المعهد الملكي للثقافة الأمازيغية
ⵉⵔⵉⵎ ⴰⴳⵍⴷⴰⵏ ⵏ ⵜⵙⴷⵓⵙⵜ ⵜⴰⵎⴰⵣⵉⵖⵜ
INSTITUT ROYAL DE LA CULTURE AMAZIGHE

Les technologies de l'information

Opportunités pour le développement de l'amazighe

Coordination
Ait Ouguengay Youssef

2012

Les Technologies de l'information: Opportunités pour le développement de l'amazighe

المملكة المغربية
المعهد الملكي للثقافة الأمازيغية
ⴰⵙⵉⵏⴰⴳ ⴰⴳⵍⴷⴰⵏ ⵏ ⵜⵙⵏⵉⵍⵜ ⵜⴰⵎⴰⵣⵉⵖⵜ
INSTITUT ROYAL DE LA CULTURE AMAZIGHE

Le Centre des Etudes Informatiques, des Systèmes d'Information et de Communication (CEISIC)

Le 3ème atelier international sur les TIC et l'amazighe
TICAM'08

Les Technologies de l'information: opportunités pour le développement de l'amazighe

Actes de l'atelier

Rabat, 24 – 25 Novembre 2008
ⵕⵕⴱⴰⵟ, 24-25 ⵏⵓⵡⴰⵏⴱⵉⵔ 2008

Publications de l'Institut Royal de la Culture Amazighe

Centre des Etudes Informatiques, des systèmes d'Informations et de Communication

Série : Colloques et séminaires N° 32

Titre	:	Les Technologies de l'information: opportunités pour le développement de l'amazighe
Coordination	:	Ait ouguengay Youssef
Edition	:	Boumediane Mounia
Éditeur	:	Institut Royal de la Culture Amazighe
Réalisation et suivi	:	Centre des Etudes Informatiques, des Systémes d'Information et de Communication.
Couverture	:	Unité d'édition - CTDEC
Dépôt légal	:	2012 MO 1393
ISBN	:	978-9954-28-117-8
Imprimerie	:	El Maarif Al Jadida - RABAT
Copyright	:	IRCAM ©

Atelier International sur l'amazighe et les nouvelles technologies de l'information

Depuis l'avènement de l'Internet, le devenir de l'informatique est complètement défini par les mutations que subit la grande toile mondiale. Cette dernière évolue la façon dont les gens communiquent. Qu'il s'agisse des services commerciaux, des accès aux réseaux de données ou du développement des logiciels, les utilisateurs des produits informatiques ont besoins de retrouver leur identité dans les outils qu'ils manipulent et de communiquer en leur langue qu'ils connaissent le plus.

L'intégration de la langue et de la culture à travers l'internationalisation du logiciel informatique trouve là son champ d'application par excellence. En fait, l'internationalisation est un travail technique qui prépare un logiciel à son localisation et son adaptation à une ou plusieurs langues et cultures particulières.

Entre concepts de développements informatiques et spécificités de la langue cible, se pose des difficultés de cohérence de sens des mots et des expressions et de mise en forme (la diréctionalité par exemple) du texte traduit et intégré dans le logiciel.

Pour arriver à sa fin, l'internationalisation met en œuvre plusieurs aspects de la technologie d'information qui s'entre crochent pour arriver aux fins de la localisation informatique comme le codage informatique utilisé (Unicode, ISO 10646), les techniques de la traduction appliquées aux TICs (règles de programmation, etc.), les normes de la terminologie respectés (ISO 704, etc.) et les normes de traitement informatique de(s) langue(s) en question.

Quant au contexte de la langue et de la culture amazighes, il est à notre sens plus que jamais favorable pour émerger une nouvelle expérience dans le domaine du logiciel informatique, d'autant que le Tifinaghe, l'alphabet de l'amazighe, est Intégré dans l'Unicode / ISO 10646 et un chantier de recherche très important est ouvert pour l'amazighe et des efforts d'intégration de l'amazighe dans les nouvelles technologies d'information ne cesse d'émerger. Le présent recuil, rassemble les travaux qui ont été exposé durant cette 3éme Edition du colloque TICAM'08 sur l'état des lieux et les opportunités du développement de l'amazighe dans les technologies de l'information.

Table des matières

Indiquer la langue, l'écriture, le pays dans des documents informatiques

Patrick Andries
Conseils Hapax, Québec, Canada
Membre du consortium Unicode
patrick@hapax.qc.ca

Résumé: Dans cette communication nous verrons comment préciser la langue, le dialecte ainsi que d'autres informations culturelles comme le système d'écriture utilisé ou le pays d'où provient le scripteur, et les appliquer à des données Unicode. Ces informations peuvent se révéler précieuses pour une kyrielle de processus de traitement des documents.

In this paper, we will review how to specify the language, the script and the country of an electronic document, or parts thereof. This information, as we will see, is valuable for a series of automated text processes.

Mots-clés. Unicode, langue, écriture, pays, dialecte, locale, amazighe, informatique, français, chinois,tifinaghe, Unicode, arabe, ISO 639, ISO 3166, RFC 5646, RFC 4646, ISO 15924, BCP 47, IETF.

1. Introduction

Préciser la langue, ainsi que d'autres informations culturelles comme le pays d'où provient le scripteur, et les appliquer à des données Unicode permet :

- de résoudre les ambiguïtés d'affichage (exemple : « désunifier » les caractères CJC),
- de rendre la synthèse vocale possible,
- d'obtenir les bonnes ressources linguistiques d'un programme internationalisé
- d'afficher la bonne version linguistique d'une page internet,
- de mieux effectuer la coupure de lignes et de mots,
- d'obtenir de meilleurs résultats quand on classe, cherche ou trie ces données,
- et de permettre la correction orthographique.

En règle générale, on peut dire que la langue est orthogonale au codage de caractères : connaître le codage d'un texte ne permet pas d'en deviner la langue. C'est ainsi qu'un texte codé en Latin-1 peut très bien être en anglais, en espagnol ou en français, voire les trois à la fois.

Nous allons passer en revue ci-dessous les normes qui établissent la manière de préciser la langue et d'autres métadonnées culturelles d'importance.

2. ISO 639 – indicatifs de langue

L'ISO 639 est une norme internationale qui définit des indicatifs pour la représentation des noms de langues. Elle est actuellement composée de 3 parties. Sa première partie, l'ISO 639-1[1] (alpha-2), utilise des codets ou indicatifs sur 2 caractères, et les associe à des noms de langue en français et en anglais. L'ISO 639-2[2] (alpha-3) utilise des codets sur 3 caractères et connaît deux formes de codages possibles : ISO 639-2/B (bibliographique) et ISO 639-2/T (terminologique). En règle générale, les indicatifs bibliographiques ressemblent à ceux définis par la norme américaine Z39.53 et s'inspirent des noms qui désignent ces langues en anglais, alors que les indicatifs terminologiques ressemblent aux noms que ces langues se donnent. Enfin, l'ISO 639-3[3] complète l'ISO 639-2 ; ces indicatifs à trois lettres sont tirés de la base de données de l'Ethnologue[4].

Tableau1.Les différentes parties de l'ISO 639

Partie	Type d'indicatif	Nombred'indicatifs	Exemples
ISO 639-1	à deux lettres	136	« fr » pour le français, « wa » pour le wallon, « ar » pour l'arabe.
ISO 639-2	à trois lettres	484	« fre » et « fra » pour le français, « ber » pour le berbère
ISO 639-3	à trois lettres	7581	« fra » pour le français, « rif » pour le rifain.

L'ISO 639 n'est pas un registre stable. On y ajoute, de temps à autre, de nouveaux indicatifs et plusieurs langues en ont changé. C'est le cas de l'hébreu, de

[1] Liste non officielle sur <http://fr.wikipedia.org/wiki/Liste_des_codes_ISO_639-1>.
[2] Liste sur <http://www.loc.gov/standards/iso639-2/php/French_list.php>.
[3] Liste sur <http://www.sil.org/ISO639-3/codes.asp>.
[4] Voir <http://www.ethnologue.com/>.

l'indonésien et du yidiche qui, pour deux d'entre eux, sont passés à des indicatifs qui s'inspirent de leur nom ou graphie en anglais plutôt que dans la langue en question (de « iw » à « he » et de « ji » à « yi »). Les logiciels doivent prendre en charge les deux versions de ces indicatifs et les considérer comme synonymes.

Tableau2. Quelques indicatifs tirés de l'ISO 639-1

Indicatif ISO 639-1	Langue
ar	arabe
de	allemand
el	grec
en	anglais
es	espagnol
fr	français
it	italien
nl	néerlandais
pt	portugais
ru	russe

Vingt-deux langues ont dans l'ISO 639-2 deux codets (indicatifs) différents : l'un bibliographique (ISO 639-2/B) et l'autre terminologique (ISO 639-2/T). C'est le cas du français qui s'écrit « fre » (bibliographique) ou « fra » (terminologique). Il faut considérer, à nouveau, ces indicatifs comme synonymes. En pratique, ces doublons sont rarement utilisés puisque ces langues ont également un indicatif de deux lettres (ISO 639-1) et que les normes et standards préconisent alors l'utilisation de celui-ci.

L'ISO 639-2 prévoit une zone à usage privéqui s'étend de « qaa » à « qtz », l'ISO n'affectera ces codets à aucune langue normalisée. Ces indicatifs peuvent être utilisés par accord commun entre des tiers.
Dans les applications qui utilisent les codets de langue ISO 639, il est préférable d'utiliser en premier lieu un indicatif alpha-2 de l'ISO 639-1, s'il existe. Si ce n'est le cas, on choisira le codet alpha-3 de l'ISO 639-2/T. Enfin, en dernier ressort, on pourra utiliser les indicatifs alpha-3 de la norme ISO 639-3.

2.1 Macrolangue

Certains des indicatifs de l'ISO 639-3 correspondent à des macrolangues. Une macrolangue est un ensemble de langues étroitement apparentées ou de dialectes fortement divergents. On compte 56 langues dans ISO 639-2 qui sont considérées comme des macrolangues dans ISO 639-3. L'arabe (« ar » dans ISO 639-1 et « ara » dans ISO 639-2) est une de ces macrolangues dans l'ISO 639-3 (« ara »). Le chinois est également considéré comme une macrolangue (« zh ») et une des langues de cette macrolangue est « cmn », appelée le mandarin. Notons que, dans le cas de l'arabe et du chinois, les communautés locales ont du mal à admettre le verdict des linguistes. Ces érudits considèrent habituellement que l'arabe n'est pas une langue, mais que le marocain et le syrien sont des langues distinctes. Les arabophones, en revanche, ne partagent pas cette analyse.

Tableau3.Les « langues berbères » dans l'ISO 639-3

Indicatif ISO 639-3	Langue
shi	tachelhit ou chleuh
tzm	tamazight (centre du Maroc)
rif	tarifit ou rifain
cnu	chénoui
shy	tachaouit, chaoui
kab	kabyle
thv	tamachek (Sud algérien)
thz	tamachek (Agadez)

L'ISO 639-1 et 2 reflétaient en gros les besoins des bibliothécaires, soucieux de simplifier la classification en évitant de multiplier les langues. L'ISO 639-3, pour sa part, adopte plutôt le point de vue des linguistes qui tendent à voir nettement plus de langues. Ce débat entre les « fusionneurs » (les bibliothécaires) et les « diviseurs » (les linguistes) se poursuit. L'intégration de l'ISO 639-3 illustre l'importance des linguistes (et du SIL d'où proviennent ces indicatifs) dans la conception du registre des langues.

Quant à l'ISO 639-5, il définit des codes alpha-3 (à trois lettres) pour les familles ou groupes de langues. Certains de ces codes font également partie de l'ISO 639-2, où se mêlaient aussi des codes pour des macro-langues et langues individuelles, mais pas de façon assez précise ni complète.

Tableau4.Exemples de famille et groupes de langues dans l'ISO 639-5[5]

Indicatif ISO 639-5	Groupe	Hiérarchie
afa	Langues afro-asiatiques	afa
ber	Langues berbères	afa : ber
cdc	Langues tchadiques	afa : cdc
cus	Langues couchitiques	afa : cus
egx	Langues égyptiennes	afa : egx
ine	Langues indo-européennes	ine
itc	Langues italiques	ine : itc
omv	Langues omotiques	afa : omv
roa	Langues romanes	ine : itc : roa
sem	Langues sémitiques	afa : sem

3. ISO 3166 – indicatifs de pays

Les indicatifs de pays normalisés par l'ISO 3166-1[6] forment un deuxième type de renseignement culturel. L'ISO 3166-1 comprend trois répertoires différents : alpha-2, alpha-3 et numérique-3. L'ISO 3166-1 alpha-2 définit une série d'indicatifs de pays sur deux lettres. Alpha-3 précise des codets de pays sur 3 lettres. Il existe également une norme ISO 3166-2, codée à l'aide de quatre lettres, qui permet de désigner des subdivisions de pays.

Tableau5. Quelques indicatifs de pays ISO 3166-1 à deux lettres

Pays	Indicatif
Algérie	DZ
Belgique	BE
Canada	CA
Égypte	EG
Espagne	ES

[5] Pour une liste complète : <http://www.loc.gov:8081/standards/iso639-5/fr.php> .
[6] On peut consulter gratuitement la liste officielle complète à l'adresse suivante : <http://www.iso.org/iso/fr/country_codes/iso_3166_code_lists/french_country_names_and_code_elements.htm>

Pays	Indicatif
France	FR
Libye	LY
Mali	ML
Malte	MT
Maroc	MA
Mauritanie	MR
Niger	NE
Tchad	TD
Tunisie	TN

Les codets ISO 3166-1 (sur deux lettres) correspondent habituellement aux noms de domaines de tête par pays[7] (en anglais, « Country Coded Top LevelDomains », abrégé en *ccTLD*) utilisés dans l'attribution des adresses internet (voir le « ma » dans « ircam.ma »). Il existe quelques exceptions dont la plus notable est sans doute celle du Royaume-Uni qui a un indicatif 3166 égal à « GB » (Grande-Bretagne), mais dont le domaine internet est « .uk » (« United Kingdom », Royaume-Uni).

On remarque que, à l'instar de l'ISO 639, l'ISO 3166 n'est pas un ensemble figé d'indicatifs, certains pays naissent (par exemple à la suite de l'éclatement de l'Union soviétique), d'autres changent de nom et voient leur indicatif tomber en désuétude ou être repris par un autre pays. C'est le cas des Territoires des Afars et Issas, aujourd'hui Djibouti, dont le codet AI a été repris par l'île d'Anguilla. Il en va de même du codet GE, anciennement attribué aux îles Gilbert et Ellice devenues Tuvalu, et maintenant utilisé par la Géorgie. On imagine facilement les problèmes de données historiques qui utiliseraient les codets 3166 pour indiquer un pays. Ce genre de complication est d'ailleurs à la source d'une récente norme internet sur laquelle nous reviendrons bientôt : le RFC 5646.

