionnels permettent essentiellement de répondre aux questions suivantes :

- **Quelles sont les caractéristiques de la ressource ?** Cela consiste à préciser le type de ressource, la clientèle visée et l'environnement technologique dans lequel elle peut être utilisée.
- **Comment est gérée la propriété intellectuelle ?** Cela permet d'identifier les règles ou conditions à respecter (copyright, droits d'usage, coût) et d'identifier les auteurs et producteurs associés à la création de la ressource.
- **Comment classifier cette ressource ?** Cela consiste à classifier la ressource en fonction des catégories reconnues et à fournir des mots-clés pour en faciliter le repérage par la suite.

En plus de ces vingt-trois (23) éléments devant obligatoirement ou conditionnellement être documentés, le profil d'application NORMETIC v. 1.1 comprend également neuf (9) *éléments recommandés* qu'il est suggéré de documenter si l'information est jugée nécessaire ou pertinente et vingt-six (26) *éléments facultatifs* qui comprennent d'autres éléments dont la documentation est laissée à la discrétion de l'utilisateur.

Pourquoi une description normalisée selon le profil *Normetic* ?

Il est agréable de pouvoir puiser dans le patrimoine éducatif d'une variété de pays francophones par exemple. Cet exercice n'est toutefois pas sans embûches : comment distinguer le diplôme d'études collégiales du Québec du bac français ? Il faut savoir également que le cours secondaire est dispensé à la polyvalente au Québec, au collège et au lycée en France, à l'athénée en Belgique et au gymnase en Suisse...

Seule l'utilisation d'une description normalisée permet d'éviter ces écueils. Ainsi, en spécifiant l'âge des apprenants à qui s'adresse la ressource, on évite les difficultés liées à l'utilisation des nomenclatures locales. Le standard utilisée dans ***Eurêka*** est IEEE LOM (Learning Object Metadata), qui présente 58 descripteurs documentés regroupés en différentes catégories (titre, description, format, cycle de vie, etc.).[10] Au Québec, nous proposons le profil d'application ***Normetic*** version 1.1, qui exige l'utilisation de 21 des descripteurs du LOM.[11]

Eurêka comporte un système d'aide pour faciliter l'interprétation des descripteurs normalisés. Lorsque le pointeur de la souris passe sur le nom d'un descripteur dans une fiche de résultat, il se transforme en point d'interrogation et une fenêtre flottante présente ce descripteur.

Les fonctions d'édition de fiches ***Eurêka*** offertes aux organismes francophones qui recensent des ressources d'enseignement et d'apprentissage de même que la capacité du système de consulter d'autres banques compatibles à la norme LOM font en sorte que le dépôt ***Eurêka***, qui contient déjà plus de 5000 ressources (REA), s'enrichira continuellement.

[10] Selon la norme IEEE 1484.12.1-2002 - Learning Object Metadata - http://ieeeltsc.org/wg12LOM/

[11] http://profetic.org/normetic2004/

Les ressources d'enseignement et d'apprentissage (REA) et les objets d'apprentissage

Depuis quelques temps l'expression « objet d'apprentissage » a cédé le pas au terme « ressource d'enseignement et d'apprentissage » (MLR: Metadata for Learning Resource)[12]. Pourtant ces deux expressions ne sont pas équivalentes. Qu'est-ce qu'une ressource d'enseignement et d'apprentissage ? Le groupe québécois de travail sur les normes (GTN-Québec) propose la définition suivante de REA. [13]

> Une ressource d'enseignement et d'apprentissage (REA) c'est « toute entité numérique ou non numérique, conçue ou pouvant être utilisée pour des fins d'apprentissage, d'éducation ou de formation ». [14] Un objet d'apprentissage (OA) est un élément d'apprentissage résultant du morcellement d'un contenu plus général. Un objet d'apprentissage c'est la plus petite partie cohérente d'une REA. Une REA ce peut être un logiciel d'édition et de communication (portail, moteur de recherche, répertoire, logiciels outils, applicatif de formation) ainsi que les données, les informations et les oeuvres numérisées (données statistiques ou informationnelles, références générales, oeuvres littéraires, artistiques ou autres) utiles à l'enseignant ou à l'apprenant dans le cadre d'une activité d'enseignement ou d'apprentissage (R. Bibeau, 2005b).

En pratique, une REA sur support numérique est souvent un fichier réutilisable pouvant être intégré dans une leçon ou un cours. Une animation Flash, une présentation PowerPoint, une composition musicale en MP3, un texte en format RTF, un scénario pédagogique ou une activité éducative en format PDF (R. Bibeau, 2005a), une collection d'images en format JPG, une séquence vidéo en format AVI, QuickTime ou MPEG, peuvent tous être considérés comme des REA.

Le fonctionnement du dépôt *Eurêka* [15]

« Tout le secret d'un bon dépôt de REA consiste en une description précise qui facilite le repérage, et les concepteurs d'***Eurêka*** l'ont compris. Un coup d'œil sur une fiche fournit immédiatement toute l'information pertinente, et ce, bien au-delà du titre et de la description de la ressource. Par exemple, des informations techniques (format, taille du fichier) ou pédagogiques (type de ressource, tranche d'âge, contexte), des renseignements sur les droits (coût, droits d'auteur), sur la contribution (auteur, établissement d'enseignement, etc.) et sur le classement de cette REA, etc. » (P.-J. guay, 2005).

Dans la fiche présentée à la figure 1, ci-dessous, un simple clic sur l'icône permettra de lancer le téléchargement du fichier de cette ressource menant à une

[12] jtc1 iso/iec sc36 http://mdlet.jtc1sc36.org/doc/SC36_WG4_N0103.pdf

[13] http://www.normetic.org

[14] http://www.profetic.org:16080/normetic2004/IMG/pdf/2005-GTN-16-10-2.pdf

[15] http://eureka.ntic.org/

page Web. Car il faut bien comprendre qu'un dépôt d'objets d'apprentissage ne contient habituellement pas les objets d'apprentissage (les unités de cours, les leçons, les exercices, les activités, les travaux pratiques, les descriptions de projets éducatifs, les références, les animations, les simulations, etc.) proprement dits mais seulement un descriptif indexé, normalisé et l'hyperlien vers la ressource elle-même qui demeure emmagasiné sur le serveur des propriétaires ou des mandataires de la ressource. Si des conditions d'accès à la ressource (abonnement ou coût d'utilisation) sont réclamés par les auteurs ou leurs ayants droits, ce sera à eux de collecter ces droits et non pas au dépôt d'objets comme ***Eurêka*** ; un peu comme dans une bibliothèque, ce n'est pas au documentaliste de percevoir les droits d'utilisation d'une œuvre quelconque.

« Dans la partie droite de la fiche descriptive ***Eurêka*** de la figure 1, « on trouve de l'information de nature pédagogique. La section du bas indique deux chemins possibles de navigation thématique conduisant à cette ressource. En suivant ces chemins, on pourra probablement trouver d'autres ressources apparentées et tout aussi pertinentes. Dans ***Eurêka***, l'utilisateur peut consulter la carte de visite de chaque organisme ou de chaque personne ayant contribué à la création d'une ressource. Un clic, et la carte de visite électronique en format vCard de l'auteur apparaît ! » (*ibid*.).

Figure 1 : Exemple de fiche de ressource dans Eurêka[16]

Recherche libre par mots clés ou recherche guidée par arborescence

On distingue deux façons de naviguer dans Internet : la recherche libre par mots clés (inscrire quelques mots décrivant ce que l'on cherche), et la recherche guidée par arborescence et par thème (naviguer dans des menus et des catégories). Les deux modes sont disponibles dans *Eurêka* : le mode recherche libre, simple ou avancée, permet de spécifier les paramètres de façon à restreindre la recherche à un niveau particulier ou de la limiter à l'intérieur d'un thème précis. Le mode de navigation guidée par thématique permet de fureter en descendant ou en remontant le long d'arborescences thématiques, un peu comme lorsqu'on bouquine dans les rayons d'une bibliothèque (Figure 2).

[16] http://eureka.ntic.org/display_lo.php?action=show&lom_id=1787

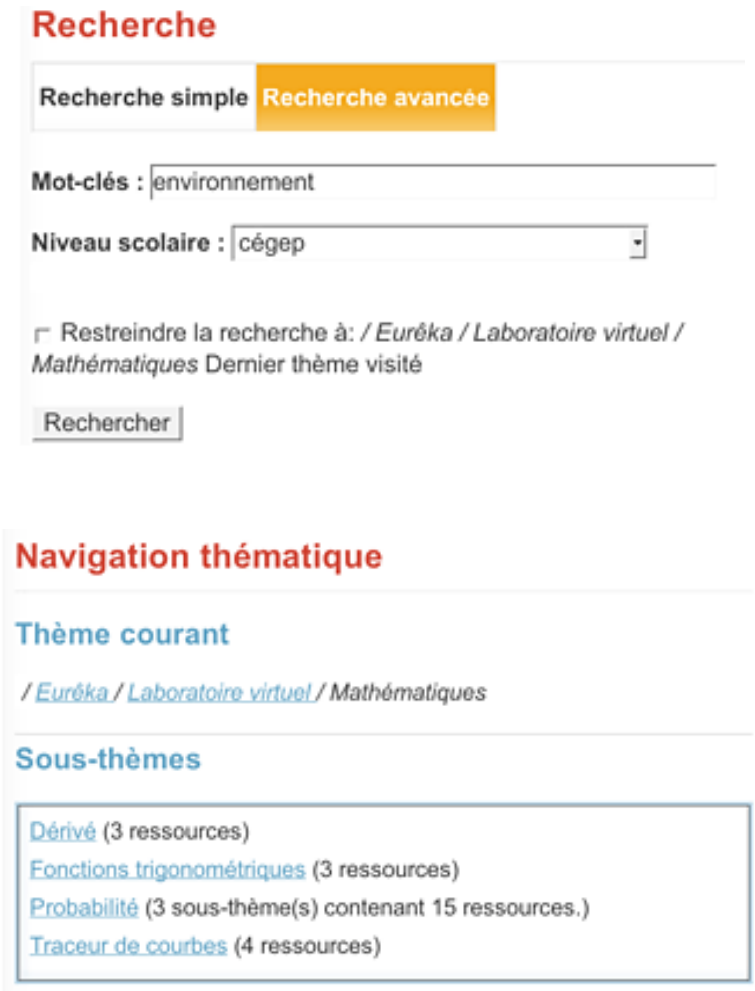


Figure 2 : Le répertoire de ressources d'enseignement et d'apprentissage Eurêka, *développé par la Vitrine APO, offre deux modes de navigation : la recherche par mots clés en mode simple ou avancé et la navigation guidée le long d'arborescences thématiques* (P.-J. guay, *op.cit*)

« Les résultats d'une recherche sont classés par pertinence selon le ou les descripteurs où une correspondance a été établie en fonction du nombre total de correspondances dans une fiche donnée. Le bouton  permet d'afficher la liste des champs dans lesquels une correspondance a été obtenue et souligne les occurrences. » (*ibid*.).

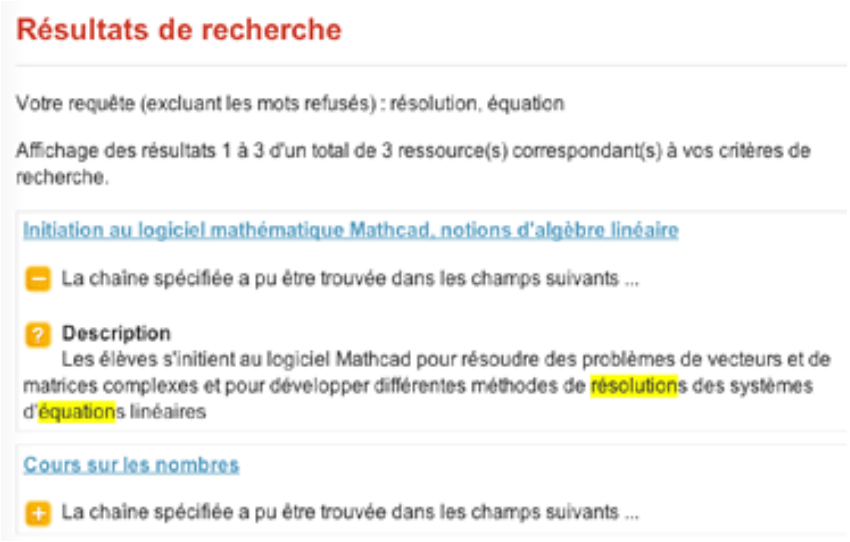


Figure 3 : Le détail des résultats de la recherche des mots résolution et équation est affiché afin d'évaluer la pertinence d'une ressource avant de consulter la fiche complète

Qui contribue à Eurêka ?

Les ressources indexées et cataloguées dans ***Eurêka*** sont décrites et validées par plusieurs partenaires TIC des ordres d'enseignement primaire-secondaire et collégial. On y trouve les logiciels du Centre Collégial de Développement de Matériel Didactique (CCDMD),[17] les cours du Cégep@distance,[18] des ressources offertes dans la plateforme de formation en ligne DECclic,[19] les rapports de recherche PAREA du Centre de documentation collégial, et naturellement, les répertoires de la Vitrine APO (Bibliothèque virtuelle des périodiques, Laboratoire virtuel, Index de sites francophones éducatifs et guide *Internet et éducation*.[20]

Parce qu'elles sont décrites dans un format normalisé, d'autres ressources sont également accessibles. C'est ainsi qu'on y trouve la collection RESPEL, le répertoire de ressources pédagogiques en ligne de la communauté wallonne de Belgique, partenaire du projet.[21]

Les enseignants sont également invités à enrichir le dépôt en remplissant un formulaire de suggestions qui sera traité par un des responsables de l'édition de

[17] http://www.ccdmd.qc.ca/

[18] http://www.cegepadistance.ca/

[19] http://www.decclic.qc.ca/

[20] htp://ntic.org

[21] http://www.enseignement.be/respel/RespelRech/RechMotsCle.aspx

ressources. Il n'est même pas nécessaire que la ressource suggérée soit déjà hébergée sur un serveur Web ; elle peut être directement déposée dans ***Eurêka***.[22]

Conclusion

La façon moderne de chercher des contenus éducatifs validés (REA) sur Internet est d'interroger un dépôt ou une moissonneuse de REA. Eurêka est un dépôt de métadonnées décrivant des REA disponibles sur divers serveurs un peu partout sur Internet. Certes, le fait de Produire des REA conformes aux standards et spécifications impose des efforts additionnels et des compétences nouvelles qui ne sont pas par défaut dans les établissements de formation mais seule l'utilisation d'une description normalisée permet d'éviter les écueils des vocabulaires et des classifications « locales ». Ainsi, en spécifiant des descriptifs standards de ces ressources, on évite les difficultés liées à l'utilisation d'une nomenclature locale. Le standard utilisé dans Eurêka est IEEE LOM (Learning Object Meta data), qui contient 58 descripteurs regroupés en différentes catégories (titre, description, format, cycle de vie, etc.). Plus précisément, est utilisé au Québec le profil Normetic, qui exige l'inscription de vingt-trois (23) des descripteurs du LOM.

Références bibliographiques

Bibeau, R. (2002), « Un guide de rédaction et de présentation d'un scénario pédagogique ». Thot-Cursus, http://thot.cursus.edu/rubrique.asp?no=16778.

Bibeau, R. (2005a), « Répertoire de répertoires de scénarios pédagogiques et d'activités d'apprentissage avec les TIC », Thot-Cursus, http://thot.cursus.edu/rubrique.asp?no=20976.

Bibeau, R. (2005b), « Les TIC à l'école : proposition de taxonomie et analyse des obstacles à leur intégration », Revue de l'ÉPI, décembre. http://www.epi.asso.fr/revue/articles/a0511a.htm.

Crepuq-Novasys (2003), *La description normalisée des ressources : vers un patrimoine éducatif,* Montréal, CREPUQ-Novasys, p. 17-18.

GTN-Q (2005), *Le profil d'application NORMETIC ; La description normalisée des ressources***,** ISO/IEC JTC1/SC36, WG4/N1035; WG4/N1036.

GUAY, P.-J. (2005), « Pédago-Dépôt : une visite d'Eurêka », Le bulletin Clic, no. 59, http://clic.ntic.org/clic59/eureka.html..

[22] http://eureka.ntic.org

Asinag-Asinag, 9, 2014, p. 197-206

Les Technologies de l'Information et de la Communication (TICs) au service de l'amazighe

Patrick Andries[1]
Lahbib Zenkouar[2]
Entretien réalisé par le Comité de Rédaction

Les réalisations de l'IRCAM, en termes de promotion de l'amazighe dans les TIC, ont été couronnées par l'homologation de l'ISO/Unicode et l'intégration directe dans les systèmes opérationnels les plus communs. À votre avis, dans quelle mesure ces avancées pourront-elles participer à la revitalisation de l'amazighe ?