La norme **ISO 3166-2**(seconde partie de la normeISO 3166), édictée par l'Organisation internationale de normalisation, permet de désigner les principales subdivisions administratives d'unpayspar uncodeten quelques chiffres ou lettres complétant le code ISO 3166-1 du pays.

[7]Pour plus de détails sur ces termes techniques, voir l'autre communication de ces mêmes actes : *Demain, encore plus de tifinaghes sur Internet.*

Tableau6. Quelques codets de l'ISO 3166-1 et de l'ISO 3166-2

Indicatif	Signification
FR-01	Département de l'Ain en France
MA-CAS	Province de Casablanca
MA-OUA	Province d'Ouarzazate
MA-TET	Province de Tétouan
MA-01	Région Tanger-Tétouan
MA-06	Région de Meknès-Tafilalet

4. M.49 – Indicatifs de pays et de régions

Les codets ONU M. 49 sont des indicatifs à trois chiffres attribués par la division de la statistique de l'ONU. Ils ne correspondent pas toujours à des pays ; le codet 001, par exemple, représente le monde entier, 002 l'Afrique et 015 l'Afrique septentrionale. Le codet à trois chiffres d'un pays défini dans l'ISO 3166-1 numérique-3 est identique au codet M.49 défini pour le même pays.Toutefois, certains indicatifs M.49 n'ont pas de correspondance ISO 3166-1 quand ils représentent une région supranationale ou infranationale[8].

Tableau7. Quelques indicatifs M.49

Indicatif	Signification
012	Algérie
434	Libye
466	Mali
478	Mauritanie
504	Maroc
562	Niger
788	Tunisie

[8] Liste complète officielle ici : <http://unstats.un.org/unsd/methods/m49/m49alphaf.htm>.

5. ISO 15924 – indicatifs d'écriture

L'ISO 15924 définit un indicatif pour près de 150 écritures différentes. Nous en reproduisons quelques-uns dans le tableau 4[9]. Rappelons que le français, par exemple, relève de l'écriture latine (Latn) et l'amazighe normalisé au Maroc du tifinaghe (Tfng).

Tableau 8.*Quelques indicatifs d'écriture*

Indicatif	Signification
Arab	Arabe
Copt	Copte
Cyrl	Cyrillique
Grek	Grec
Hani	Idéogrammes japonais
Hans	Idéogrammes chinois simplifiés
Hant	Idéogrammes chinois traditionnels
Latf	Latin brisé (« fraktur » ou gothique)
Latn	Latin
Tfng	Tifinagh (tifinar).

6. RFC 5646 – étiquettes de langue

L'Internet Engineering Task Force, en abrégé IETF, littéralement le « Détachement d'ingénierie d'Internet », est un regroupement informel d'experts qui élabore les standards de l'Internet. Les *Request for comment* (RFC, les *demandes de commentaires*) sont une série de documents portant sur l'Internet émis par l'IETF. Peu de ces RFC sont des standards, mais tous les standards de l'Internet sont enregistrés en tant que RFC. Chaque RFC porte un numéro séquentiel unique. Une fois adopté, un RFC n'est jamais modifié ou retiré. Si un RFC doit être modifié, on publie un autre RFC (avec un autre numéro) qui le remplace.

[9] La liste complète peut être consultée à l'adresse suivante :
<http://www.unicode.org/iso15924/iso15924-fr.html>

C'est ce qui s'est produit en 2009 quand le RFC 5646 a complété et remplacé le RFC 4646 qui définissait déjà des étiquettes linguistiques. La plus grande innovation du RFC 5646 est l'introduction des milliers de codets de l'ISO 639-3 et ceux de l'ISO 639-5 qui définit les ensembles de langues (par exemple « afa » pour « les langues afro-asiatiques »). Son prédécesseur, le RFC 4646, remplaçait déjà un standard précédent sur le même sujet, le RFC 3066 qui lui-même remplaçait le RFC 1766. Toutes les étiquettes de langue définies selon le RFC 1766 restent conformes aux RFC 4646 et 5646. Les étiquettes de langue du RFC 5646 consistent en une série de sous-étiquettes séparées par des traits d'union, sans égard à la casse. Voici un aperçu des règles de construction des étiquettes définies par le RFC 5646 (nous expliquerons cette « grammaire » ci-dessous) :

```
Une étiquette de langue consiste en
étiqlang           ; étiquette générative
              -ou- usage-privé        ; une étiquette privée
              -ou- patrimonial        ; anciennes valeurs
                                      ; enregistrées

étiqlang       = (langue
                  ["-" écriture]
                  ["-" région]
                  ("-" variante)*
                  ("-" extension)*
                  ["-" usage privé])

langue = [a-wA-W]{2,3}; codet ISO 639 le plus court
["-" extlang]; suivi d'un extlang optionnel ou
|[a-wA-W]{4}; pour normalisation ultérieure[10] ou
| [a-wA-W]{5,8}        ; une valeur enregistrée auprès de
                              ; l'IANA[11].

extlang = [a-wA-W]{3}; jusqu'à 3 codets ISO 639 séparés
                            ; par un "-", en pratique un seul
                            ; codet est utilisé

écriture     = "Latn", "Cyrl" ; codets ISO 15924 (4 lettres)

région       = "US", "CS", "FR"... ; codets ISO 3166 à 2 lettres
               "419", "019"...     ; codets ONU M.49 à 3 chiffres

variante     = "rozaj", "1996"...  ;  plusieurs sous-étiquettes
                         ; permises de 4 à 8
                                   ;alphanumériques

extension    = une lettre suivie de sous-étiquettes, plusieurs
```

[10] Réservé pour faire face à l'épuisement possible des codets ISO 639 à 3 lettres.
[11] Pour les très rares cas où l'IANA accepterait d'enregistrer une langue refusée par l'ISO 639. À l'heure actuelle, aucune sous-étiquette de ce type n'a été attribuée.

```
extensionssont permises pour une même étiquette de
langue.

usage-privé  = "x-" suivi de sous-étiquettes, autant que
nécessaires, peuvent être en début ou en fin
d'étiquette, mais pas au milieu.

patrimonial  = étiquettes reprises de l'ancien registre (RFC 3066)
etqui ne sont pas des doublons(une liste fermée).
```

Chaque type de sous-étiquette a une longueur précise et des restrictions quant à son contenu. L'étiquette commence toujours par la sous-étiquette « langue » qui peut être un codet ISO 639 ou une autre valeur enregistrée auprès de l'IANA[12]. Elle peut être suivie de différentes sous-étiquettes. Il existe à l'heure actuelle cinq types de sous-étiquettes qui peuvent suivre l'indicatif de langue : l'écriture, la région, les variantes, les extensions et l'usage privé. L'ordre, la longueur et le contenu de chaque sous-étiquette sont bien établis.

Toutes les sous-étiquettes légitimes sont consignées dans un registre unique tenu à jour par l'IANA, plutôt que les différentes agences de mise à jour des normes ISO comme c'était le cas pour le RFC 3066. Pour la sous-étiquette « langue », l'IANA n'enregistre qu'un seul codet par langue, alors que l'ISO pouvait en avoir normalisé plusieurs (trois pour le français par exemple : « fr », « fra » et « fre »). Si un codet ISO à deux lettres est disponible, celui-ci apparaîtra dans le registre plutôt que le codet à 3 lettres. Voici l'entrée pour la sous-étiquette de langue « es » dans le registre de l'IANA[13] :

```
%%
Type: language
Subtag: es
Description: Spanish
Description: Castilian
Added: 2005-10-16
Suppress-Script: Latn
%%
```

Laligne « Suppress-Script » indique qu'il ne faut pas mentionner l'écriture de cette langue quand elle est égale à « Latn ». Ceci afin de décourager la prolifération d'étiquettes pléonastiques comme « es-Latn ». Le « Suppress-Script » est également utile pour éviter que de nombreux anciens analyseurs d'étiquettes de langue conçus pour le RFC 3066 dont la syntaxe était, *grosso modo*, « langue » ou « langue-région » ne se plantent quand ils sont confrontés à des étiquettes de langues inutiles. C'est le cas, par exemple, pour les documents à l'étiquette « fr-

[12] L'IANA est un organisme américain responsable de la gestion de l'espace d'adressage IP d'Internet et d'autres ressources requises par les protocoles de communication sur Internet.
[13] Le registre se retrouve ici : <http://www.iana.org/assignments/language-subtag-registry>.

CA » (français du Canada) valable déjà avec le RFC 3066 et qu'il est inutile de changer en « fr-Latn-CA » car cela pourrait causer des problèmes aux anciens analyseurs.

6.1. Extlang

L'indicatif de langue ISO 639 peut être suivi d'un « extlang » après un « - ». Le premier indicatif est la sous-étiquette principale de langue. L'extlang, s'il est présent, doit précéder toute autre sous-étiquette comme l'écriture ou la région.

Exemples d'étiquette linguistique qui utilisent une sous-étiquette extlang :

- `zh-yue` (chinois cantonais)
- `ar-ary` (arabe marocain)

Le but principaldes combinaisons`<langue principale>-<extlang>` est de prendre en charge des formes historiques d'étiquette linguistique avant, notamment, l'inclusion des milliers de codets de l'ISO 639-3 dans le RFC 5646.Pour chaque forme `<langue principale>-<extlang>`, il existe une forme équivalente qui n'utilise qu'un indicatif `<langue principale>` sans extlang. Cette forme simple est recommandée et il faut l'utiliser dans la mesure du possible. C'est ainsi qu'on préfèrera `yue` à `zh-yue` pour le cantonais et `ary` à `ar-ary` pour l'arabe marocain.

Voici l'entrée du registre qui correspond à l'extlang pour l'arabe algérien du Sahara (`aao`) :

```
%%
Type: extlang
Subtag: aao
Description: Algerian Saharan Arabic
Added: 2009-07-29
Preferred-Value: aao
Prefix: ar
Macrolanguage: ar
```

La `<langue principale>`utilisée avec un extlang est une macrolangue[14] qui comprend un certain nombre de langues ou de dialectes fortement divergents. La sous-étiquette de la macrolangue peut s'utiliser seule, toutefois si son sens n'est pas suffisant clair, le lecteur confronté à un tel document pourrait bien ne pas nécessairement pouvoir le lire.

[14] Il existe des macrolangues dont les langues constitutives ne peuvent servir d'extlang, c'est le cas du cri au Canada (la macrolangue) et le cri des Plaines ou le cri de Moose (les langues englobées par la macrolangue). On pourra écrire « cr » (le cri macrolangue), « crk » (cri des Plaines), mais pas « cr-crk ».

C'est ainsi que `zh` signifie « chinois», un concept qui recouvre de nombreux « dialectes »[15] qui ne sont pas nécessairement compréhensibles entre eux. Quand on utilise « `zh` » sans plus, on fait le plus souvent référence à la variante dominante dans l'ensemble chinois, à savoir le mandarin (`cmn`), bien qu'il s'agit là d'une convention tacite sur laquelle BCP 47 ne dit mot.

Pourquoi les formats avec et sans extlang sont-ils permis ? La principale raison, comme nous l'avons évoqué, est que le RFC 4646, le prédécesseur du RFC 5646, contenait déjà des étiquettes du type `zh-yue` où `zh` était la langue principale et `yue` désignait une variante. C'était une des seules manières[16] de désigner le cantonais qui ne bénéficiait d'aucun indicatif ni dans l'ISO 639-2 ni dans l'ISO 639-1. Avec l'adjonction de l'ISO 639-3 lors de l'adoption du RFC5656, il existe désormais un codet précis qui désigne directement le cantonais (`yue`) qui peut désormais servir dans la sous-étiquette de `<langue principale>`.

Bien que les formes sans extlang soient recommandées, les formes utilisant la macrolangue (avec ou sans extlang) sont toutefois permises et il existe des circonstances où ce choix est sans doute le plus opportun.

C'est le cas, notamment, pour des données ou des applications qui utilisent déjà la sous-étiquette « `ar` » (la macrolangue arabe) et qui préfèreront sans doute continuer d'utiliser cet indicatif plutôt que le nouvel indicatif plus précis « `arb` » (arabe standard moderne).

Notons enfin que le modèleextlang est bien adapté à la négociation de langues (HTTP) où la recherche d'un document en « `ar-ary` »(arabe-arabe marocain) fournira un document identifié comme simplement en arabe (`ar`), solution probablement acceptable si aucun document équivalent en arabe marocain (`ary`) n'est disponible. Notons que l'emploi de cette technique de recherche de document par troncatures successives à droite peut fournir un résultat incompréhensible en fonction de la macrolangue utilisée comme langue primaire (l'indicatif le plus à gauche).

6.2. Valeurs à ne pas utiliser

La plupart des valeurs « patrimoniales » incluses pour des raisons historiques et celles dites redondantes comprennent un champ qui indique qu'elles ne doivent

---

[15] Pour les Chinois, le cantonais est par exemple un dialecte. Bien qu'en réalité, d'un point de vue linguistique, il y a plus de différences entre le cantonais et le mandarin qu'entre l'italien et le français, même si l'intercompréhension à l'écrit entre ces deux langues chinoises est assez bonne grâce aux idéogrammes.

[16] On utilisait aussi zh-HK par exemple pour désigner la macrolangue chinoise de Hong Kong, où le cantonais est dominant. Cette étiquette est toujours valable, même si elle n'est pas conseillée.

plus être utilisées et qu'on leur préfère désormais un autre indicatif. L'entrée du registre ci-dessous est de ce type.

```
%%
Type: grandfathered
Tag:i-lux
Description: Luxembourgeois
Preferred-Value: lb
Deprecated: 1998-09-09
Comments: replaced by ISO code lb
```

Elle peut se lire ainsi : le codet « i-lux »pour le luxembourgeois est de type patrimonial (*grandfathered*), il faut éviter de l'utiliser, on lui préfère la valeur « lb » depuis le 9 septembre 1998.

6.3. Règle d'or

La règle d'or quand on crée des étiquettes linguistiques se résume à toujours utiliser la forme la plus courte possible. Il faut éviter de préciser les sous-étiquettes comme la région, l'écriture ou les autres sous-étiquettes facultatives, sauf si elles apportent une information utile. Ainsi faut-il utiliser « ja » pour le japonais et non « ja-JP » à moins que vous teniez absolument à dire qu'il s'agit du japonais parlé au Japon... Le tableau 9.5 énumère quelques étiquettes linguistiques.