Patrick Andries

Ces avancées peuvent participer à cette revitalisation dans la mesure où, d'une part, elles éliminent sur le plan pratique des obstacles à la production de textes amazighes à l'aide des outils informatiques modernes et, d'autre part, au niveau

[1] Membre expert du Consortium Unicode, il est l'un des rédacteurs de la proposition de normalisation des tifinaghes dans Unicode et l'ISO/CEI 10646. Il a également contribué à la normalisation de nombreux caractères dans ces normes. Il est le rédacteur et le webmestre du site dédié à Unicode http://hapax.qc.ca. Auteur de l'ouvrage « Unicode 5.0 en pratique » et de plusieurs contributions scientifiques dans le domaine de la typographie, il a dirigé le développement d'un navigateur internet multilingue et d'une bibliothèque logicielle d'internationalisation. Il exerce actuellement le métier de conseiller spécialiste dans la publication électronique, la gestion documentaire et l'internationalisation des applications.

[2] Lahbib Zenkouar est enseignant-chercheur à l'Ecole Mohammadia d'Ingénieurs et Professeur de génie électrique à cette école. Il est également chef de l'équipe travaillant sur les Techniques de Communications et Radiocommunications (TCR) au Laboratoire d'Electronique et Communication (LEC). Il est ex-directeur du Centre des Etudes Informatiques, des Systèmes d'Information et de Communication (CEISIC) de l'Institut Royal de la Culture Amazighe (IRCAM).

L. Zenkouar a conduit plusieurs projets concernant l'informatisation de l'écriture amazighe tifinaghe, notamment plusieurs normes dont la norme de codage 10646. Son domaine de recherche et développement concerne les systèmes d'information et de communications et l'informatisation de la langue amazighe.

L. Zenkouar est titulaire d'un Doctorat de l'université des sciences et techniques du Languedoc en microélectronique à Montpellier (France) et d'un Doctorat en sciences appliquées en télécommunications de l'Institut d'Electricité de Montefiore à Liège (Belgique).

symbolique, elles annoncent au grand public que cette langue peut s'exprimer par ces outils associés à la modernité, que cette langue peut s'inscrire dans cette modernité.

Si ces avancées techniques sont nécessaires aujourd'hui, elles ne sont toutefois pas suffisantes pour assurer la revitalisation de l'amazighe.

Lahbib Zenkouar

Ces réalisations sont primordiales pour le devenir de la langue amazighe car elles permettent à l'écriture amazighe d'être reconnue et manipulée par tous les outils de traitement électronique de l'information quel que soit leur type d'interface, notamment les ordinateurs et actuellement les portables mobiles qui en plus de la communication téléphonique intègrent de plus en plus quasiment toutes les fonctionnalités de l'ordinateur.

Par ailleurs, cette reconnaissance tout à fait technique basée sur l'attribution d'un code électronique à chaque symbole ou glyphe de l'alphabet amazighe, constitue une consécration internationale de l'écriture amazighe tifinaghe et marque de manière irréversible son inscription dans le patrimoine universel des écritures de notre humanité en tant qu'attribut propre à la civilisation amazighe.

Cette norme a permis l'élaboration de la norme de tri, basée sur les codes électroniques attribués aux lettres de l'écriture amazighe par la norme ISO-UNICODE, en plus d'une algorithmique propre à cette norme, permettant un aménagement du tri qui tient compte des caractéristiques propres de la langue parlée au niveau local.

Le codage ISO-UNICODE a permis également l'élaboration de la norme des claviers amazighes. Cette dernière a tenu compte de l'habitude acquise par les scripteurs et chercheurs amazighes imprégnés par le clavier azerty qui a précédé et permis de travailler normalement malgré une occupation locale du plan multilingue de base (BMP) supportant cette première version de fortune. Sur ce volet, la logique d'adoption des claviers a été respectée, et se révèle une question d'habitude plutôt que d'une étude détaillée sur la fréquence des lettres dans un texte.

Ce bref historique a permis d'aboutir à des claviers standards respectant la norme ISO-9995 des claviers et s'appuyant sur un codage réservé exclusivement à la langue amazighe.

Il convient de souligner que la norme de tri élaborée par l'IRCAM, comporte des clauses informatives de grande importance puisque celles-ci fournissent les correspondants latins et arabo-araméens du répertoire Tifinaghe. Bien évidemment, dans le cas de ces normes informatives, la langue parentale est soit le latin soit l'arabe. La seule écriture répertoriée sur le registre strictement amazighe, comme je l'ai souligné précédemment, est bien évidemment l'écriture tifinaghe et constitue

de ce fait une symbolique identitaire de l'appartenance amazighe que nous constatons à travers toute l'Afrique du Nord. Conclure que cette immense symbolique est partie des services de l'IRCAM est une fierté pour tous les chercheurs de l'Institut. Cet état d'esprit est une forme de cette revitalisation de l'amazighe. L'IRCAM prouve ainsi, grâce à son pragmatisme, son efficacité et son utilité sur à peine une décennie.

En outre, l'établissement de cette correspondance entre ces glyphes permettra de traduire en tifinaghe les textes berbères écrits en lettres arabes et enrichira ainsi la bibliothèque amazighe qui en a grand besoin.

De même, ils permettront une correspondance réciproque quasiment bijective entre l'alphabet amazighe écrit en tifinaghe et l'alphabet en latin, utilisé pour transcrire l'amazighe.

La stratégie poursuivie par le centre des technologies de l'information et partant de l'IRCAM, d'élaborer des polices de qualité et de styles différents a permis de nourrir cette écriture et constitue actuellement et à juste titre un atout de taille dans la communication visuelle de cette écriture dans l'espace public et les édifices de l'Etat.

Sur ce chapitre, la huitième chaîne de télévision nationale « Tamazight » aurait eu beaucoup de mal si elle n'avait pas disposé de ces polices réalisées par l'équipe des technologies de l'information et de la communication de l'IRCAM, suite aux normes adoptées. Nous constatons avec émerveillement que tous ces efforts ont permis de revitaliser pleinement l'écriture amazighe tifinaghe sur ce moyen de communication public pénétrant tous les foyers du Royaume.

Tous les éléments que je viens de citer sont des preuves de la revitalisation de la langue amazighe. Il reste cependant à faire converger grâce à ces atouts technologiques, les travaux des chercheurs de l'IRCAM sur la production d'un dictionnaire global, moderne et contemporain de notre ère. Tous les efforts des chercheurs doivent être dirigés dans ce sens.

Par ailleurs, il faut associer intiment à cet effort, celui des ouvrages pédagogiques d'apprentissage de l'écriture et de la langue amazighe.

En septembre 2012, Microsoft a annoncé, au siège de l'IRCAM à Rabat, la naissance de son nouveau système d'exploitation Windows 8 qui intègre l'amazighe dans son répertoire linguistique. Quelles sont, selon vous, les retombées de cet événement ?

Patrick Andries

Windows 8 a, en effet, amélioré la prise en charge de l'amazighe et du tifinaghe sur ce système d'exploitation très répandu. C'est une bonne nouvelle. On pouvait désormais choisir un clavier « tifinagh » livré d'office avec le système d'exploitation pour saisir des textes identifiés comme étant de l'amazighe, ou plutôt comme du « tamazight » selon la nomenclature choisie par Microsoft.

Il faut cependant apporter un bémol. Sous le capot, Microsoft a décidé pour Windows 8 que ce « tamazight » correspondrait à un indicatif de langue erroné. L'organisation internationale de normalisation, l'ISO, publie plusieurs listes d'indicatifs de langues que les ordinateurs utilisent pour étiqueter les textes informatiques. La série de normes internationales qui identifient les langues est connue comme l'ISO 639. Un texte français informatisé sera donc accompagné d'une étiquette ISO 639 valant « fr » pour préciser qu'il est écrit dans cette langue. Un texte anglais sera décoré d'un indicatif ISO 639 « en ». C'est très utile pour toute série de processus comme la vérification orthographique ou la traduction automatique.

Dans le cas de Windows 8, Microsoft a choisi d'utiliser l'indicatif [tzm] pour indiquer le « tamazight ». Or cette valeur ne correspond dans l'ISO 639 qu'à un des parlers amazighes du Maroc : la variante de l'Atlas central. Heureusement, grâce aux efforts de l'IRCAM, ce choix malencontreux sera corrigé dans la version 8.1 : l'amazighe marocain standardisé qui correspond au nouvel indicatif ISO 639 [zgh] sera pris en charge. Cet indicatif correspond à la norme amazighe commune préconisée par l'IRCAM pour l'ensemble des parlers amazighes du Maroc. Le nom non qualifié de « tamazight » ne sera plus non plus associé à l'indicatif [tzm]. Cet indicatif sera désormais décrit dans l'interface de Windows 8.1 de manière plus correcte comme l'amazighe de l'Atlas central.