Tableau 9.*Quelques étiquettes linguistiques*

Étiquette	Explications
fr	Le français.
ja	Le japonais.
i-enochian	Exemple d'étiquette patrimoniale : l'énochien[17].
zh-Hant	Le chinois en écriture chinoise traditionnelle.
zh-Hans	Le chinois écrit à l'aide de l'écriture chinoise simplifiée.
sr-Cyrl	Le serbe en cyrillique.
sr-Latn	Le serbe en latin.
sl-Latn-IT-nedis	Le dialecte slovène de Nadiza écrit en latin tel qu'utilisé en Italie (étiquette non

[17] L'énochien ou « langue des anges » est une langue occulte proposée par les alchimistes anglais John Dee et Edward Kelley au XVI^{e} siècle. Il possède son propre alphabet qu'Unicode n'a pas inclus (on le considère comme une simple transposition de l'alphabet anglais). Le codet i-enochian est un indicatif linguistique, pas un indicatif d'écriture.

Étiquette	Explications
	recommandée puisque le « Latn » est redondant pour le slovène, en d'autres termes l'entrée pour « sl » dans le registre de l'IANA a un « Suppress-Script » pour « Latn »).
sl-IT-nedis	Comme ci-dessus, mais sans l'écriture redondante et déconseillée. Cette étiquette est donc meilleure que celle ci-dessus.
de-CH-1901	L'allemand en Suisse écrit avec la variante orthographique de 1901.
zh-yue-Hant-HK	Le cantonais en écriture chinoise traditionnelle à Hong Kong (« zh-yue » est ici un code patrimonial).

6.4. Stabilité garantie

Les changementsles plus importants par rapport au précurseur des RFC 5646 et 4646, c'est-à-dire le RFC 3066, ont trait au fait que la syntaxe du RFC 5646 est désormais plus rigoureuse que celle du RFC 3066, que l'IANA tient à jour de manière permanente, stable et gratuite un seul registre unifié et qu'aucun codet ne disparaîtra ou ne sera affecté à une autre entité. L'Internet impose que « cs-CS », une fois valide, reste valide bien que la Tchécoslovaquie ait disparu, en tant qu'entité politique distincte, après l'enregistrement des codets en question. Le registre de l'IANA continue de suivre les normes de l'ISO mentionnées ci-dessus, mais il ne supprime jamais une sous-entrée et des règles claires ont été établies si des conflits entre ce registre et les listes de l'ISO venaient à surgir.

7. RFC 4647, trouver un document dans une langue

Dernier élément de cette série de normes et standards, le RFC 4647[18]. Ce document décrit une syntaxe appelée un *choix de langues* pour construirela listedes préférences linguistiques d'un utilisateur. En d'autres termes, comment préciser que l'on préfère avoir d'abord des documents en français de France, puis n'importe quelle variante française d'un document équivalent, puis les documents correspondants en arabe.

Ce RFC décrit également plusieurs mécanismes pour comparer ces listes de choix de langue et les associer à des étiquettes de langues. Deux types de mécanismes de correspondance sont définis : le filtrage et la consultation. Le filtrage produit un ensemble (éventuellement vide) d'étiquettes de langues tandis

[18] Cf. <http://abcdrfc.free.fr/rfc-vf/rfc4647.htm>

que la consultation produit une seule étiquette de langue. Les applications possibles comprennent la négociation de langue ou la sélection de contenu.

7.1. Filtrage

Dans le cas du filtrage, on cherche à trouver toutes les étiquettes linguistiques qui correspondent à un critère : trouver tous les documents en finnois (`fi`) par exemple.

L'utilisateur précise la valeur la plus générale qui constitue une réponse correcte, c'est ainsi que « `de` » (allemand) correspond à :

- « `de` » (allemand),
- « `de-CH` » (allemand utilisé en Suisse)
- « `de-CH-1996` » (allemand utilisé en Suisse, orthographe de 1996)

Il existe deux types de filtrages : un filtrage de base et un filtrage étendu. Le filtrage de base exige des préfixes communs :« `de-CH` » correspond donc à « `de-CH` » ou « `de-CH-1996` », mais pas à « `de-Latn-CH` ».

Le filtrageétendu avec l'aide du joker « * » permetd'accepter toutes les valeurs d'une sous-étiquette« `de-*-CH` » correspond à « `de-CH` », « `de-CH-1996` »,« `de-Latn-CH` »mais pas à « `de` » car il manque l'élément « `CH` ».

7.2. Consultation

Dans le cas de la consultation, on recherche le meilleur document parmi une liste de documents. L'utilisateur précise le choix de languele plus précis possible, car il ne veut qu'un document en retour.« `de-CH` » peut rendre « `de` » ou « `de-CH` », mais pas « `de-CH-1996` ». Si aucune étiquette de langue ne correspond à la demande, la valeur par défaut est retournée.

Quelques applications de ce mécanisme pour trouver la meilleure version linguistique :

- La sélection d'un gabarit contenant le texte d'une réponse électronique automatique.
- La sélection d'un élément contenant du texte à inclure dans une page Web particulière.
- La sélection d'une chaîne de texte à inclure dans un journal des erreurs.
- La sélection d'un fichier son à jouer en invite d'un système téléphonique.

- La recherche par repli des informations « locales » d'un programme informatique : les messages à afficher, les conventions de tri, le calendrier à présenter dans la langue de l'utilisateur.

Lors du mécanisme de consultation, le choix de langues est tronqué progressivement à partir de la fin jusqu'à ce qu'une étiquette de langue correspondante soit trouvée dans la base des documents. Par exemple, en commençant avec le choix « `zh-Hant-HK` » (chinois, écriture traditionnelle, Hong Kong), la consultation recherche progressivement du contenu comme indiqué ci-dessous :

```
Exemple d'un schéma de repli de consultation

Choix à réaliser :zh-Hant-HK
   1. zh-Hant-HK
   2. zh-Hant
   3. zh
   4. (défaut)
```

Le comportement de repli autorise de la flexibilité dans la recherche d'une correspondance. Sans repli, le contenu par défaut serait immédiatement retourné si un contenu correspondant exact n'était pas disponible. Grâce au repli, un résultat correspondant au mieux à la demande de l'utilisateur peut être fourni.

8. BCP 47

Pour éviter qu'une norme ne mentionne un RFC qui risque de devenir désuet, l'IETF recommande aujourd'hui aux auteurs de normes de faire référence à des *Best Current Practice* (« les Meilleures pratiques actuelles ») plutôt qu'à d'autres RFC. Les numéros de BCP ne changent pas, mais leur dernière version suit les derniers développements dans leur domaine.

Le BCP 47 regroupe les meilleures pratiques actuelles liées à l'indication de langue d'un document, d'une partie de document ou d'un objet. Le BCP 47 actuel est composé des RFC 5646 et 4647. La version précédente du BCP 47 comprenait les RFC 4646et 4647 qui remplaçaient déjà le RFC 3066, désormais désuet, lequel avait déjà remplacé le RFC 1766.

9. Tout cela est bien beau, mais est-ce utilisé ?

Ces étiquettes de langues sont utilisées par de très nombreux produits, standards et normes parmi lesquels on peut nommer : XML, HTML, RSS, MIME, SOAP, SMTP, LDAP, CSS, XSL, CCXML, Java, C#, ASP, perl, Apache, IE, Firefox…

Certains processus comme le mécanisme qui sert à préciser la locale POSIX (la langue et les conventions locales) d'un processus utilisent encore le RFC 1766 (le bisaïeul du RFC 5646) :

```
LANG=fr_FR
setenv LC_COLLATE=de-DE@phone
```

Ce genre de syntaxe est très fréquent sur les machines Unix. La première ligne indique la langue à utiliser notamment pour l'affichage des messages que le système affichera. La deuxième précise que le tri devra se faire en considérant les données comme de l'allemand (« de ») d'Allemagne (« DE ») en utilisant la convention de tri du bottin téléphonique (« @phone »). Dans les deux cas, on remarque que les codes de langue et de pays correspondent respectivement à l'ISO 639 et l'ISO 3166 (mais dans leur version de 1988).

9.1 Recherche de la bonne version linguistique d'un page Web

Le protocole HTTP permet à un navigateur internet d'indiquer quelle langue l'utilisateur préfère lire à l'aide de l'entête `Accept-language` qui accompagne chaque demande de document. La négociation de langue est très avantageuse pour naviguer sur les sites multilingues : le client obtient directement la version qui lui convient, sans perdre son temps à louvoyer parmi des pages qu'il comprend mal ou pas du tout, à la recherche d'un hyperlien vers la bonne langue. Le site <http://www.debian.org/> et celui du W3C (mais de manière très parcellaire[19]) l'utilisent.

L'`Accept-Language`de HTTP 1.1 permet de préciser un choix de langues codées sous la forme proposée par le RFC 1766. Toutefois, une révision de ce protocole (dénommé pour l'instant « HTTPbis ») permettra de définir celles-ci selon les RFC 5646 et 4647 (la dernière version du BCP 47 donc).

Comme nous l'avons vu, la recherche du bon document Internet à l'aide d'`Accept-Language` correspond à une consultation en termes du RFC 4647.

9.2 Recherche des ressources de programmes

Le même mécanisme de consultation et de repli pour trouver la ressource qui correspond le mieux au profil linguistique de l'utilisateur est utilisé par de nombreux environnements de développement comme Java ou C#.

Le programme écrit dans ces langages pourra ainsi accéder automatiquement aux valeurs qui correspondent le mieux à l'utilisateur. Un programme correctement conçu pourra ainsi aller chercher et afficher les messages, les libellés, les menus,

[19] La section des communiqués de presse du W3C utilise cette technique, voir par exemple <http://www.w3.org/2010/08/woff-pr.html>.

les avertissements dans la langue qui correspond le mieux à l'utilisateur par le même processus que nous avons décrit dans la section *7.2 Consultation* ci-dessus.

9.3 Indiquerla langue, l'écriture, le pays dans HTML et CSS

On recommande d'ajouter, dans les documents HTML, un attribut `lang` à la balise html. On préfère le mettre sur la balise `html` plutôt que le body, car il existe au moins un passage de texte qui ne serait pas balisé en choisissant `body` : le titre (`title`). L'exemple ci-dessous déclare un document écrit en français de Belgique :

```
<html lang="fr-BE">
```

En XML, et donc en XTHML, on utilisera l'attribut`xml:lang` pour préciser la langue :

```
<htmlxml:lang="fr-BE" xmlns
="http://www.w3.org/1999/xhtml">
```

Ces attributs peuvent s'utiliser sur la quasi-totalité des balises où il est raisonnable de préciser unelangue. La valeur de ces attributs correspond à un codet défini par le BCP 47. Parmi les balises où ces attributs sont fréquemment utilisés : `<p>` (le paragraphe) et `<span>` (le bout de texte). Par contre, on ne peut inclure d'attribut de langue sur `<br>` (passage à la ligne) ou `<hr>` (ligne horizontale, filet) qui n'ont pas de valeur linguistique.

Après avoir déclaré la langue globale du document, il est facile de mentionner qu'un passage est dans une autre langue par l'adjonction d'un attribut `lang`, `xml:lang` ou les deux sur la balise qui comprend le texte dans la langue étrangère :

```
<html lang="fr">
<body>
<p>Triple-patte s'exclama alors :<spanlang="la">Timeo Danaos
et dona ferentes</span>.Homère aurait dit :
<spanlang="grc">Φοβοῦ τοὺς Δαναοὺς καὶ δῶρα φέροντας</span>.</p>
```

Le langage de « stylage » CSS utilisé avec HTML connaît un type de sélecteur d'intérêt en matière linguistique : la pseudoclasse. La pseudoclasse est un sélecteur spécial qui définit une condition difficilement exprimable par les sélecteurs standard ou qui n'est pas liée à la structure du document. Une pseudoclasse nous intéresse particulièrement, il s'agit de `:lang(`*l*`)` où *l* représente une langue.

La pseudo classe`:lang(l)` sélectionne des éléments d'une langue « l » sans que ces éléments aient nécessairement un attribut `lang`, il suffit que la langue courante

soit celle mentionnée entre les parenthèses de `lang()`. Rappelons que les éléments HTML/XHTML héritent de l'attributlang de leurs éléments supérieurs (« leurs parents »). Le stylage lié à la langue permet de régler avec précision la présentation d'une page ou d'un passage selon la langue en question. On peut de la sorte choisir une police particulière pour des passages particuliers : des polices touarègues pour mieux rendre l'original en tamachek, une belle police coufique en arabe, etc. La recherche de correspondance est à nouveau une consultation au sens du RFC 4647, on cherche la meilleure correspondance avec comme stratégie de repli la troncature successive par la droite de l'attribut de langue (hérité au besoin) de l'élément que l'on doit « styler ».

C'est ainsi que dans le code ci-dessous :

```
<style>
:lang(ar) { color: blue; }
:lang(ar-MA) { color: red; }
</style>
<body>

<p>Il se tourna vers nous et dit, philosophe, <span lang="ar-
MA">inch Allah</span>.</p>

…
<p>Il se tourna vers nous et dit, philosophe, <span
lang="ar">inch Allah</span>.</p>

<p>Il se tourna vers nous et dit, philosophe, <span lang="ar-
ary">inch Allah</span>.</p>
```

Le premier « inch Allah » sera en rouge : la meilleure correspondance est celle pour `ar-MA`. Le deuxième sera en bleu, car « `ar` » est la meilleure valeur. Quant à la troisième invocation, elle s'affichera en bleu, car aucune pseudoclasse pour « `ar-ary` » n'a été définie, mais il en existe une pour « `ar` », qui correspond par troncature par la droite.

10. Conclusion

Indiquer la langue d'un texte ou d'un passage de ce texte s'avère utile pour de nombreux processus, qu'il s'agisse de la vérification orthographique, du repérage de texte comme Google, de trier ce texte ou encore d'en permettre la synthèse vocale.

La syntaxe des étiquettes linguistiques est définie par le BCP 47 qui est actuellement constitué des RFC 5646 et 4647. Chaque étiquette linguistique est composée d'une ou plusieurs « sous-étiquettes » séparées par des traits d'union. Si l'on excepte les étiquettes à usage privé et les étiquettes patrimoniales conservées

pour des raisons de compatibilité arrière, les sous-étiquettes doivent se présenter dans l'ordre suivant :

- une sous-étiquette qui représente la langue
- une sous-étiquette optionnelle qui représente une langue plus précise (« extlang ») quand la première sous-étiquette fait référence à une macrolangue.
- une sous-étiquette optionnelle qui représente l'écriture
- une sous-étiquette optionnelle pour la région
- une série optionnelle de sous-étiquettes représentant les variantes
- une série optionnelle de sous-étiquettes représentant des extensions
- une série optionnelle de sous-étiquettes à usage privé.

Exemples :

es	représente l'espagnol
fr-CA	le français au Canada
yue-Hant-HK	le cantonais écrit en chinois traditionnel à Hong-Kong

À ce stade, il n'existe pas d'indicatif particulier pour représenter l'amazighe marocain commun enseigné dans les écoles marocaines et promu par l'IRCAM. C'est une lacune qu'il faudra sans doute combler dans les années à venir. La question a déjà été abordée dans plusieurs cercles d'experts, une des questions qui risque de se poser lors de la codification d'un nouvel indicatif pour l'amazighe sera de voir s'il faut le considérer comme une macrolangue (regroupant les langues tamazight, rifaine, etc.) ou comme une nouvelle langue supplantant ou regroupant des dialectes, un peu comme l'allemand standard le fit.

11. **Remerciements**

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 cette communication a été présentée.Nous remercions également M. Ouguengay, son équipe, ainsi que Mme Aïcha Bouhjardirectrice du CAL et son équipe, pour leur appui essentiel pour mettre à jour les normes de l'ISO et de l'IETF afin d'inclure les informations nécessaires pour pouvoir étiqueter correctement les textes électroniques écrits en amazighe.

Bibliographie

Andries, P. (2008*)*. *Unicode 5.0 en pratique*,Dunod éditions, Paris.

Bortzmeyer, S. (2009), *RFC 5646: Tags for IdentifyingLanguages*, disponible à <http://www.bortzmeyer.org/5646.html>.

Phillips, A et Davis, M. (2009), *RFC 5646*, disponible à <http://www.rfc-editor.org/rfc/rfc5646.txt>.

POUR LA PROMOTION DE LA LANGUE AMAZIGH

YAHYA HLAL
Ex Professeur à l'EMI Rabat-Maroc

I. INTRODUCTION

La langue est le moyen de communication, le moyen d'expression et le moyen d'identification. Nul doute que c'est à travers la langue que la culture d'une société est la mieux exprimée. Le système de valeur en est intimement lié. La vraie valeur des concepts ne peut ne saisir que dans le contexte de cette langue. Un même mot peut dans une langue provoquer le respect et la vénération alors que dans une autre langue ce même mot peut provoquer le mépris.

La survie d'une langue ne peut être assurée que si elle se donne les moyens de pouvoir assumer les nouveaux objets et concepts, de pouvoir communiquer à grande échelle, et de pouvoir suivre l'évolution rapide des progrès scientifiques et technologique. Il se dessine désormais deux grandes classes de langues : celles qui ont ce pouvoir et les autres. On les appelle parfois les info-riches et les info-pauvres.

La langue Amazigh n'échappe pas à cette règle. Sa promotion culturelle et technologique ne peut se concevoir sans une prise de conscience de la problématique ; puis d'imaginer une stratégie à moyen et long terme pour assurer cette promotion. Se fixer des objectifs concrets et se donner les moyens politiques et matériels sont les conditions indispensables pour se lancer dans une telle entreprise aussi noble que nécessaire.

Il va de soi que l'adoption des technologies d'aujourd'hui, facilite beaucoup les choses en termes d'outils pour réaliser les projets à entreprendre en vue de la promotion de la langue Amazigh.

Ces outils devront toucher les différents aspects qui touchent à la langue. Je citerai à titre d'exemples les aspects suivants :

- Terminologie et aide au néologisme (Unification de la terminologie)
- Outils d'aide à la traduction depuis et vers l'Amazigh
- Outils D'analyse et de génération de la langue Amazigh
- Recherche d'informations véhiculées par le texte naturel Amazigh

- Outils pédagogiques pour l'enseignement de l'Amazigh à tous les niveaux
- Outils de communication sur Internet
- Correcteur orthographique et grammatical
- Dictée automatique
- Reconnaissance de l'écriture (OCR)
- Reconnaissance de la parole
- Etc.

La langue Amazigh n'a pas besoin de redécouvrir tous les outils. Certains outils, en matière d'ingénierie des langues, peuvent être étudiés pour les adapter à l'amazigh. Je ferais une proposition dans ce sens pour aider à mettre en place un véritable laboratoire pour la promotion technologique de la langue amazigh.

II. OUTILS DE BASE POUR LE DIALOGUE AMAZIGH

Le dialogue suppose un émetteur et un récepteur. Ce dernier reçoit une chaîne linéaire de caractères supposée contenir de l'information dans le cadre d'un contexte langagier (langue A, Amazigh en l'occurrence). La compréhension consiste à transformer cette structure linéaire en une structure non linéaire mettant en évidence les éléments de compréhension (un graphe où les nœuds constituent les concepts impliqués et les arcs constituent les relations qui les lient. Ce processus de transformation constitue ce qu'on appelle Analyse du discours. Le récepteur ayant compris; il s'apprête à répondre. A partir de ce qu'il veut transmettre, et qui se trouve dans un état interne à lui (structure non linéaire), il choisit le système langagier dans lequel il veut s'exprimer puis il transforme la structure interne en une chaîne de caractères linéaire devant être reçue par l'interlocuteur. Ce processus de transformation constitue ce qu'on appelle Génération du discours.

❖ **Processus d'analyse :**

On convient, en matière d'analyse automatique du discours, de distinguer quatre étapes d'analyse : Morphologie, syntaxe, sémantique et pragmatique.

❖ **Analyse morphologique :**

On demande à l'analyse morphologique de procéder à la décomposition du mot textuel en tous les éléments premiers qui entre dans sa composition : éléments en état de préfixation, mot lexical (qui lui-même peut être formé d'éléments premiers : racine ou base, préfixe et/ou suffixe, schème etc.), éléments en état de suffixation; puis d'associer aux éléments différentes informations d'ordre grammaticale ou sémantique hors contexte.

❖ **Analyse syntaxique :**

L'analyse syntaxique peut s'envisager sous deux aspects :

-Un aspect qui fait jouer le coté positionnel des mots les un par rapport aux autres en vue de lever les éventuelles ambiguïtés issues de l'étape morphologique**.**

-Un aspect plus élaboré qui consiste à construire les arbres syntaxiques des phrases. Cela permet une certaine interprétation du discours en se basant sur les fonctions syntaxiques mises en évidence (nœuds des arbres). On pourra poser des questions du type qui a fait quoi ? et comment ? Les phrases ambigües conduisent à mettre en évidence plusieurs arbres syntaxiques.

❖ **Analyse sémantique :**

A ce niveau on s'intéresse aux relations lexicales sémantiques qui lient les concepts de façon à résoudre les éventuelles ambiguïtés issues de l'étape syntaxique.

❖ **Analyse pragmatique :**

On a recours à cette ultime étape lorsque l'analyse sémantique n'est pas suffisante pour lever les ambiguïtés. On a recours alors à des considérations extralinguistiques (données historiques, politiques, ou connaissance du monde environnant d'une façon générale) . L'exemple suivant montre de quoi il s'agit : Dans la phrase « le meurtre de Jacques est horrible » on ne sait pas si Jacques est le meurtrier ou la victime. Maintenant si dans la phrase on remplace jacques par Archimède alors quiconque connaît l'histoire d'Archimède saura qu'il a été la victime.

❖ **Processus de génération:**

En génération il y a lieu de distinguer deux aspects différents : Génération lexicale et génération textuelle.

❖ **Génération lexicale :**

Cela consiste à générer à partir d'éléments premiers (racine et schème en arable à titre d'exemple) un mot lexical conforme aux règles morphologique de la langue en question. Cela est important dans le domaine de l'aide au néologisme où il y lieu de créer des mots nouveaux correspondants à des concepts nouveaux dont les noms existent dans d'autres langues (anglais, français etc.)

❖ **Génération textuelle :**

Cela consiste à générer les mots du discours, comme c'est le cas dans le processus de traduction automatique. Là il s'agit à partir d'un mot puisé d'un dictionnaire, suite à un transfert lexical (School → Ecole par exemple), de générer la forme du mot compte tenu des attributs contextuels (nombre, détermination, etc..).

Ces outils d'analyse et de génération constituent les primitives de bases pour tout dialogue ou applications qui requièrent ce genre de traitement.

II. QUELQUES APPLICATIONS TYPES

Terminologie et aide au néologisme

La gestion de la terminologie est essentielle dans la promotion de la langue. La gestion s'entend ici au sens large du terme. Elle doit permettre entre autre la disposition d'outils pour accéder au patrimoine terminologique suivant les critères les plus variés pour répondre aux besoins de tous les utilisateurs dans tous les domaines scientifiques, économiques, culturels, etc.. Ainsi que la confection et la diffusion sous toutes les formes adéquates de lexiques dont a besoin la société dans ses différents secteurs éducatifs, administratifs et autres.

Les outils d'aide au néologisme sont importants pour permettre aux spécialistes du néologisme de produire la terminologie adéquate dans les différents domaines d'évolution de la langue, tout en respectant les lois morphologiques de la langue et ne pas céder à la facilité qui consiste à donner une sonorisation amazighe aux termes étrangers.

Outils d'aide à la traduction pour l'Amazighe

Ces outils peuvent s'envisager en relation avec les outils de terminologie multilingue.

Ces outils peuvent être dans un premier temps très élémentaires. Une aide de traduction proche du mot à mot peut rendre service, particulièrement dans les domaines scientifiques et technique. Le résultat sera repris par le traducteur pour apporter éventuellement les améliorations syntaxiques et de style pour produire le texte final.

Traitement du discours amazighe

Il s'agit de pouvoir procéder à l'extraction de l'information véhiculée par le discours. Cette information pourra être de différents niveaux, selon le besoin de l'applicatif, morphologique, syntaxique ou sémantique.

Il entre dans cette classe de traitements les moteurs de recherche de documents amazighes, leur traitement dans le cadre d'indexation pour alimenter des bases textuelles, traitement statistique selon des critères morphologiques, interprétation de texte dans le cadre de système expert basés sur une approche linguistique.
Les fonctionnalités liées au moteur de recherche devraient s'envisager dans le cadre multilingue (arabe, français et anglais entre autre). Voici, à titre d'exemple quelques fonctionnalités qu'on devrait envisager :

+ Recherche multicritère (Titre, Auteur, Date, etc.)

+ Recherche type ratissage : un mot sera recherché systématiquement partout où il peut se trouver (tout champs confondu)

+ Emploi des connecteurs logiques

+ Lutte contre le silence:

+ Classe de caractères (o = ô -> impot = impôt)

+ Famille de mots en français (imposable=imposition)

+ Recherche par concept (impôt=redevance=IGR= ضرائب)

+ Recherche par Racine en arabe ou équivalent amazighe (si un principe équivalent existe)

+ Recherche par Hyper lien dynamique

+ Recherche par requête préétablie

+ Recherche par mots clés contrôlés ou non

IV. LABORATOIRE POUR Le TRAITEMENT AUTOMATIQUE DE L'AMAZIGHE

Il est important pour la promotion technologique le l'Amazighe de penser à une structure spécialisée apportant la logistique nécessaire aux décideurs pour aider à réaliser les objectifs à court, moyen et long terme.

Dans le cas du Maroc, il sera judicieux d'utiliser les compétences existantes qui ont été utilisées pour la promotion de la langue arabe. Durant plus de vingt ans, au niveau de l'EMI, le LIT2A a développé une activité de recherche qui s'est appuyée sur la formation des ingénieurs et l'enseignement de 3ème cycle, et qui a donné lieu à la réalisation de nombreux outils dans le cadre de projets de fin d'étude et des thèses de 3ème cycle.

L'idée est d'opérer un transfert de technologie rapide dans ce domaine qui pourra s'articuler sur deux axes essentiellement :

- Une formation approfondie sur les différents aspects de l'ingénierie des langues. Cela concernera aussi bien les outils linguistiques de base que les applications connexes dont on a évoqué quelques aspects plus hauts.
- Un encadrement dans le cadre de l'activité de ce laboratoire pour s'approprier ces outils linguistiques et les adapter partiellement ou profondément pour confectionner une réelle ingénierie de la langue Amazighe.

V. CONCLUSION

Il est essentiel pour une langue comme l'amazighe qui s'est donnée comme objectif une promotion réelle dans tous les domaines de prendre conscience de l'enjeu de cette entreprise Cela devra se traduire par des décisions adéquates en vue de réaliser les objectifs fixés. Parmi ces décisions, s'employer à utiliser avec énergie les moyens technologiques qui s'offrent aujourd'hui. Ces moyens se déploient aussi bien sur le plan des outils de traitement et de communication que sur celui de la mise en place de compétences capables de mener une activité de recherche et de production en matière de promotion technologique de la langue Amazighe.

WEB 2.0 & WEB 3.0 popularisation de la création et de la promotion culturelle et scientifique : cas des Wikis Opportunité pour Tamazight

Hammou FADILI [(1)(2)]
(1) FMSH, Maison des Sciences de l'Homme, Paris, France
(2) CNAM, Conservatoire Nationale des Arts et Métiers, Paris, France
hammou.fadili@(msh-paris.fr,cnam.fr)

Résumé : L'objectif principal de cet article est de présenter quelques éléments relatifs aux technologies web2.0 et web3.0 pouvant aider à la popularisation de la création et de la promotion culturelle et scientifique. Dans la première partie, on va donner quelques éléments caractérisant la technologie web2.0 ainsi qu'une technologie essentielle la constituant, à savoir la technologie « wiki ». La deuxième partie sera consacrée à la description de quelques éléments sur l'évolution du web actuel vers la version web3.0. On profitera de cette partie pour donner quelques éléments caractérisant cette version du web en se focalisant sur la future version de la technologie wiki appelée wiki sémantique. Nous essayerons de montrer à travers ces différents éléments, l'apport de ces technologies pour la production de contenus électroniques ainsi que leur promotion.
Mots-clés. Web 2.0, Web 3.0, Wikis, Wiki sémantique, agents.

Summary: The goal of this article is to show some elements of the web2.0 and web3.0 technologies that could help to popularize creation and promotion of scientific and cultural contents. We will describe some characteristics of web2.0 by focusing on wiki technology that constitutes an essential element of this version (current) of the web. In the second part of the article, we will describe some elements that will constitute the future version of the web called web3.0 by showing the expected evolution for the wiki technology to the semantic one. The third part will be devoted to show the possibilities given by these technologies to improve creation and promotion of scientific and cultural contents.

Key words: Web2.0, web3.0, wikis, semantic wiki, agents.

1. INTRODUCTION

Les développements récents des technologies de la communication et de l'information, notamment ceux concernant la mise en réseau des ordinateurs, des logiciels spécialisés et des applications, intéressent plus particulièrement la recherche et la culture, par la gestion, la veille et la diffusion de l'information. Les exemples d'applications les plus connus renvoient aux domaines de l'édition en ligne, de la diffusion en ligne, du conseil en ligne ou encore du "travail collectif (du groupeware, en anglais) à distance.