Il faut enfin comprendre que dire que « l'amazighe est dans le répertoire linguistique » ne signifie pas que Windows corrige désormais, par exemple, l'orthographe amazighe ou que l'interface graphique de Windows est en amazighe. Non, il s'agit plutôt d'une étape nécessaire pour permettre la vérification orthographique et la traduction de l'interface : Windows sait maintenant comment désigner l'amazighe, il lui a attribué un code dans son répertoire linguistique. Une place est libre pour les outils et ressources amazighes, on peut donc maintenant greffer des outils linguistiques amazighes à Windows.

Lahbib Zenkouar

La retombée de cet événement, comme vous le dites, est la conséquence intégrale des aspects débattus dans la réponse à la première question. Il faut souligner, pour les profanes, que Microsoft ne peut réaliser que ce qui est inscrit dans les normes, notamment pour le codage Unicode. Mais, également pour le clavier amazighe marocain et pan-amazighe qui sont réalisés par ce fabricant de logiciels conformément à la norme. Il faut préciser que ces normes peuvent subir des amendements pour les rendre pratiques en fonction de l'évolution des techniques de communication et de l'interaction homme-machine. Il faut aussi souligner que les systèmes d'exploitation tels que présentés par Microsoft tendent vers de nouvelles configurations des OS (Operating System) qui s'apparentent aux formes des interfaces actuelles des portables téléphoniques mobiles.

Ce qui laisse supposer que les polices tifinaghes vont occuper les tablettes et en général, ces nouvelles architectures des téléphones mobiles.

La recherche scientifique dans le domaine de la langue et de la culture amazighes est appelée à évoluer au niveau de l'université marocaine. Quelles pistes envisagez-vous dans ce domaine ? Pensez-vous que ce domaine soit d'intérêt ?

Patrick Andries

À mon sens, le domaine du « génie linguistique » et de la recherche sur la langue amazighe est très prometteur et tout aussi original. Il reste tant de terrains à défricher, tant d'applications pratiques à créer et à diffuser.

Il suffit de penser aux outils de vérification orthographique puis grammaticale de l'amazighe ou à la mise à la disponibilité des usagers de ressources terminologiques en ligne comme le Trésor de la langue française[3], Termium[4] au Canada ou le Grand dictionnaire terminologique au Québec[5].

Ces sujets – qui reposent sur un travail linguistique fondamental où l'IRCAM a sans doute un rôle de coordination irremplaçable – sont du plus grand intérêt pour les jeunes chercheurs linguistes et informaticiens marocains attirés par des sujets d'envergure qui pourront avoir un effet des plus concrets sur les locuteurs amazighophones.

[3] http://www.cnrtl.fr/definition/

[4] http://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-fra.html?lang=fra

[5] http://gdt.oqlf.gouv.qc.ca/

La recherche scientifique dans le domaine de la langue et de la culture amazighes est appelée à évoluer au niveau de l'université marocaine. A votre avis, quelle serait la meilleure démarche pour consolider la place de la langue amazighe et assurer la continuité de son développement au sein des structures et laboratoires de recherche œuvrant dans le domaine des TICs ?

Lahbib Zenkouar

Cette question est d'une grande importance pour le devenir de la langue et de la culture amazighes. Il faut absolument se pencher sur cette problématique pour pouvoir tirer les conclusions adéquates de manière à assurer le développement de ce secteur. La recherche universitaire est une grande opportunité pour une véritable revitalisation dans l'espace universitaire de la langue et de la culture amazighes qu'il convient d'encadrer et de diriger.

Il faut encourager la recherche extra-IRCAM tout azimut, tout en essayant de développer des axes prioritaires sur les quels se pencheront plusieurs laboratoires. Mais comment procéder ?

A mon avis, il faut tenir un workshop international dont le seul thème est le développement de la recherche concernant l'amazighe et l'organiser de manière à répondre aux objectifs de cette recherche.

Je pense qu'il faut s'inspirer fortement de l'Institut pour l'arabisation ou créer un homologue dont l'objectif est l'amazighisation ou, à défaut, un comité constitué des chercheurs de plusieurs centres et dont la mission aura pour objectif la réflexion sur les moyens de transmission de la langue amazighe.

Nous constatons aujourd'hui un processus de virtualisation du savoir et des communautés à travers l'Internet et les réseaux sociaux. Le support informatique prime ainsi, de plus en plus, sur les autres moyens de communication. Que pensez-vous de la place de l'amazighe dans ce nouvel espace ? Quelles sont les contraintes qui pèsent sur la numérisation dans le domaine amazighe et quels seraient, d'après vous, les moyens nécessaires pour les surmonter ?

Patrick Andries

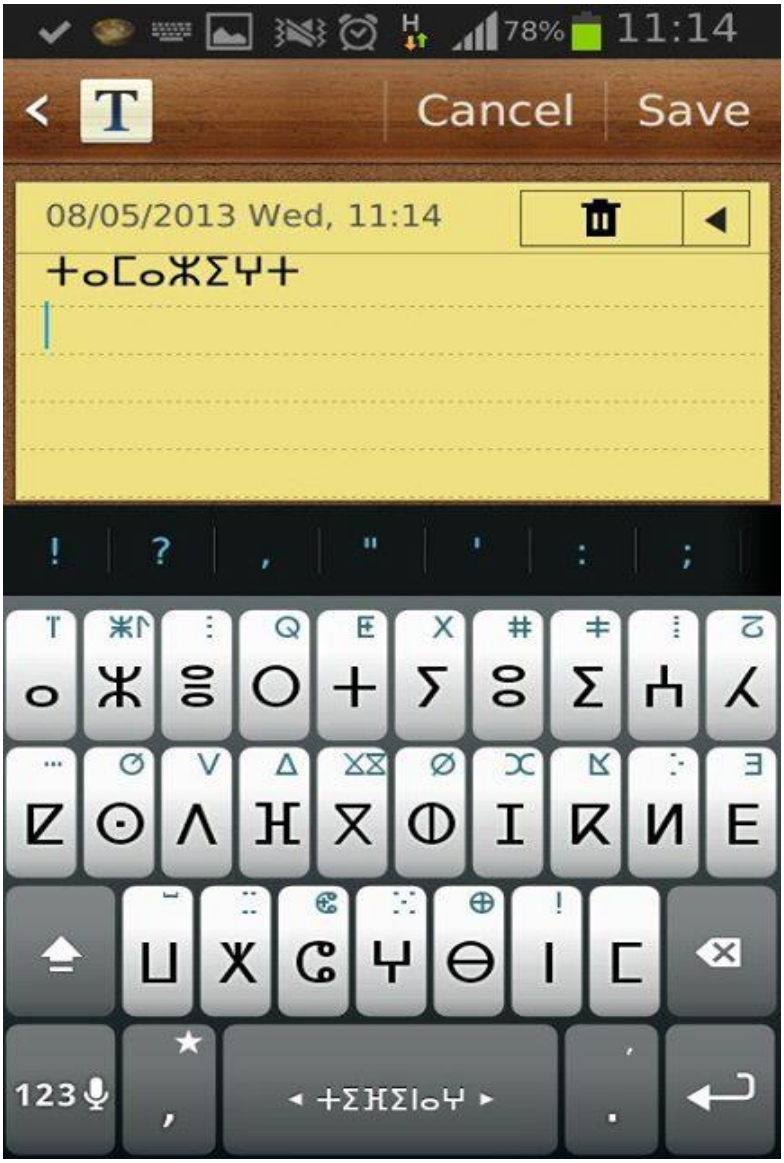
Les contraintes qui pèsent sur cette numérisation de l'amazighe sont celles où les outils numériques sont mal adaptés à l'amazighe et affichent un retard par rapport à des langues comme le français ou l'arabe.

Nous avons déjà parcouru un bon bout de chemin : un système d'exploitation comme Windows comprend d'office une police tifinaghe, un clavier tifinaghe, les noms de fichiers peuvent être en tifinaghe. On peut également s'échanger des

courriels, afficher des pages internet en tifinaghe grâce aux normes informatiques qui prennent désormais en charge les tifinaghes.

Il faudra bien sûr, comme nous l'avons vu ci-dessus, développer de nouveaux outils qui permettent la production de qualité de texte amazighe plus facilement : correcteur orthographique, grammatical, ressources lexicographiques en ligne, à savoir les outils informatiques utilisés pour produire du matériel numérique dans des langues concurrentes. C'est une œuvre de longue haleine, à mon sens essentielle.