Ces technologies, issues des premières versions de l'Internet sont arrivées à maturité et se spécialisent autour des standards du Web 2.0 - version actuelle du WEB. Elles sont centrées sur l'utilisateur et permettent de faciliter l'utilisation de l'Internet en se faisant aider par les agents logiciels connectés dans le réseau mondial, comme les programmes interactifs d'aide à la création et à la diffusion de contenus. Cette version du Web est, elle-même, amenée à évoluer, on parle déjà du WEB 3.0 pour la future version qui aura pour rôle de concilier l'internet et l'intelligence artificielle. Des éléments sur cette partie seront détaillés un peu plus tard dans cet article.

L'objectif principal de cet article est de présenter quelques éléments relatifs aux technologies du web 2.0 et du web 3.0 pouvant aider à la popularisation de la création et de la promotion culturelle et scientifique. Nous définirons quelques éléments caractérisant le web 2.0 ainsi qu'une technologie essentielle le constituant, à savoir la technologie « wiki ». Puis on présentera quelques éléments sur l'évolution du web vers sa version 3.0. En profitant de cette section, nous donnerons quelques éléments caractérisant la future version du web en se focalisant parallèlement sur la future version de la technologie « wiki » appelée « wiki sémantique ». Nous essayerons de montrer à travers ces différents éléments, l'apport de ces technologies pour la production et la promotion de contenus électroniques culturels et scientifiques. Ceci en mettant en avant la technologie WIKI permettant de mettre en place des outils simples et populaires intéressants pour les utilisateurs du grand public pouvant les inciter à constituer des groupes et des communautés virtuelles favorisant la création et la diffusion de contenus dans leurs domaines d'intérêts. Tous ces éléments ainsi que leurs contextes d'utilisation vont être décrits comme suivant :

Nous commençons par donner quelques éléments de définition du web 2.0, version actuelle du web considéré comme dynamique, coopératif et interactif. On s'intéressera ensuite à la notion du wiki dans le contexte du web 2.0 en définissant les caractéristiques essentielles de cette technologie. Nous détaillerons ensuite les fonctionnalités qui peuvent être offertes par les wikis et les domaines d'utilisation où les wikis ont fait leurs preuves en se servant d'un exemple montrant l'efficacité

d'une telle technologie pour l'apprentissage des langues. Nous décrirons ensuite quelques éléments pouvant définir la prochaine version du web appelée web3.0 permettant de faire cohabiter le web et l'intelligence artificielle en faisant un parallèle avec l'évolution des wikis vers les wikis sémantiques. Enfin, nous montrerons que tous ces éléments constituent une chance et une opportunité pour le développement d'une culture (surtout si celles manquant de moyens) par la popularisation de la création et de la promotion de ses contenus en utilisant Internet d'une manière simple et coopérative.

2. Notions du web 2.0

Le web 2.0 peut être défini comme un ensemble d'outils, normes, protocoles,… permettant, d'une part les interactions entre les utilisateurs ou groupes d'utilisateurs via des applications particulières, capables de tisser des liens entre individus et communautés par la création des « réseaux sociaux » autour de centres d'intérêts communs, mais aussi entres les applications par échanges de données basées sur des protocoles normalisés : Web services, les flux RSS, OAI...

Le Web2.0 est considéré comme dynamique, interactif, et coopératif. Il permet :

- la co-création : la production de contenus par les utilisateurs via des réseaux sociaux en utilisant des applications et outils dédiés : Wikis, Blogs, sites dynamiques…,
- la décentralisation : la création peut se faire par des personnes éloignées, physiquement distribuées,
- la compilation de contenus : génération de contenus à partir d'autres contenus via des applications respectant des normes et des protocoles permettant la recherche et les échanges des données,
- le support des applications internet riches et interactives de type CMS par exemple,
- la création des systèmes émergents ou applications distribuées dont le comportement global est la résultante des comportements élémentaires des services constituants,
- …

Ce qui caractérise le WEB 2.0 des versions antérieures, c'est l'interactivité et la collaboration permettant la production coopérative de contenus et faisant d'un utilisateur d'Internet un lecteur et un créateur au même temps. Les technologies utilisées dans le WEB 2.0 évoluent en supportant de nouvelles normes qui constitueront les bases de la future génération du WEB. Elles permettent donc la mise en place d'un certain nombre d'outils et de technologies (dont la technologie wiki qui est un élément essentiel et représentatif de cette technologie) profitables à la prochaine version du WEB.

Dans ce qui suit, nous allons montrer l'intérêt de cette technologie ainsi que sa philosophie à travers la technologie Wiki, décrite dans les paragraphes suivants.

2.1. Qu'est ce qu'un WIKI

Un wiki est un CMS (système de gestion de contenu) de sites Web qui rend les pages Web librement et également modifiables par tous les visiteurs autorisés [Wikipedia, 2009].
Un Wiki est en général une application ouverte, universelle et multilingue en ligne où les lecteurs peuvent modifier les pages qu'ils sont entrain de consulter. Ce qui veut dire que les parties espace de travail « Backend » et espace consultation « Frontend » sont confondues. L'interface de gestion est simplifiée, réduisant le nombre de champs des formulaires, même lorsqu'il s'agit de la gestion des contenus structurés selon des schémas complexes. On utilise pour cela les langages WIKIML (Wiki Markup Language) ou XML (Extented Markup Language) pour structurer les données d'une manière simple via des applications simples comme les WYSIWYG (What You See Is What You Get). Il suffit d'insérer le contenu souhaité dans le seul champ du formulaire, puis procéder à la description de sa présentation ou à sa structuration en lui associant des annotations par l'intermédiaire de l'interface WIKI qui est facile à utiliser et ergonomique.
La famille des WIKI se situe entre la famille des CMS (Content Management System ou Système de Gestion de Contenu) et celle des BLOGS, qu'on peut définir respectivement comme suivant :

- un CMS est une application WEB de gestion de contenus basée en général sur des templates ou modèles. Il est basé sur des architecture riches et parfois dans certains cas d'utilisation lourdes distinguant bien la partie consultation de contenus dédiée aux visiteurs « Frontend » de la partie gestion / administration dédiée aux auteurs administrateurs des sites « Backend »,
- un Blog est Wiki particulier appartenant à son « propriétaire », contrairement à un WIKI qui est un espace ouvert et géré par plusieurs personnes, qu'on pourrait désigner comme un blog de groupe.

2.1.1. Les fonctionnalités essentielles d'un Wiki

La technologie Wiki est basée sur des solutions : légères, moins couteuses, accessibles à tous. Son utilisation ne nécessite pas de formation particulière. Nous proposons ci-après un résumé des fonctionnalités que pourrait offrir un WIKI, sous forme d'une liste non exhaustive suivante. Il permet :

- à des personnes autorisées ou pas, d'éditer et de publier facilement et rapidement des contenus en ligne car cette technologie est facile à utiliser ne nécessitant pas de formation pour son utilisation,
- d'aider à la création de groupes d'utilisateurs autour de sujets ou de thématiques particuliers pouvant former des « sociétés virtuelles » sur Internet respectant des règles de groupes suivant le modèle des sociétés réelles,
- d'inciter à la création en encourageant la production de nouveaux contenus du fait de la popularité des outils de types Wikis : Wiki, blogs, forums…,
- d'être un levier pour la création de groupes par l'augmentation de la production de la « littérature grise » et favoriser l'émergence d'une «intelligence collective»,
- de faire évoluer les contenus via des processus de travail collaboratif par mutualisation de compétences,
- d'entretenir et sauvegarder la mémoire d'un groupe, d'un projet, d'une institution. Le WIKI d'une institution contient d'une manière naturelle des contenus reflétant la vie et l'évolution de celle-ci,
- une gestion complète des versions et des historiques des contenus constituant un des aspects positifs de la sécurité des contenus,
- une gestion des notifications permettant aux utilisateurs intéressés par des thèmes particuliers d'être alertés à chaque fois que des créations, modifications ou suppressions de contenus liés à leurs thématiques favorites ont lieues,
- une utilisation sûre, car c'est une technologie qui est en production depuis plusieurs années, et qui a déjà fait ses preuves en termes de gestion & consultation de contenus collectifs,
- assurer une meilleure diffusion des contenus, du fait qu'elle est très populaire et largement consulté,
- …

2.1.2. Domaines potentiels d'utilisation

Du fait que les Wikis sont facilement modifiables, ils ont été adaptés et utilisés dans plusieurs domaines. Ils sont très utilisés comme petits et moyens sites Web où on n'a pas besoin de gros CMS ou d'applications sophistiquées pour leurs mises en place. Ils sont généralement utilisés comme intranets de laboratoires et d'équipes de recherche, de projets,…, sous forme, de bases de connaissances et de documentation partagées et d'espaces de rédaction collaboratifs. On les a adaptés également pour mettre en place des espaces d'échanges et de diffusion de supports pédagogiques et pour gérer les communications entre les enseignements et les apprenants. Un autre domaine important d'utilisation des WIKIs est l'édition/publication électronique où il suffit de quelques développements légers pour mettre en place des processus d'édition et de publication électroniques. Les

différents circuits des traitements et d'approbations des données ainsi que les différentes vues de publication peuvent y êtres aussi implémentés. Des utilisations en tant que bibliothèques numériques ont été également mise en place dans beaucoup d'institutions pour la diffusion d'informations structurées comme les informations bibliographiques par exemple, où il suffit de les associer à une gestion avancée des métadonnées implémentée nativement par les Wikis sémantique. Un autre aspect très populaire d'utilisation des Wikis est leur adaptation en tant que BLOGS ou sites d'information personnels. C'est cette utilisation qui a rendu leur utilisation populaire au niveau du grand public et a fait de cette technologie un moteur pour l'Internet moderne. Aussi, l'utilisation des Wikis en tant qu'espaces privés est presque systématique dans toutes les entreprises et institutions quelles que soient leurs tailles.
Ci-après, quelques exemples de Wikis publics :

- iTalki : le réseau social d'apprentissage linguistique qui recense plus de 108 langues dont Tamazight : http://www.italki.com/learn-Tamazight.htm
- Wikipédia et ses variantes, Wikilivres, Wiktionnaires, Wikisources (articles),…qui est le WIKI le plus connu et le plus populaire de l'Internet. C'est le deuxième site consulté après Google.

2.1.3. Un exemple

L'exemple présenté dans ce paragraphe concerne un WIKI mondialement connu, utilisé pour l'apprentissage des langues. Ce choix est motivé par les thématiques traitées dans ce colloque relatives à la langue et à la culture Amazighe. Il s'agit d'Italki, espace où l'on peut trouver tout ce dont on a besoin pour apprendre une langue. C'est un réseau social et une ressource en ligne pour l'apprentissage des langues étrangères. Selon un de ses concepteurs, avec italki, aucune langue ne sera perdue dans l'histoire. Il est conçu autour des services suivants :

- « partenaires Linguistiques », c'est un espace où on peut trouver des partenaires linguistiques pour une langue donnée,
- « réponses », c'est un espace pour poser des questions sur une langue,
- « connaissances », espace pour stocker et consulter les pages web dédiées à l'apprentissage d'une langue,
- « dossiers », espace pour stocker et consulter les manuels, documentation partagés,
- « ressources », espace pour stocker et consulter les ressources comme les sites WEB,
- « groupes », ensemble des groupes crées autour de l'apprentissage d'une langue,
- « index linguistique », c'est l'index de toutes les langues prises en charge dans le WIKI.

Tamazight est prise en charge dans ce WIKI. On y a constitué plusieurs groupes, de forums,… pour la création et la collecte d'information et de documents sur l'apprentissage de Tamazight.

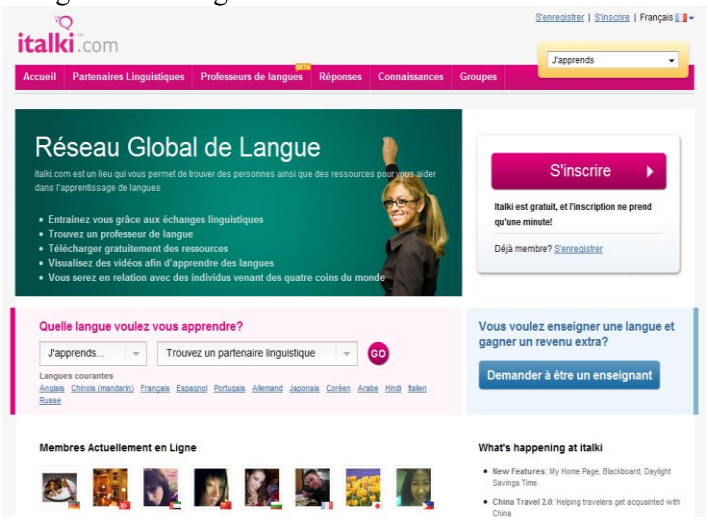


Figure1. Le WIKI iTalki

3. Le WEB sémantique et la notion d'agents

Le WEB sémantique est une notion très importante de l'Internet moderne, considéré comme intelligent où on ne se contente pas de stocker et diffuser des données, mais on s'intéresse à leur compréhension en effectuant des raisonnements sur leurs sens par des machines et agents logiciels. On peut le définir comme un ensemble de technologies permettant aux machines d'effectuer des traitements (dans certains cas difficiles pour l'homme, par exemple : indexation, compréhension et recherche sémantiques) sur des données en s'appuyant sur les concepts comme, l'expression du sens, La représentation des connaissances, Les ontologies, Les agents, L'évolution de la connaissance… (Tim Berners-Lee, co-inventeur avec Robert Cailliau du World Wide Web).
Un agent est une entité du système modélisé, situé dans un environnement, doté de capacités d'adaptation et d'autonomie lui permettant d'atteindre ses objectifs.

Il existe plusieurs types d'agents :

- Réactifs qui ne font que réagir aux stimuli qu'ils perçoivent d'une manière mécanique. Le comportement du système émerge des réactions simples des agents,
- cognitifs disposant de capacités de raisonnement sur sa représentation du monde où ils évoluent. L'une des architectures les plus connues pour ce modèle est l'architecture BDI (Belif, Desir, Intentions) où les agents sont constitués principalement des caractéristiques suivantes :

- Les croyances qui représentent la connaissance sur l'état de l'agent et de l'environnement où il évolue, c'est à dire ce que l'agent connaît sur lui même et sur le monde où il évolue,
- Les buts qui représentent la connaissance sur les motivations et les objectifs de l'agent,
- Les intentions qui représentent les informations sur les choix des plans que l'agent peut faire pour satisfaire des exécutions possibles.