Il faut aujourd'hui s'assurer que les nouvelles plateformes, sans doute aussi importantes que les PC, voire plus, offrent les mêmes outils que ceux disponibles sur Windows. Il faut donc que les téléphones intelligents et les tablettes numériques au Maroc proposent au strict minimum un clavier tifinaghe, des polices tifinaghes. Certaines initiatives privées existent déjà ; elles permettent ainsi l'ajout facile d'un clavier et de polices tifinaghes à des téléphones cellulaires Galaxy sous Android (voir ci-dessous).



Enfin, peut-être, faudra-t-il donner un coup de pouce aux producteurs de contenu numérique amazighe en fournissant une aide technique et des contenus de référence en ligne dont les producteurs de contenu pourraient s'inspirer ou qu'ils pourraient même copier gratuitement pour faciliter et baisser le coût de production des contenus amazighes. On peut notamment penser à la publication en ligne de textes classiques, de contes, de proverbes, d'informations historiques ou de

chansons (également sous format vidéo) en amazighe. Le rôle de facilitateur d'une institution comme l'IRCAM est, bien sûr, primordial dans ce domaine.

Lahbib Zenkouar

La virtualization de l'imprimerie a eu lieu il y a plus d'une décennie. Dans les programmes d'action de l'IRCAM, une grande partie des projets est consacrée à cet aspect. Mais il y avait tellement d'aspects qu'une certaine confusion pouvait régner à ce moment-là ou, du moins, laisser paraître une forme de confusion pour les profanes.

A ma connaissance, tous les écrits de l'IRCAM sont faits sous impression numérique. La bibliothèque de l'IRCAM est, par conséquent, de plain-pied dans la virtualisation tout en s'inscrivant également sur le support papier et, donc, de l'écrit classique qu'il ne faut à aucun prix abandonner ou laisser en second lieu. L'écrit classique sur papier doit être la priorité « number one » de la bibliothèque amazighe. Bien évidement, la virtualisation permettra une plus large diffusion et un bénéfice maximum au profit du public intéressé ; ce qui constitue le principal objectif de l'IRCAM.

Pour revenir à un aspect de la virtualisation, je donnerai pour exemple ce qui est fait sur le site de l'école amazighe qui a permis de mettre à disposition les ouvrages didactiques et qui a contribué ainsi à faire bénéficier la communauté amazighophone des énormes efforts pédagogiques d'apprentissage réalisés par l'IRCAM.

Il faut, à mon avis, continuer à améliorer les services offerts par les logiciels relatifs à l'édition, à la collecte et à la classification des ouvrages traitant des aspects de la langue et de la culture amazighes. Il faut développer la recherche sur l'indexation des textes, la recherche thématique, le traitement automatique des langues naturelles, la recherche de l'information. Bien évidemment, cet aspect des choses se fera dans le cadre de la recherche développée à l'intérieur et à l'extérieur de l'IRCAM.

Sur d'autres aspects de la virtualisation, il faut créer un dialogue avec la communauté amazighe et créer des forums spécialisés. Il reste à développer cette approche et à la rendre ciblée et pertinente.

En outre, tous ces aspects sont complémentaires et peuvent être traités par des laboratoires universitaires et répondraient ainsi aux problématiques posées dans la question précédente.

Sur un autre plan, il faut numériser les écrits anciens pour éviter leur perte, améliorer leur présentation grâce aux logiciels de traitement d'images et essayer de les traduire en tifinaghe pour enrichir et agrandir la bibliothèque amazighe.

Je pense que tous les corpus qui permettront de travailler et de réfléchir en amazighe et sur l'amazighe doivent être initiés et soutenus quelle que soit leur forme d'expression, qu'elle soit orale, écrite ou visuelle. C'est, à mon avis, un énorme chantier numérique.

En conclusion, j'ai la ferme conviction qu'à l'aube de la nouvelle constitution qui consacre la langue amazighe langue officielle du Royaume du Maroc, l'IRCAM, pionnier valeureux de cet effort culturel, peut se révéler insuffisant en nombre de chercheurs. En plus de cette illustre institution à laquelle il faut, à mon avis, attribuer un rôle d'expertise et de coordination, il faut créer des institutions régionales, développer et soutenir la recherche dans les universités, créer des centres de recherches spécialisés pour arriver à suivre et à mettre sur pied dans les plus brefs délais une langue efficace capable de donner corps à cette nouvelle réalité de la langue amazighe.

Comptes rendus

Compte rendu de l'ouvrage d'Ahmed Boukous intitulé : *Revitalisation de la langue amazighe* et sous-titré : *Défis, enjeux et stratégies*, publié par l'Institut Royal de la culture amazighe, série : Etudes n° 22, Imprimerie Top Press, Rabat, 2012.

Physiquement, le livre d'Ahmed Boukous comporte 354 pages. Après les premières pages s'ouvrant sur un bel exergue correspondant à un passage emprunté à Amin Maalouf préconisant le droit de chaque citoyen à reconnaître une part de lui-même dans la culture mondiale émergente dans le nouveau millénaire, puis après le rappel de quelques titres publiés par l'auteur, vient un sommaire, des pages présentant le protocole de transcription ainsi que le système de transcription adoptés. Ensuite un prologue d'une dizaine de pages (pp. 1-10) met en contexte la problématique abordée dans sa complexité et résume les axes essentiels des thématiques abordées.

L'ouvrage s'organise en deux grandes parties. Dans la première (pp. 11 – 120), composée de 5 chapitres, l'auteur fait le diagnostic de la situation linguistique de l'amazighe en relation avec les autres langues au Maroc, aussi bien sur le plan national que supranational. Il dresse un bilan des divers facteurs qui menacent la survie de l'amazighe ou entravent son évolution. Situation que résume bien la notion d'*attrition* langagière.

La deuxième partie (pp. 121 – 310), composée de 7 chapitres, passe en revue les éléments porteurs d'espoir susceptibles d'assurer la *revitalisation* de l'amazighe dont les plus essentiels sont représentés par :

- l'importance du cumul réalisé par les travaux de description de la langue et de la culture amazighes depuis le protectorat à nos jours (après séparation prudente entre ce qui relève de l'idéologie et ce qui relève de la science) ;
- le succès des stratégies adoptées au niveau de l'action de la société civile et de la réponse positive apportée par l'Etat aux aspirations des défenseurs de l'amazighe (fait que résume bien la notion de *résilience*) ;
- les efforts prometteurs de codification, normalisation et standardisation de l'amazighe ;
- la possibilité de contourner les étapes intermédiaires entre le local et le global grâce au concept de *glocalisation* dont l'application permet « de fournir un levier de maîtrise des effets de la dominance induits par la globalisation et ceux des mécanismes de résilience générés par les spécificités locales » (p. 313). Ensuite, vient un épilogue servant de conclusion à l'ouvrage (pp. 311 – 326).

Le livre se termine par une bibliographie bien fournie, un index des termes, une table des figures et une table des matières.

Sur le fond, la description que fait l'auteur de la situation linguistique marocaine est devenue un classique dans les travaux de sociolinguistique sur le Maroc, notamment la hiérarchie proposée des strates linguistiques qui y coexistent : la strate locale (l'amazighe), la strate centrale (l'arabe), la strate supercentrale (le français) et la strate hypercentrale (l'anglais).

Le livre explore des notions qui pourraient sembler, de prime abord, relever d'une simple terminologie subjective (résilience, revitalisation, attrition, glocalisation, obsolescence,…), mais qui constituent de véritables concepts puisés dans la littérature spécialisée. Le pari de cette situation est justement de garder suffisamment de distance par rapport à sa subjectivité afin d'asseoir ses questionnements et son argumentation sur des bases objectives. De sorte que l'auteur s'efforce, dans le style adopté, « d'établir l'équilibre entre la règle de la méthode scientifique et les principes de l'éthique et de l'équité dont le chercheur citoyen ne peut se départir » (p. 10).

Ceci a une incidence indéniable sur son style d'écriture où chaque mot est pesé, où il s'interdit tout dérapage émotif, rattrapé cependant par le sens commun que pourraient suggérer des termes comme « revitalisation », « mise à mort », « mise en danger » des langues.

Le livre présente une multiplicité d'angles de vue, sur la situation des langues au Maroc, et, en particulier, celle de l'amazighe. Le même objet est abordé selon plusieurs points de vue ; le politologue, le sociologue, le sociolinguiste, le linguiste, y trouvent tous leur compte.