Un système multi-agents est un ensemble d'agents partenaires partageant des ressources et compétences complémentaires, similaires ou dissemblables et coopérant afin d'atteindre des objectifs partagés.
Un Système Multi-Agents Adaptatif (AMAS) est un système autonome qui doit faire face à des situations imprévues qui ne peuvent pas être résolues de manière algorithmique. Le comportement global (comportement émergent) d'un AMAS est le résultat de la coopération définissant l'organisation entre agents, ce qui revient à dire que, pour changer la fonction globale d'un AMAS, il suffit de changer l'organisation des agents le composant, dits agents AMAS. Les agents dans ce contexte, doivent faire face en permanence aux changements liés à l'environnement et aux situations non prévues. La communication au sein d'un AMAS entre agents se fait par l'intermédiaire d'un protocole d'interactions qui constitue un ensemble de règles de conduite que les agents doivent respecter entre eux afin de structurer leurs échanges. Et la mission de l'Internet futur, est de concilier ces deux notions dans le contexte du WEB, dans la perspective d'un WEB intelligent appelé WEB 3.0.

4. WEB 3.0

Le WEB 3.0 est le nom prévu pour la prochaine version de l'internet. C'est un concept émergeant qui s'articule autour du WEB sémantique et de l'intelligence artificielle qui pourrait être de type AMAS.

Il est considéré comme l'internet 3ème génération, version évoluée de la version statique du WEB vers le WEB intelligent en passant par le WEB dynamique classique. Il aura pour mission de faire cohabiter le web est l'intelligence artificielle distribuée (IAD) en utilisant un ensemble, d'outils, protocoles, normes, standards… permettant à des machines et à des « agents web intelligents » d'effectuer, des raisonnements, des traitements…en ligne d'une manière coopérative et automatique sur des contenus WEB.

Ce qui aura l'avantage d'alléger l'utilisation de l'internet futur qui, sans l'intégration de tels outils, peut conduire à des blocages des traitements et à la

saturation du réseau mondiale ; car la quantité des données que génère Internet par les utilisateurs (humains) et par des agents web (machines) est très importante et peut atteindre des niveaux qui ne pourraient pas êtres gérés par les outils et technologies actuelles.

C'est grâce aux systèmes d'inférences basés sur des règles, qu'on pourrait remédier à ces problèmes de saturation des traitements, pour la prise en charge de la totalité des données présentent sur Internet, constituant ainsi la base de connaissances sur laquelle s'effectuent les raisonnements. Ceci pourrait être rendu possible en structurant et en annotant sémantiquement les données en utilisant les technologies telles que le web sémantique et la représentation du sens (sémantique), qui intègrent des technologies avancées telles que : xml, métadonnées, RDF, TAL, extraction, ontologies, compréhension automatiques de contenus… et permettre ainsi à des outils intelligents de prendre en charge beaucoup de tâches automatisables et difficiles à effectuer par un utilisateur humain.

Le WEB 3.0 est une évolution du WEB 2.0 et des propriétés qui le caractérisent. Ce qui veut dire que les aspects liés à l'interactivité et à la coopération feront toujours partie du WEB futur. C'est-à-dire aussi que dire que les utilisateurs ne seraient pas remplacés complètement par les machines, mais, auront, en plus des rôles joués dans la version actuelle du WEB, des rôles supplémentaires liés en particulier : au contrôle, à la supervision, à la configuration…des applications et des agents. Ils seront en retrait dans certains cas pour laisser les machines effectuer les premiers traitements avant leur validation. Ainsi, les machines joueront un rôle important comme outils d'aide pour effectuer des traitements ou des prétraitements sur Internet quand l'intervention humaine n'est pas nécessaire.

4.1. Wiki sémantique

La technologie Wiki sémantique sera au cœur du WEB 3.0 comme l'a été la technologie Wiki par rapport au WEB 2.0 et hérite donc de ses propriétés i.e. qu'elle peut combiner les aspects WEB et l'intelligence artificielle dans un contexte WIKI.

Un Wiki sémantique est un Wiki particulier doté d'outils et de méthodes lui permettant de formaliser et représenter le sens. Il peut être utilisé dans des applications gérant des schémas complexes des données structurées. Les technologies utilisées dans un Wiki sémantique se basent sur les Systèmes de Gestion des Connaissances, TAL, Ontologies, Inférences, Analyse & extraction des données, Annotation et Recherche Sémantiques…

Les utilisateurs peuvent coopérer pour créer non seulement des contenus mais aussi pour leurs associer de la sémantique. Cette tâche peut être considérée comme une tâche difficile qui nécessite dans certains cas beaucoup d'échanges, de coopérations et de discussions pour déduire et effectuer les meilleures annotations.

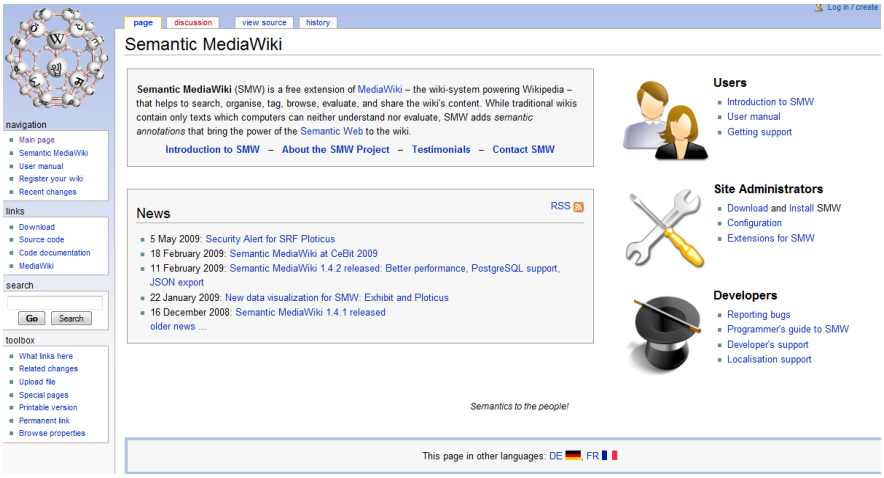


Figure2. MediaWiki sémantique

5. Opportunité pour Tamazight

On vient de décrire dans les paragraphes précédents les rôles que pourraient jouer les outils et technologies constituants les WEB 2.0 et WEB 3.0 pour la création et la diffusion des contenus.

Dans le premier cas on peut aider à la création et à la diffusion des contenus par intéressement et à la constitution de groupes et de communautés d'intérêts communs.

Dans le second, on permet à des machines et à des agents intelligents de générer et de produire de grandes quantités de contenus d'une façon automatique, enrichissante pour une culture données.

Tamazight doit donc en profiter pour augmenter la quantité de ses contenus et sa présence sur Internet qui feront sa richesse. Pour se faire il faut que les institutions travaillant sur la culture Amazighe mettent à disposition du grand public, des applications dynamiques et interactives de types wikis, blogs, forums…traitant des sujets et des thèmes attractifs et intéressants pour inciter les utilisateurs à participer à titre individuel ou au sein de groupes à la création de contenus. Les utilisateurs Imazighen doivent eux aussi de leur côté avoir le

réflexe d'aller s'exprimer sur Internet sur tout ce qui est en rapport avec leur culture et leur langue : écrire, bloquer, commenter, discuter,...de leurs sujets favoris. D'une manière général, Tamazight doit intégrer l'ère de la « cyberculture » pour profiter de la vague actuelle du Web 2.0, Web dynamique et interactif, dont les Wikis font partie, où le lecteur est auteur et où le visiteur est acteur créateur de contenus, afin de se préparer pour la prochaine révolution de l'internet qui se fera autour du WEB 3.0, où les machines et les « agents intelligents » pourraient prendre la relève dans certains cas d'utilisation pour des traitements spécifiques et aider ainsi à son développement par la création, génération et compilation automatique de contenus. C'est une chance à ne par manquer pour son développement à condition que ces nouvelles technologies soit largement utilisées et aussi vulgarisées au niveau du grand public.

6. Conclusion

Dans cet article, nous avons montré l'intérêt que pourrait représenter les nouvelles technologies en l'occurrence le WEB 2.0 et prochainement le WEB 3.0 pour la création et la promotion des contenus. Les contenus culturels et scientifiques constituent les éléments essentiels et la richesse d'une culture et d'une civilisation, donc, l'utilisation de telles technologies pourrait être moteur pour son développement. Ceci est d'autant plus vrai, si on ne dispose pas de moyens suffisants pour son épanouissement. Pour se faire, les responsables et les institutions ont le devoir de soutenir les efforts visant à développer et à utiliser les nouvelles technologies pour la création et la préservation du patrimoine scientifique et culturel. Ils doivent également rendre accessible tous ces éléments en mettant en place des systèmes permettent de donner l'accès aux informations existantes et aux outils permettant la création d'une manière facile et intuitive de nouveaux contenus.

7. Bibliographie

[1] BERNERS-LEE Tim, HENDLER James and LASILLA Ora, The Semantic Web, Scientific American, May 2001.
Wikipédia : http://fr.wikipedia.org/wiki/Accueil

[2] Agostinelli, S. (2006). Quelles formes de partage
les wiki autorisent-ils ? Dans A. Piolat
(Éd.), Lire, Ecrire, Communiquer et Apprendre sur internet (pp. 401-418). Marseille : Solal.

[3] Karayan, R. (2003). La révolution wiki est en vue, Le journal du net. Retrouvé Août 10, 2003, de www.journaldunet.com /0308/030811Wiki.shtml.

[4] [Berners-Lee et al., 2001] Tim Berners-Lee, James A. Hendler et Ora Lassila (2001). The SemanticWeb. Scientific American, 284(5):34–43.

[5] [O'Reilly, 2005] Tim O'Reilly (2005). O'Reilly Network : What Is Web 2.0 : Design Patterns and Business Models for the Next Generation of Software. http://www.oreillynet. com/lpt/a/6228.

[6] [Passant et Laublet, 2008b] Alexandre Passant et Philippe Laublet (2008b). Ontologies et Web 2.0. In IC2008, 19èmes Journées Francophones d'Ingénierie des Connaissances.

[7] Bernatchez J. Le blogue comme instrument d'apprentissage: bilan d'une expérience réalisée à l'École nationale d'administration publique. ENAP - Université du Québec, 2006. [en ligne] http://www.uquebec.ca/capres/fichiers/art_ENAP-juin06.shtml (consulté le 15/01/2007)

[8] VÖLKEL M., KRÖTZSCH M., VRANDECIC D., HALLER H. & STUDER R. (2006). Semantic Wikipedia. Proceedings of the 15th international conference on World Wide Web, p. 585–594.

[9] BUFFA M., GANDON F. L., ERETEO G., SANDER P. & FARON C. (2008). Sweetwiki : A semantic wiki. Jornal of Web Semantic, 6(1), 84–97.

L'enseignement à distance de l'amazighe dans un contexte de diaspora : pluralité de situations et convergences opératoires

HOCINE SADI
Université d'Evry Val d'Essonne
Hocine.sadi@free.fr

Résumé : Des circonstances particulières, liées à l'enseignement du berbère en France nous ont confronté aux TICE. Les méthodes de travail, d'élaboration des ressources pédagogiques, la mise en ligne de ces ressources et le dispositif d'échange avec les apprenants sont autant de questions qui intéressent toute entreprise d'EAD. En outre, comment parvenir à des ressources interopérables et réutilisables dans un environnement éclaté, instable et qui de surcroît souffre d'une absence de structure normative ?
Mots clés : Enseignement à distance, diaspora, amazighe, TICE, normes.

Abstract: Particular circumstances, related to the teaching of Berber in France confronted us with ICT in education. Methods of work, development of the teaching resources, online availability of these resources and the kind of sharing device with learners/students/pupils are as many concerns of interest related to any business of distance learning (e-learning). Furthermore, how to manage re-usable and inter-operable resources in a fragmented and unstable environment, which in addition suffers from an absence of normative structure?
Keywords: Distance learning, e-learning, Diaspora, amazighe, ICT in education, standards.

1. Introduction : gage de modernité, la caution scientifique

Pour situer ma contribution à ces troisièmes ateliers internationaux organisés par le CEISIC de l'IRCAM, je commencerai par indiquer le lieu d'où j'interviens. L'on verra incidemment, sur un cas concret mais non singulier, comment les nouvelles technologies de l'information de la communication et de l'éducation (TICE) ont pu croiser un itinéraire d'un professeur de mathématiques militant de la culture amazighe, susciter chez lui un investissement scientifique.

Tout au long d'une carrière professionnelle clairement inscrite dans les mathématiques, j'ai manifesté un intérêt constant pour les études berbères. Dès mon entrée comme étudiant à l'Université d'Alger à la fin des années soixante, j'assistais aux cours de berbère que Mouloud Mammeri y dispensait. Plus tard, inscrit en troisième cycle de mathématiques à l'Université Paris 6, je suivais les

conférences données par Lionel Galand au début des années soixante-dix à l'Ecole pratique des hautes études à la Sorbonne. Tout en militant à l'Académie berbère- « Agraw Imazighène » à la même période, je participai à la création du Groupe d'études berbères de Vincennes en 1972 aux côtés de Mbarek Redjala qui le dirigera.

Alors que j'occupais les fonctions de maître assistant en mathématiques à l'Université de Tizi-Ouzou (1979-1983), nous créâmes en compagnie de collègues universitaires et de quelques étudiants la revue Tafsut en 1981. Avec Ramdane Achab et Mohand Laïhem, nous nous mîmes à la confection d'un lexique de terminologie mathématique de quelques 2000 termes. Le fascicule qui avait bénéficié du soutien de Mouloud Mammeri paraît en 1984 dans la série scientifique et pédagogique de Tafsut. En 1990, je publiai un livre de mathématiques récréatives en amazighe (kabyle) « Tusnakt s wurar » qui reprenait la terminologie mise au point dans le lexique édité par Tafsut.

Il peut être utile pour la compréhension des enjeux de l'époque de souligner le caractère volontariste de cette démarche qui consistait à inscrire l'amazighe dans un champ, celui des mathématiques, où il était totalement absent. Ce volontarisme ne visait pas, surtout à l'époque, à satisfaire un quelconque besoin pédagogique, notre motivation répondait à un autre souci, celui de voir l'amazighe s'approprier une discipline qui incarne par excellence l'abstraction, les sciences et donc la modernité. Notre travail scientifique se voulait une réponse à une politique qui, pour exclure l'amazighe de la sphère officielle, invoquait l'inaptitude de cette langue à entrer dans la modernité et par là-même à survivre dans ces temps modernes, comprenions-nous. Notre action s'inscrivait dans un courant de renouveau du militantisme culturel et faisait suite à la diffusion, dans un contexte de semi-clandestinité, du lexique de mots techniques élaboré au début des années soixante-dix sous la direction de Mouloud Mammeri à Alger.

Avant d'aborder dans les pages qui suivent l'aspect nouvelles technologies à proprement parler et en finir avec ce qui peut paraître comme de la préhistoire, je signalerai que, déjà en 1990, j'avais utilisé le logiciel Latex pour transcrire la langue amazighe dans mon livre « Tusnakt s wurar » (mathématiques récréatives). Rappelons que même lorsque l'on emploie les caractères latins pour transcrire l'amazighe, le système habituellement retenu incorpore des signes diacritiques (point souscrit à certaines consonnes, chevron sur d'autres, …) et deux lettres grecques, à savoir l'epsilon et le gamma.

Durant longtemps, et pour beaucoup encore aujourd'hui, la question des polices de caractères a été pour tous les amazighisants une source de tracasseries sans fin. Pourtant, dès 1990, le logiciel Latex nous avait permis d'écrire tous les signes rencontrés dans la transcription de la langue amazighe avec un résultat typographique excellent. La solution n'a cependant pas fait école car Latex, logiciel conçu pour écrire les mathématiques, s'avère d'un usage assez lourd pour

une écriture usuelle, puisque toutes les lettres avec des marques diacritiques sont obtenues en saisissant des codes plus ou moins longs. Voici un exemple de codes pour la lettre « s » muni de trois marques diacritiques : le point souscrit, la cédille et le chevron :

Caractère	Code Latex
ṣ	\d s
ş	\c s
š	\v s

Nous avons donné les codes pour la lettre s, mais le procédé fonctionne pour toutes les lettres de l'alphabet latin dans leurs versions minuscule et majuscule. De plus, si l'on veut éviter l'espace entre le d et le s dans le code « \d s », on pourra écrire « \d{s} ». Les lettres grecques sont obtenues en écrivant leurs noms précédés du signe de commande réalisé par la barre oblique inverse « \ », ainsi γ est obtenu avec le code « \gamma » et Γ avec « \Gamma ». Outre cette lourdeur, Latex s'utilise avec un éditeur de textes qui produit de l'ASCII pur à l'exclusion de tout autre logiciel de traitement de texte. Performant mais rigide et compliqué, Latex reste irremplaçable encore aujourd'hui pour écrire des textes amazighes mathématiques en caractères latins.

C'est dans le cadre d'une mission dont je fus chargé de 2005 à 2008 par le ministère de l'éducation nationale français, que je me consacrai pleinement à l'enseignement de la langue berbère au Centre national d'enseignement à distance (CNED). L'objet de ma mission était d'assurer une préparation à l'épreuve facultative de langue berbère du baccalauréat français présentée chaque année par environ deux mille candidats. Je me suis alors retrouvé brusquement immergé dans le monde des TICE au service de la langue amazighe. Confronté à de nombreux problèmes dont celui de la transcription, les techniciens du CNED n'ont pas pu m'aider à les résoudre. Ces difficultés m'ont amené à engager une réflexion dans ce domaine et j'entamai un travail de recherche sur les nouvelles technologies et l'enseignement à distance de la langue amazighe au laboratoire Paragraphe de l'Université de Paris 8.

Dans les pages qui suivent, je vais présenter la mission qui m'a été confiée et voir comment les nouvelles technologies se sont imposées à moi dans le contexte précis de la diaspora marqué par la dispersion du public. Je réexaminerai dans la section 4 la question des polices de caractères latins à la lumière de la solution apportée par Unicode. Mais, bien entendu, j'aborderai la question des TICE sur un plan plus général, celui des normes et de l'interopérabilité ainsi que de leur pertinence en fonction de l'environnement pédagogique.

2. Environnement linguistique en diaspora, cadre institutionnel et statut juridique de l'amazighe.

Même si l'on ne dispose pas de statistiques très précises, la France est sûrement le pays, à l'exception de ceux d'Afrique du Nord, qui compte le plus grand nombre de locuteurs amazighophones. L'on peut avancer que plusieurs centaines de milliers d'entre eux sont de nationalité française tandis que les autres, nombreux également, ont gardé leur nationalité d'origine. On peut aussi affirmer que toutes les variétés de l'amazighe sont représentées en France : des parlers libyens en passant par les différents touaregs, les parlers de Tunisie, à ceux d'Algérie et du Maroc. Il serait possible de dessiner une carte de France des parlers berbères (variétés marocaines, et particulièrement le chleuh, dans le Sud-Ouest et le Nord de la France tandis que le kabyle domine en Ile de France et le chaoui plutôt dans le Sud-Est, …)

Cependant, il n'est pas sûr que cette situation perdure car une étude de l'Institut national des études démographiques (INED) menée en France montre que l'amazighe ne se transmet que très faiblement d'une génération à l'autre. La décroissance continue depuis trois ans du nombre de candidats au baccalauréat berbère corrobore l'observation de l'INED, même s'il convient d'être prudent dans l'analyse de cette baisse, trop brusque pour s'expliquer par des considérations relevant d'une autre échelle de temps (cf. graphique de la section 3.). Il n'entre pas dans le cadre de cette communication d'aller rechercher plus avant les raisons de ce phénomène ; mais au vu de certains éléments, il semble que la permanence du fait linguistique amazighe en France doive davantage au flux de nouveaux migrants en provenance d'Afrique du Nord qu'à la transmission intergénérationnelle en immigration.

Pour compléter cet aperçu de la situation de l'amazighe en France avant d'aborder le statut juridique de la langue, examinons sa position dans les circuits de l'édition et des médias.

L'édition concerne essentiellement la chanson. Des manifestations importantes (concerts, galas) sont régulièrement organisées un peu partout dans les principales villes de France. Il existe une télévision et une radio berbère (BRTV, créée en 2000) dont la transmission est assurée par satellite. En outre, plusieurs autres radios locales assurent des émissions traitant du monde amazighe et parfois même en langue amazighe (généralement dans les variétés chleuh ou kabyle).

Les nouvelles technologies de l'information ont été investies par le réseau associatif et des privés. De nouvelles télévisions apparaissent assez régulièrement sur la Toile. Il convient toutefois de tempérer cette donnée en signalant que ces initiatives sont éphémères et que les « programmes » (il serait plus juste de parler de contenus) sont extrêmement pauvres.

La présence amazighe se manifeste principalement sur les sites d'internet où il a été recensé près de 500 sites « amazighes ». Mais là encore, cette donnée doit être

nuancée en ce sens que les sites qui utilisent la langue amazighe sont une minorité et ceux qui offrent une version complète de leur contenu en langue amazighe se comptent … sur les doigts des deux mains !

2.1 Statut de la langue

L'ouverture en 1981 du droit associatif français aux étrangers a profité au mouvement culturel amazigh en France et un cours de langue amazighe a pu être assuré, dès 1984, au lycée Honoré de Balzac à Paris dans le cadre d'activités culturelles par l'association « ABRID-A ». Mais le statut de l'amazighe est demeuré inchangé en France jusqu'en 1999. À cette date, le berbère sera intégré par le gouvernement français, aux côtés de l'arabe dialectal, dans la liste des 75 langues de France en vue de l'adoption de la Charte européenne des langues minoritaires. Non ratifiée par le Parlement en raison de son inconstitutionnalité, la Charte n'aura aucun impact juridique mais le débat qui l'a entourée produira une incidence sur une institution comme la Délégation générale à la langue française (DGLF) qui deviendra la Délégation générale à la langue française et aux langues de France qui prend en charge la langue berbère.

Le 20 mars 2002, alors que nous occupions la fonction de conseiller technique au cabinet du ministre de l'éducation nationale français, sera publiée une circulaire dans le Bulletin officiel de l'éducation nationale (BOEN) qui fait pour la première fois référence à l'enseignement du berbère dans le secondaire. Dans le même mouvement, en vue de répondre à la demande de mise en place d'une préparation à l'épreuve de berbère au baccalauréat, sera créée la mission au CNED (2005-2008) et une convention sera signée avec l'Institut national des langues et cultures orientales (INALCO) en 2005.

Concluons cet état des lieux par une mesure qui sort du cadre purement français qui vient d'être évoqué puisqu'elle relève d'accords internationaux. Depuis 1999, le berbère est enseigné dans le cadre de conventions portant sur l'Enseignement des langues et culture d'origine (ELCO).

2.2 Qu'est-ce que les Elco ?

Il s'agit d'accords bilatéraux signés entre la France et certains pays lui ayant fourni une main d'œuvre. Ces accords visaient à assurer aux enfants d'immigrés un lien avec leur culture d'origine afin de leur faciliter le retour au pays parental. Nous ne nous attarderons pas sur la philosophie largement obsolète des ELCO, nous n'en parlons ici que sous l'incidence que ces conventions ont et peuvent avoir sur l'enseignement de l'amazighe en France. Voici la liste des pays ayant signé une convention ELCO :

- Portugal (1973)
- Italie, Tunisie (1974)
- Maroc, Espagne (1975)
- Yougoslavie (1977)
- Turquie (1978)
- Algérie (1981)

Les textes de ces conventions bilatérales prévoient que le pays d'origine fixe les programmes et prend en charge les enseignants, tandis que le ministère français ouvre gracieusement les portes de ses établissements et participe conjointement avec les autorités du pays d'origine à l'inspection des enseignants.

Bien que la langue amazighe ne soit nullement inscrite dans le cadre de la convention bilatérale ELCO entre l'Algérie et la France, le Haut commissariat à l'amazighité (HCA), créé en 1995, avec la mission d'introduire la langue amazighe dans l'enseignement public algérien, a fait valoir que l'amazighe avait sa place dans les accords ELCO dans la mesure où ceux-ci portaient sur les langues et cultures d'origine auxquelles appartient l'amazighe. Et dès 1999, la langue amazighe sera prise en charge par l'Ambassade d'Algérie à Paris sous l'égide du responsable des ELCO, mais cette prise en charge se fera de manière quasi-clandestine et sera circonscrite … au cadre associatif ! Les enseignants de la langue amazighe ont saisi le ministère pour que la langue amazighe soit prise en compte au même titre que la langue arabe. La tutelle algérienne de l'éducation nationale semble considérer que l'intégration officielle de l'amazighe aux ELCO nécessite un avenant à l'accord bilatéral.

Nous n'avons évoqué ici que le cas de l'Algérie pour lequel nous possédons des informations. Quant au Maroc, nous ne disposons à notre niveau d'aucun élément faisant état d'une démarche pour intégrer l'amazighe dans les ELCO.

LANGUES ENSEIGNEES	EFFECTIFS ELEVES 2007	ÉVOLUTION 2001/2007
ARABE MAROCAIN	29 292	+ 8,4%
TURC	18 604	+ 25,5%
ARABE ALGÉRIEN	12 336	+ 49 %
PORTUGAIS	9 324	- 0,05 %
ARABE TUNISIEN	5 474	- 0,5 %
ITALIEN	1 864	- 82 %
ESPAGNOL	1 386	- 2 %
SERBE	149	+ 4%
CROATE	21	
TOTAL	**78 450**	**+ 4 %**

Statistiques des ELCO pour les premier et second degrés
(Source DGESCO Bureau des écoles A1-1)

3. La mission d'enseignement du berbère au sein du CNED.

Comme cela a été signalé plus haut, l'objet de la mission était de fournir une préparation à l'épreuve facultative de berbère au baccalauréat français. Pour en évaluer le besoin pédagogique, il convient d'examiner plus précisément le contenu de l'épreuve et son public.

3.1 Le cadre réglementaire de l'épreuve.

Régies par la note de service n°2003-115 du 17-7-2003 publiée dans le BOEN n°30 du 24 juillet 2003 qui précise que certaines langues (dont fait partie le berbère) « soit ne sont enseignées que dans un nombre limité d'établissements, soit ne font pas l'objet d'un programme national et d'un enseignement réglementaire dans le système éducatif français », les épreuves de berbère sont élaborées par l'INALCO conformément à une convention signée entre cet Institut et la Direction des enseignements scolaires (DESCO) relative aux langues dites rares.

La même note définit ainsi l'épreuve elle-même :

> « L'épreuve d'une durée de deux heures vise à évaluer le degré de compréhension par le candidat d'un texte d'une longueur de vingt à trente lignes et la qualité de son expression personnelle dans la langue vivante étrangère. Le texte rédigé en langue contemporaine peut être d'origines diverses (extraits de journal, de revue, de nouvelle, de roman, etc.). Il doit être immédiatement intelligible à des locuteurs de la langue

considérée sans référence à un contexte culturel extérieur au texte.
Il est demandé aux candidats de traduire quelques lignes du texte (dix au maximum) et de répondre en langue étrangère à des questions portant sur le texte. Le barème est de 5 points pour la traduction et de 15 points pour les questions. »

Par ailleurs, l'épreuve de berbère possède la particularité de se présenter dans trois variétés linguistiques différentes : kabyle, chleuh et rifain, cette dernière variété n'a été introduite au baccalauréat qu'à partir de 1999.

Avant 1995, l'épreuve était orale et devant « l'explosion » du nombre de candidats passé de quelques dizaines dans les années soixante-dix à plus de deux mille au début des années 90 sous l'effet de la politique du regroupement familial mise en place par le gouvernement français, le ministère qui était confronté à des difficultés croissantes au moment de recruter les examinateurs pour les oraux décide, au risque de voir l'épreuve disparaître, de transformer celle-ci en épreuve écrite.

Ce passage à l'écrit changera fondamentalement la donne pédagogique car le berbère est une langue essentiellement orale et une majorité de candidats découvrent le berbère écrit le jour de l'examen. Dès lors, un impératif pédagogique apparaît : fournir aux futurs candidats les notions de base pour l'écriture et la lecture du berbère.

3.2 Les contraintes.

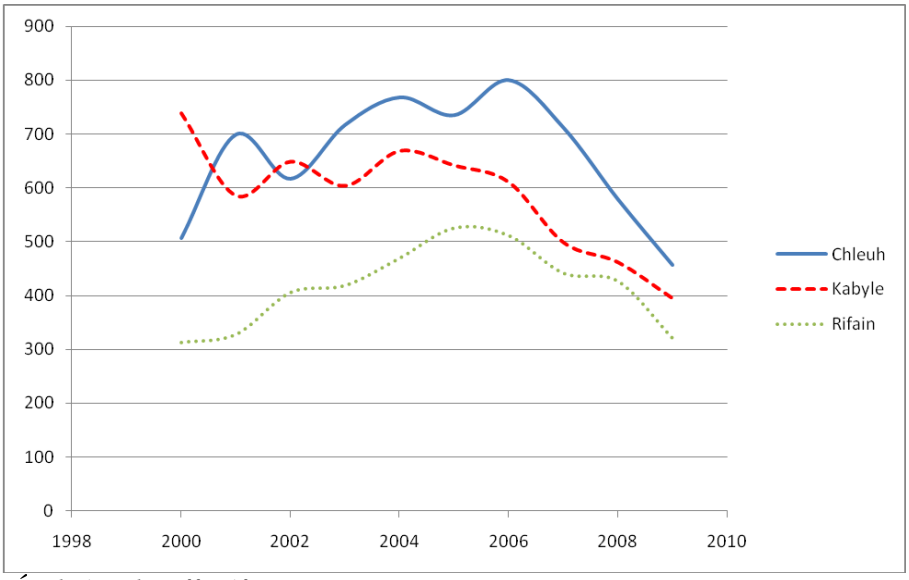
Le choix de l'option enseignement à distance s'est imposé à nous en raison des contraintes qui pèsent sur cette mission. Elles sont de deux ordres :

La première, qui tient au statut du berbère en France, est déjà, à elle seule, déterminante. Dans l'enseignement public en France, il n'existe pas de poste d'enseignant de berbère dans le secondaire ou le primaire. Les solutions qui peuvent être trouvées ne sont que des palliatifs.