La documentation et les références bibliographiques utilisées sont d'un intérêt primordial aussi bien du point de vue de leur pertinence que de leur actualité, les travaux classiques ne sont pourtant pas négligés.

La question de la variation linguistique est abordée sur la base d'un examen minutieux du système phonologique et morphologique des parlers étudiés en relation avec l'éclairage apporté par les variables sociolinguistiques considérées. On reconnaît le travail de l'homme de terrain.

Par ailleurs, le dialectologue est satisfait de recommandations, concernant les travaux hérités du colonialisme, de ne pas jeter le bébé avec l'eau du bain.

Le livre lui-même reflète l'approche dialectique entre l'objet d'analyse conçu comme une entité abstraite et son ancrage dans la société : incessant va et vient entre les préoccupations académiques chiffrées et les notions d'éthique et d'équité.

Le style est dominé par une approche « économiste » de la langue : notion de marché, échange, gestion, offre, globalisation ainsi que le concept de conflit, violence symbolique (inspiré de Bourdieu). De telle manière qu'en plus de servir d'outil de positionnement dans la hiérarchie sociale, la langue est « un indicateur de développement cognitif de l'individu ». Ce qui met en parallèle le développement socio-économique avec le degré d'accès à la connaissance. L'auteur recourt abondamment à la matrice SWOT utilisée dans le domaine de la gestion des entreprises.

Face au « poids » des forces sociales pratiquement inconscientes qui relèguent l'amazighe dans une position de langue minorée, l'auteur entrevoit, pour faire pencher la balance, l'opportunité d'un recours volontariste à l' « émergence de la conscience identitaire communautaire » tout en signalant le danger probable du communautarisme.

Suite à l'institutionnalisation de l'amazighe comme langue officielle, à côté de l'arabe standard, il appelle à plus d'implication des décideurs politiques, la société civile ayant accompli, en grande partie, sa tâche.

Le cheminement de la pensée de l'auteur se fait avec prudence, la progression se fait du simple au complexe, anticipant les objections des contradicteurs éventuels. Il est tenu compte d'un préalable important : celui de la nécessité de séparer le discours *épilinguistique* (registre des représentations que se fait le locuteur de sa langue et explications spontanées relevant d'un amateurisme de bonne foi mais souvent passionnel) du discours *métalinguistique* (fondé sur une approche plus prudente, demandant sans cesse à l'évidence de se justifier).

Le projet de codification et de standardisation de l'amazighe ayant suscité beaucoup de polémique, l'auteur, sans contourner la difficulté, expose pas à pas les diverses étapes de la réflexion à ce sujet, l'adoption et l'adaptation de l'alphabet tifinaghe, les diverses niveaux des propositions de standardisation : supranational, national, régional et local, ainsi que leurs avantages et inconvénients. On peut souligner une conception révolutionnaire de la standardisation entrevue par l'auteur : une *standardisation raisonnée*. Celle-ci, loin de limiter l'avenir de la sauvegarde de l'amazighe au seul locuteur amazighophone appelé à agir au niveau de la transmission générationnelle ainsi qu'à l'action militante sur le plan civil et national, et loin de le faire dépendre uniquement de « la conscience et de l'engagement de la communauté concernée » (p. 8, citant Landry et *al.* : 2005), pourrait répondre « aux besoins des apprenants dont l'amazighe, local ou régional, n'est pas langue première, d'autant plus que la Constitution consacre le caractère national de l'amazighe » p. 257).

Ainsi pourrait se dessiner, à notre sens, un espoir de voir s'élargir la base de la masse parlant l'amazighe en misant aussi, pour la promotion de cette langue, de façon parallèle à la transmission générationnelle, sur les locuteurs dont l'amazighe n'est pas la langue maternelle. Cette voie pourrait favoriser une généralisation de l'enseignement de l'amazighe dans une perspective de standardisation qui serait envisagée comme « la construction » du standard national à partir de la capitalisation des convergences interdialectales et l'enrichissement du vocabulaire commun par le moyen de la néologie lexicale et de la terminologie pour créer les technolectes nécessaires » (p. 245).

Par ailleurs, à l'issue de cette brève synthèse, nous nous permettrons de formuler une remarque. Comme corrélat à cette conception innovante de la standardisation, nous pensons que, bien que compréhensible sur le plan historique, la réflexion proposée par l'auteur, sur l'amazighité (pp. 317-321), et ce malgré la référence au discours royal du 17 octobre 2001 où il est affirmé que « l'amazighité occupe une position centrale dans l'identité nationale, qu'elle constitue une culture en partage entre les différentes sensibilités du peuple marocain » (p. 138), cette réflexion donne l'impression de se limiter au seul rapport entre amazighe et amazighité et ne développe pas la conception originale mentionnée plus haut, susceptible, pour sa part, de contribuer au renforcement de la masse parlante de l'amazighe standard en cours de construction. Sur le plan de la recherche en dialectologie également, elle ne permet pas de combler la fracture entre la recherche sur l'arabe marocain et celle portant sur les diverses variétés de l'amazighe, qui semblent se tourner le dos.

Ainsi, le sentiment pour certains locuteurs d'appartenir au monde de l'amazighité bien qu'ayant l'arabe comme langue maternelle (soit par rupture de la courroie de transmission générationnelle, soit du fait de l'appartenance à un espace historiquement arabisé mais où le substrat amazighe est patent à tous les niveaux : phonique, morphosyntaxique et lexical, sans oublier le toponymique et le patronymique), ne trouverait pas son expression dans l'amazighité réduite au seul rapport entre l'amazighe et la communauté amazighophone. En effet, il nous semble que les travaux d'investigation sur l'arabe marocain sont susceptibles d'apporter des éclairages sur des états de la langue amazighe concernant son système phonologique, morphosyntaxique et lexical. La recherche sur la structure et l'histoire de l'arabe marocain relève aussi, pour une grande part, à notre sens, de la réflexion sur l'amazighité, entendue en tant qu'ensemble de valeurs culturelles qui constitue le socle de notre personnalité et qui transcende l'amazighophonie.

Fouad Brigui
Université Sidi Mohammed Ben Abdallah, Fès

Résumés de thèses

Amrouche Mustapha (2012), *Reconnaissance de caractères, de textes et de documents basée sur les modèles de markov cachés*, Université Ibn Zohr, Faculté des Sciences d'Agadir.

Mots clés : Reconnaissance automatique de l'écriture, Ecriture Amazighe, Tifinagh, Approches syntaxiques, automates à états finis, Réseaux de neurones.

Les travaux de recherche que nous avons menés s'intéressent au développement des méthodes de reconnaissance de caractères manuscrits et imprimés et de textes en tenant compte du contexte par combinaison de niveaux d'analyse et de connaissances morphologiques. Nous proposons ainsi deux approches de reconnaissance de l'écriture arabe et amazighe.

En effet et en premier temps, nous avons développé une approche de reconnaissance de caractères isolés, basée sur les primitives directionnelles obtenues à l'aide de la technique des fenêtres glissantes à partir de la transformée de Hough de caractère. L'approche conçue adopte une modélisation markovienne de type modèle discriminant qui consiste à associer un ou plusieurs modèles par classe. Selon cette méthode, la reconnaissance s'effectue en estimant les probabilités d'émission de la suite d'observations de la forme à reconnaître par les différents modèles préalablement construits. La forme à reconnaître est affectée à la classe dont le modèle qui maximise la probabilité. Cette approche est pratiquement utilisée dans le cas où le nombre de classes à reconnaître est relativement limité (application à vocabulaire limité). Toutefois, elle devient coûteuse en temps de calcul et espace mémoire quand ce nombre dépasse le millier, puisque chaque classe possède au moins un modèle qui lui est propre. Nous évaluons le système de reconnaissance proposé sur des bases de données de caractères arabes et amazighes.

L'approche proposée donne de bons résultats, bien qu'elle ne tienne pas compte des caractéristiques morphologiques de l'écriture étudiée. En effet, nous avons évalué les performances de notre système sur la base des caractères Tifinaghs **AMHCD**[1] **[1]** avec deux variantes. La première adopte la modélisation discrète des probabilités d'émission, la seconde utilise les HMMs continus.

A partir de la base **AMHCD,** nous avons constitué deux parties : apprentissage et test.

Pour la partie *apprentissage,* nous avons 17160 exemples de caractères, soit 2/3 de la base **AMHCD** ; pour la partie *test*, 8580 exemples caractères, soit 1/3 de la base ladite base. Avec la modélisation discrète, nous avons obtenu un taux de reconnaissance de 90, 4%. Dans le cas continu, on procède par une modélisation des densités des probabilités par des gaussiennes. Nous avons effectué une série d'expérimentations sur la totalité de la base AMHCD. A l'aide de ces expériences,

[1] Youssef Es Saady, Ali Rachidi, Mostafa El Yassa and Driss Mammass. Article: AMHCD: A Database for Amazigh Handwritten Character Recognition Research. International Journal of Computer Applications 27(4):44-, August 2011. Published by Foundation of Computer Science, New York, USA.

nous avons évalué le taux de reconnaissance de notre approche en fonction de nombre d'états par modèle HMM et de nombre de gaussiennes. En effet, nous avons utilisé cinq topologies qui varient entre 6 et 14 états pour étudier l'influence de ces paramètres sur les performances pour une modélisation des émissions par une seule ou deux composantes gaussiennes. Le tableau ci-dessous présente les résultats obtenus sur cette base.

Nombre d'états	6	8	10	12	14
Nombre de gaussiennes	1-2	1-2	1-2	1-2	1-2
Taux de reconnaissance	96,21%	96, 56%	96, 88%	97, 38%	97, 89%

Dans un second temps, nous avons proposé une deuxième méthode pour la reconnaissance automatique hors ligne de caractères Tifinaghes imprimés. La méthode proposée est basée sur un chemin discriminant (DP-HMM) opérant sur un vocabulaire de base formé de différents graphèmes fondamentaux. Le vocabulaire est généré en se basant sur les caractéristiques morphologiques de la graphie amazighe. Un seul modèle HMM global construit et entraîné sur les éléments du lexique proposé par des primitives structurelles et géométriques. Chaque chemin au long de ce treillis représente une séquence de segments, qui constitue un caractère de l'alphabet Tifinaghe. Pour ce faire, les caractères d'entrées sont pré-classés en deux groupes (forme circulaire et non circulaire). Par la suite, ils sont décrits par leurs points d'intérêts et leurs segments. La reconnaissance s'effectue en décodant dynamiquement le chemin optimal suivant le critère de maximum de vraisemblance.

Les taux obtenus ont montré la robustesse de l'approche proposée. En effet, pour valider le système proposé, nous avons effectué des expérimentations significatives à l'aide de Toolkit (HTK) sur la totalité de la base de données de patterns de la graphie amazighe[2] ***(BD1)***. Nous avons constitué, à partir de cette base, deux ensembles distincts de données, un ensemble *A (A=2/3)* pour l'apprentissage et un ensemble *B (B=1/3)* pour les tests.

Plusieurs tests ont été effectués pour évaluer le taux de reconnaissance du système en fonction de nombre d'états et de nombre de mélange de gaussienne. Par ailleurs, nous avons effectué les premiers tests sur toute la base de patterns de la graphie amazighe (BD1 : contient 19437 caractères multi fonts c'est-à-dire. 627 échantillons x 31 classes). Le Tableau ci-dessus présente les résultats obtenus de ces tests sur la base BD1, en utilisant les modèles mono-gaussiens, les modèles à deux gaussiens et les modèles à trois gaussiens.

Nombre d'états	3	5
Nombre de mélange de gaussienne	1-2-3	1-2-3
Taux de reconnaissance	99, 38%	99,72%

[2] Y. Ait Ouguengay, M. Taalabi (2009), « Elaboration d'un réseau de neurones artificiels pour la reconnaissance optique de la graphie amazighe: Phase d'apprentissage », *Systèmes intelligents-Théories et applications*, Paris : Europia, cop. (impr. au Maroc).

Essaady Youssef (2012), *Contribution au développement d'approches de reconnaissance automatique de caractères imprimés et manuscrits, de textes et de documents Amazighes,*Université Ibn Zohr, Faculté des Sciences d'Agadir.

Mots clés : Modéles de Markov cachés, reconnaissance de l'écriture manuscrite et imprimée, primitives structurelles et directionnelles, transformation de Hough.

Cette thèse a pour objet principal la reconnaissance automatique hors ligne de l'écriture amazighe. Dans ce cadre, nous avons d'abord construit une base de données, nomée AMHCD, de caractères amazighes manuscrits composés de plus de 25.000 caractères isolés écrits par 60 scripteurs différents. Cette base a été utilisée pour évaluer et tester les résultats de nos travaux. Elle est destinée aussi à servir d'autres chercheurs dans le domaine de la reconnaissance de l'écriture amazighe manuscrite.

Ce travail de recherche propose deux approches de reconnaissance automatique de l'écriture amazighe qui ont contribué à améliorer les performances. La première approche est syntaxique ; elle utilise des automates à états finis avec des primitives structurelles pour reconnaitre les caractères amazighes imprimés. Elle s'intéresse à la forme du caractère tifinagh qui est composé de primitives structurelles telles que des segments, des points et/ou des petits cercles. Après les prétraitements, des algorithmes appropriés permettent de construire la chaîne du codage de Freeman représentant le caractère en entrée. La chaîne est utilisée dans l'entrée de l'automate maximal canonique, qui reconnaît tous les caractères amazighes segmentés pour décider la classe d'appartenance du caractère. Cet automate est construit à partir des automates spécifiques de chacun des caractères amazighes imprimés. Sur une base de 630 caractères amazighes imprimés isolés, les résultats expérimentaux montrent la solidité de l'approche. Sur 630 caractères, 589 ont été reconnus, soit un taux de reconnaissance de 93,49%. Les erreurs de reconnaissance proviennent de la forme de certains caractères non reconnus dont le squelette comporte plus des segments non orthogonaux. La limite de cette approche est qu'elle ne traite pas les caractères circulaires. De plus, les caractères amazighes manuscrits ne peuvent être pris en compte par cette approche.

Afin de remédier à ces limites, nous avons développé un deuxième système de reconnaissance de l'écriture amazighe basé sur la ligne centrale horizontale du caractère. Ce système est basé sur une approche neuronale qui utilise un réseau de neurones multicouches comme classifieur. Après des prétraitements sur l'image d'entrée, le texte est segmenté en lignes et puis en caractères isolés. Les positions des lignes centrales horizontales du caractère sont utilisées pour obtenir un ensemble de caractéristiques indépendantes et dépendantes à ces lignes. Ces caractéristiques sont liées aux densités de pixels et sont extraites sur les images binaires des caractères en se basant sur l'utilisation de la technique des fenêtres glissantes. Le système a montré de bonnes performances sur une base de 19437 paternes amazighes imprimés et sur 20150 caractères amazighes manuscrits de la base AMHCD. La base des paternes imprimés utilisée contient à peu près vingt mille patterns. Il s'agit d'une base des patterns de différentes fontes amazighes et de tailles variées. Elle contient au total 12 polices de caractères et les tailles du 10

points au 28 points pour chaque modèle. Les patterns sont fournis sous forme d’images bitonales de tailles variables.

Une amélioration de ce système a été proposée en intégrant d'autres caractéristiques basées sur la ligne centrale verticale du caractère. Cette amélioration a donné de bons résultats. En effet, pour la base des patterns imprimés, le taux de reconnaissance est 98,49% lors de l'intégration des caractéristiques basées sur la position de la ligne centrale horizontale et augmente à 99,28% lors de l'ajout des caractéristiques basées sur la position de la ligne centrale verticale. Pour la base AMHCD de caractères amazighes manuscrits, le taux augmente de 92.23 % à 96.32 % lors de l'ajout des caractéristiques basées sur la position de la ligne centrale verticale du carctère. Les causes d'erreurs sont principalement dues à une grande similarité morphologique entre certains caractères amazighes et, parfois, sur des fontes différentes.

Guide de rédaction de la revue □□□□□□-*Asinag*

Conditions générales

- Tout article proposé doit être original, accompagné d'une déclaration de l'auteur certifiant qu'il s'agit d'un texte inédit et non proposé à une autre publication.
- Le compte rendu de lecture doit avoir pour objet la lecture critique d'une publication récente (ouvrage, revue ou autres) en la situant dans l'ensemble des publications portant sur le thème concerné.
- Tout article publié dans la revue devient sa propriété. L'auteur s'engage à ne pas le publier ailleurs sans l'autorisation préalable du Directeur de la revue.
- Les textes non retenus ne sont pas retournés à leurs auteurs. Ceux-ci n'en seront pas avisés.

Présentation de l'article

- Une page de couverture fournira le titre de l'article, le nom, le prénom, l'institution, l'adresse, le numéro de téléphone, le numéro de fax et l'adresse électronique de l'auteur. Seuls le titre de l'article, le nom et le prénom de l'auteur et le nom de son institution doivent figurer en tête de la première page du corps de l'article.
- Les articles seront envoyés par courrier électronique sous forme de fichier attaché en format Word ou RTF (Rich Text Format) à l'adresse suivante : « ***asinag@ircam.ma*** ».
- L'article ne dépassera pas 15 pages (Bibliographie et moyens d'illustration compris).
- Le texte sera rédigé en police **Times**, taille 12, interligne **1**, sur des pages de format (17*24). Le texte en tifinaghe doit être saisi en police **Tifinaghe-ircam Unicode**, taille 12, téléchargeable sur le site Web de l'IRCAM « http://www.ircam.ma/lipolicesu.asp ». Pour la transcription de l'amazighe en caractères latins, utiliser une police Unicode (**Gentium**, par exemple).
- Le titre est d'environ 10 mots et peut être suivi d'un sous-titre explicatif. Il sera rédigé en gras, de police Times et de taille 14.
- Le résumé des articles ne dépassera pas 10 lignes.

Moyens d'illustration

- Les tableaux sont appelés dans le texte et numérotés par ordre d'appel (chiffres romains). La légende figurera en haut des tableaux.
- Les figures et les images sont appelées dans le texte et numérotées par l'ordre d'appel en chiffres arabes. La légende sera donnée en dessous des figures.

Références bibliographiques et webographiques

- Les références bibliographiques ne sont pas citées en entier dans le corps du texte, ni dans les notes. Sont seulement indiqués, dans le corps du texte et entre parenthèses, le nom de/des auteurs suivi de la date de publication du texte auquel on se réfère et, le cas échéant, le(s) numéro(s) de la/des page(s) citée(s). Si les auteurs sont plus de deux, indiquer le nom du premier auteur, suivi de « et al. ».

Ex. : (Geertz, 2003) ; (Pommereau et Xavier, 1996) ; (Bertrand et *al*., 1986) ; (Bouzidi, 2002 : 20).

Dans le cas de plusieurs publications d'un auteur parues la même année, les distinguer à l'aide de lettres de l'alphabet en suivant l'ordre alphabétique (1997a, 1997b, etc.).

Ex. : (Khair-Eddine, 2006a) ; (Khair-Eddine, 2006b).

Lorsque plusieurs éditions d'une même référence sont utilisées, on signalera la première édition entre crochets à la fin de la référence dans la liste bibliographique.

- Les références bibliographiques complètes, classées par ordre alphabétique des auteurs, sont fournies à la fin de l'article (sans saut de page).

✓ Les titres des ouvrages sont présentés en italique.

Les références aux **ouvrages** comportent dans l'ordre : le nom de l'auteur et l'initiale de son prénom, l'année de parution entre parenthèses, suivie, s'il s'agit de l'éditeur, de la mention (éd.), le titre, le lieu d'édition, le nom de l'éditeur. Toutes ces indications seront séparées par des virgules.

***Ex.* :** Cadi, K. (1987), *Système verbal rifain, forme et sens*, Paris, SELAF.

✓ Les titres d'articles de revue, de chapitres d'ouvrages, etc. se placent entre guillemets.

Les références aux **articles de revue** comportent (dans l'ordre) : le nom et l'initiale du prénom de l'auteur, l'année d'édition, le titre de l'article entre guillemets, le titre de la revue en italique, le volume, le numéro et la pagination. Toutes ces indications seront séparées par des virgules.

Ex. : Peyrières, C. (2005), « La recette de notre caractère », *Science & Vie Junior*, n° 195, p. 48-51.

✓ Les références aux **articles de presse** comportent seulement le titre entre guillemets, le nom du journal en italique, lieu d'édition, la date et le numéro de page.

Ex. : « Les premiers pas du supermarché virtuel », *l'Economiste*, Casablanca, 26 octobre 2007, p. 17.

✓ Les références aux **chapitres d'ouvrages collectifs** indiquent le nom et le prénom de l'auteur, le titre du chapitre, la référence à l'ouvrage entre crochets : […].

✓ Les références aux **actes de colloques** ou **de séminaires** doivent comporter le nom et la date du colloque ou du séminaire.

Ex.: Boukous, A. (1989), « Les études de dialectologie berbère au Maroc », in *Langue et société au Maghreb. Bilan et perspectives*, Actes du colloque organisé par la Faculté des Lettres et des Sciences Humaines-Rabat en octobre et décembre 1986, p. 119-134.

✓ Les références **aux thèses** : elles sont similaires aux références aux ouvrages, on ajoute l'indication qu'il s'agit d'une thèse, en précisant le régime (Doctorat d'Etat, Doctorat de 3ème cycle…) et l'université.

Ex. : Hebbaz, B. (1979), *L'aspect en berbère tachelhiyt (Maroc)*, Thèse de Doctorat de 3ème cycle, Université René Descartes, Paris V.

- Les références **webographiques** : il est nécessaire de mentionner l'URL (Uniform Resource Locator) et la date de la dernière consultation de la page web.

Ex. : http://fr.wikipedia.org/wiki/Langue_construite, octobre 2007.

Notes, citations et abréviations

- Dans le cas où des notes sont fournies, celles-ci sont en bas de page et non en fin d'article. Il faut adopter une numérotation suivie.
- Citations : les citations de moins de cinq lignes sont présentées entre guillemets « ... » dans le corps du texte. Pour les citations à l'intérieur des citations, utiliser des guillemets droits « ... "..." ... ». Les citations de plus de quatre lignes sont présentées sans guillemets, après une tabulation et avec un interligne simple.
- Toute modification d'une citation (omission, remplacement de mots ou de lettres, etc.) est signalée par des crochets […].

Sous-titres : le texte peut être subdivisé par l'utilisation de sous-titres en caractères gras.

Italique : éviter de souligner les mots, utiliser plutôt des caractères en italique.

- Si l'auteur emploie des abréviations pour se référer à certains titres qui reviennent souvent dans l'article, il devra les expliciter dès leur premier usage.

Ex. : Institut Royal de la Culture Amazighe (IRCAM).

REVUE ⴰⵙⵉⵏⴰⴳ- Asinag

Bulletin d'abonnement

Périodicité : 2 numéros par an

Bulletin à retourner à :

Institut Royal de la Culture Amazighe

Avenue Allal El fassi, Madinat al Irfane, Hay Riad. B.P. 2055 Rabat

Tél : (00212) 537 27 84 00 – Fax : (00212) 537 27-84-36

e-mail :abonnement@ircam.ma

Titre	*Maroc Prix /an	*Etranger Prix /an	Quantité	Total
□□□□□□- *Asinag*	100 Dh	30 €		

*Les frais d'expédition sont inclus dans ces tarifs (Maroc et étranger)

Nom, prénom : ……..……...

Etablissement : ……..……...

Adresse : ……..……...……..

Pays : ……..…….

Code postal : …….Ville : ... ……...

Tél. : ……..Fax : ……...……..

Je désire souscrire un abonnement à la Revue ⴰⵙⵉⵏⴰⴳ- *Asinag de :*

☐ 1 an

☐ 2 ans

Mode de paiement :

☐ Chèque bancaire à l'ordre de ……

☐ Virement bancaire

Préciser les noms et adresse de l'abonné.

Banque……N° de compte : ……

Date : Signature

مجلة أسيناگ-ⴰⵙⵉⵏⴰⴳ

قسيمة الاشتراك

تصدر هذه المجلة بمعدل عددين في السنة

ترسل قسيمة الاشتراك بالبريد العادي الى العنوان التالي :

المعهد الملكي للثقافة الأمازيغية
شارع علال الفاسي، مدينة العرفان، حي الرياض ص.ب. 2055 الرباط
الهاتف: 00 84 27 537 (00212) الفاكس: 36-84-27 537 (00212)
البريد الإلكتروني : abonnement@ircam.ma

العنوان	*المغرب الثمن / سنة	*باقي الدول الثمن / سنة	الكمية	المجموع
أسيناگ- ⴰⵙⵉⵏⴰⴳ	100 Dh	30 €		

* بما فيه مصاريف الإرسال (المغرب وباقي الدول)

الاسم و النسب:..

المؤسسة : ....

العنوان :..

البلد : ....

المدينة :...................................: الرمز البريدي.................................

الهاتف :...................................: الفاكس.................................

أريد الاشتراك في مجلة أسيناگ-ⴰⵙⵉⵏⴰⴳ لمدة: ☐ سنة

☐ سنتين

طريقة الأداء:

☐ شيك بنكي لفائدة ....

☐ التحويل البنكي

يجب تحديد اسم وعنوان المشترك.

المؤسسة البنكية.................................رقم الحساب...

التاريخ : التوقيع :