D'une autre nature, la seconde contrainte qui concerne le public visé pèse tout aussi lourdement sur l'organisation de la formation à mettre en place. L'extrême dispersion des candidats à travers les établissements de France fait qu'il est très difficile, en dehors de quelques rares cas situés dans les grandes agglomérations, d'envisager l'organisation de classes traditionnelles même en regroupant des élèves provenant d'établissements voisins. Les deux mille candidats sont répartis sur plus de cinq cents établissements disséminés à travers le territoire national français. Même l'académie de Créteil, qui arrive dans le peloton de tête du point de vue numérique, n'offre qu'une moyenne de 3,57 élèves par lycée (ne sont retenus dans ce calcul fait pour l'année 2007 que les établissements qui présentent au moins un candidat toutes variétés linguistiques confondues !)

À cette dispersion globale, s'ajoute la volatilité des effectifs d'une année sur l'autre au sein d'un même établissement qui complique singulièrement la mise en place d'un enseignement présentiel. Par exemple, le Lycée Bergson dans le dix-neuvième arrondissement de Paris qui présentait 9 (neuf) candidats en 2000 n'en présente

plus que 3 (trois) en 2004, sans que cette diminution ne traduise une tendance à la baisse du nombre global de candidats (lequel a en réalité augmenté durant cette même période !). Un autre type de fluctuation est relevé : même lorsque le nombre global de candidats dans un même établissement reste sensiblement stable, la répartition entre les différents parlers peut varier. Par exemple, on observe ainsi que le nombre de candidats à l'épreuve de chleuh peut monter au détriment du kabyle ou inversement.



Évolution des effectifs, par variante linguistique, des candidats ayant présenté l'épreuve (source DGESCO).

Ces caractéristiques font du CNED l'opérateur indiqué pour assurer la préparation à l'épreuve de berbère au baccalauréat. Enfin, l'encouragement du Ministère de l'Education Nationale à recourir aux TICE nous a conforté dans notre choix en faveur de l'EAD.

3.3 Le contenu pédagogique.

Compte tenu de ces données et vu les moyens dont dispose la mission, nous nous sommes fixé comme premier objectif la construction d'un site d'internet.

Comme l'épreuve de berbère se présente concrètement sous la forme de trois sujets indépendants- un par variante -, il nous fallait impérativement disposer d'un expert par variante. Or la mission n'était dotée que d'un poste à plein temps, celui du chargé de mission maîtrisant le kabyle.

La première préoccupation était donc de nous adjoindre deux collaborateurs externes, l'un pour le chleuh, en la personne de Abdelaali Talmenssour, alors doctorant à l'Université d'Orléans, l'autre pour le rifain, en faisant appel à

Abderahmane El Aïssati, maître de conférences à l'Université de Tilburg, tous deux rémunérés grâce au partenariat mis en place avec la DGLFLF.

Le contenu du site se scindait en deux espaces : l'un, commun à toutes les variantes, était consacré à l'information générale, l'autre, dédié aux spécificités, se subdivisait en trois sous-espaces, un par variante linguistique.

Dans la partie commune se trouvent mis en ligne une notice historique sur l'écriture du berbère qui intègre une présentation des inscriptions libyques, une étude sur la littérature berbère et quelques conseils donnés aux élèves regroupant à la fois une information administrative et des suggestions pédagogiques. Il s'agit là, bien sûr, d'une première étape appelée à être enrichie par diverses contributions.

Conçue par le chargé de mission, cette partie a été préparée avec l'aide de collaborateurs de haut niveau scientifique qui ont bien voulu apporter leur contribution bénévole.

La rémunération dont bénéficiaient les deux experts externes ne permettait pas d'exiger d'eux une implication aussi importante que celle du chargé de mission pour l'élaboration des rubriques consacrées aux différentes variantes au sein de « l'espace spécifique ». Nous avons donc procédé selon la formule suivante : le chargé de mission a préparé le schéma général et réalisé la partie relative au kabyle. Puis, les deux autres experts se sont chargés, chacun de son côté, de l'adaptation de la rubrique kabyle à leurs variantes respectives.

Chaque rubrique s'ouvre sur le point essentiel du système de transcription utilisé à l'épreuve facultative de berbère au baccalauréat. La présentation concise et complète de cette question constitue la clé de voûte de la préparation. Dans cette partie, notre souci a été de parvenir, en un minimum de lignes, à une description cohérente du système de transcription tout en restant accessible au lecteur débutant. Tous amazighophones, les candidats ignorent dans leur quasi-totalité tout de l'écriture de leur langue. Ce constat nous a conduit à adopter une méthode basée sur des exemples simples que nous avons supposés connus du lecteur. Évitant délibérément le recours abusif au concept théorique, nous avons donné les clés de lecture (et d'écriture) en ciblant les spécificités du berbère qui peuvent paraître nouvelles au débutant avant de donner à la fin un tableau récapitulatif de tous les signes utilisés.

Plutôt que de faire une unique présentation « pan-berbère », générale et systématique qui engloberait tous les cas, nous en avons fait trois, une par variante, afin d'en faciliter l'accès au candidat : celui-ci n'aura pas à filtrer, à partir d'une information globale, celle qui lui est destinée en propre. Par contre, dans le choix des exemples, nous avons privilégié ceux qui étaient communs aux trois variantes.

Cette partie a bénéficié des conseils de Lionel Galand, ancien Directeur de conférences à la quatrième section de l'Ecole pratique des Hautes études et membre correspondant de l'Institut.

Viennent ensuite les sujets classés par année suivis des corrigés et de commentaires. Les traductions françaises des textes berbères proposées dans ces annales sont reprises des ouvrages, quand ils sont bilingues, dont sont extraits les sujets. Quand seul le texte berbère est publié, notre équipe pédagogique a elle-même procédé à la traduction.

Suivant les recommandations de l'Inspection générale, nous avons donné, avec les corrigés, des éléments succincts relatifs aux textes étudiés (nature du texte, place qu'il occupe dans la littérature berbère, vocabulaire utilisé, …). Des indications bibliographiques permettent de guider le candidat désireux de s'informer au-delà des réponses apportées aux questions posées dans le sujet.

Afin d'assurer un accompagnement pédagogique personnalisé, il a été donné aux élèves la possibilité de poser des questions en ligne.

Enfin, signalons des difficultés techniques non négligeables, familières à tous ceux qui pratiquent la langue berbère écrite, apparues au moment de procéder à la saisie des textes berbères. Le système de transcription à base latine comportant deux lettres grecques et des marques diacritiques n'est disponible sur aucun clavier classique. Pour contourner cet obstacle, différents auteurs ont construit plusieurs polices de caractères dont certaines sont mises à la disposition du public. Malheureusement, dans notre cas, ces solutions se sont avérées inopérantes car inaptes à franchir le cap de la transmission par l'internet. L'échange via l'internet n'est possible que si l'usager travaille avec une machine sur laquelle est déjà installée la même police de caractères.

Nous avons construit un clavier utilisant les polices Unicode. Mais les services techniques du CNED ont préféré gérer des fichiers PDF obtenus à partir des polices Unicode. Solution qui n'a pas résolu la difficulté liée à l'échange avec les élèves… Nous reviendrons dans la section qui suit sur la question de la transcription, et tout spécialement sur le système basé sur l'alphabet latin.

4. Les questions de la standardisation, de l'interopérabilité et des normes.

À des degrés divers, aucune situation pédagogique de la langue amazighe n'échappe au problème omniprésent de la standardisation.

4.1 Des aspects généraux.

Même si l'on se place dans l'hypothèse où un seul dialecte est pris en compte, la variété linguistique, à l'intérieur même de ce dialecte qui se subdivise à son tour en plusieurs parlers, est telle que l'on ne peut ignorer cette question. L'on y est confronté en particulier lorsque l'on opère la sélection des textes à étudier et qui serviront de référents pédagogiques. L'on a pu constater que même sur les sujets proposés au baccalauréat français, où pourtant la règle s'en tient à la langue authentique, en n'autorisant que des textes extraits de journaux, livres, etc., que les

textes produits ont bien souvent été « aménagés » par les concepteurs des sujets. Parmi les interventions opérées par les auteurs des sujets, certaines sont manifestement motivées par la volonté d'écarter des expressions d'une portée par trop locale.

Au regard de la situation française, la question de la standardisation se pose bien entendu à un autre niveau pour l'Algérie ou le Maroc qui entendent enseigner l'amazighe, comme langue nationale. En Algérie, les pouvoirs publics s'engagent à mettre en place un enseignement là où la demande est exprimée alors qu'au Maroc les cours de l'amazighe sont censés s'adresser à terme à tous les élèves marocains, bien que dans ce pays le statut de langue nationale fait encore défaut à l'amazighe, du moins dans le texte de la Constitution.

Jusqu'où aller dans la standardisation ? Là encore les lois devant encadrer la réponse à cette question diffèrent entre l'Algérie et le Maroc. Depuis 2002, la Constitution algérienne dispose dans son article 3 bis :

> Art. 3 bis. — Tamazight est également langue nationale. L'Etat œuvre à sa promotion et à son développement dans toutes ses variétés linguistiques en usage sur le territoire national.

Ce qui peut laisser supposer une orientation opposée à une convergence des différents dialectes puisque l'Etat cultive toute la variété en usage sur le territoire national !

> *L'on comparera avec l'article suivant extrait du Dahir de 2001 qui a créé l'Ircam et régit l'enseignement de l'amazighe au Maroc :*
>
> *Article 2 : L'Institut, saisi par Notre Majesté à cette fin, nous donne avis sur les mesures de nature à sauvegarder et à promouvoir la culture amazighe dans toutes ses expressions.*
>
> *En collaboration avec les autorités gouvernementales et les institutions concernées, l'Institut concourt à la mise en œuvre des politiques retenues par Notre Majesté et devant permettre l'introduction de l'amazigh dans le système éducatif et assurer à l'amazigh son rayonnement dans l'espace social, culturel et médiatique, national, régional et local.*

S'il est fait mention ici de la promotion de la culture amazighe dans toutes ses expressions, il convient de noter que s'agissant de la langue, c'est le singulier qui prévaut et c'est « l'introduction de l'amazigh dans le système éducatif » qui est préconisée.

Cependant, si l'on examine les pratiques des chercheurs dans les deux pays, elles ne se calquent pas forcément sur ces textes et les oppositions sont moins fortes que ne le laisserait supposer la lecture de ces deux extraits de loi.

Au-delà du cadre juridique, les questions de standardisation touchent à d'autres domaines comme celui de la néologie et dans ce cas, la réponse à cette question est déterminée par celle apportée à une autre question fréquemment soulevée :

envisage-t-on un enseignement de l'amazighe ou bien, à terme, un enseignement en amazighe ? Par exemple, dans le cas de l'enseignement à distance, si l'on se contente d'enseigner la langue telle qu'elle existe, l'on passera par une langue de travail (français, arabe, anglais, …) pour gérer l'environnement pédagogique. Mais l'on ne peut se satisfaire de cette solution (sur le long terme) si l'on adopte une autre perspective, celle de langue d'enseignement.

Il s'agit d'un problème complexe et lourd qui a déjà suscité nombre de travaux, en particulier ceux publiés en 2004 sous l'égide de l'IRCAM. À l'évidence, le problème n'est pas purement linguistique et dépend avant tout du niveau d'implication des Etats nord-africains (volonté politique, moyens financiers, mise en place d'institutions idoines, …). En attendant que cette implication se manifeste de manière plus vigoureuse, l'absence de structures académiques reconnues consensuelles (tout particulièrement en Algérie et en diaspora) constatée dans le domaine de l'édition, de la création de néologismes a conduit à une situation anarchique dommageable. L'utilité de telles structures se fait sentir également au niveau international. Il est tout à fait souhaitable qu'une coordination effective soit assurée entre l'Algérie et le Maroc pour la création terminologique par exemple. C'est par de tels dispositifs que l'on peut infléchir le processus d'éclatement de l'amazighe et impulser une autre tendance, celle de la convergence (au moins au niveau des grands ensembles linguistiques), convergence sans laquelle cette langue ne sera jamais une langue de communication entre groupes amazighophones parlant des dialectes différents. Et dans un monde où la mobilité des personnes a atteint un niveau sans commune mesure avec ce qu'ont connu les générations précédentes qui toute leur vie durant ont pu se satisfaire de la maîtrise d'un parler très local, y-a-t-il encore aujourd'hui place pour un amazighe « poussière de parlers », pour reprendre une expression célèbre ?

4.2 L'écriture.

La question de la standardisation concerne également l'écriture. Bien qu'a priori plus technique, cette question soulève des polémiques violentes et récurrentes en raison de la charge symbolique attachée aux différents alphabets en lice pour écrire l'amazighe. L'on sait que le Maroc a opté pour le tifinaghe et que l'Algérie n'a pas arrêté son choix puisque les trois alphabets latin, tifinaghe et arabe sont utilisés dans l'enseignement avec une préférence accordée à l'alphabet latin. La polémique risque de repartir depuis que la chaîne de télévision émettant en amazighe qui vient d'être lancée en 2009 transcrit systématiquement en alphabet arabe les textes en amazighe. En France, les sujets du baccalauréat sont transcrits en caractères latins uniquement et la production littéraire se fait pour l'essentiel dans le même système d'écriture, même si les caractères tifinaghes ne sont pas totalement absents. Ce constat vaut plus généralement pour l'Europe et le Canada où réside une communauté amazighophone dynamique.

Il est clair cependant que la réalité en diaspora ne saurait échapper durablement à l'influence de ce qui se fait en Afrique du Nord où s'opéreront les choix décisifs sur le long terme.

Au regard de cet état de choses caractérisé par la pluralité, qui parfois s'apparente à des antagonismes, il peut paraître paradoxal d'affirmer que la divergence est, au plan de l'écriture, beaucoup moins importante qu'il n'y paraît.

En premier lieu, il importe de dire que les règles d'écriture ne se résument pas au choix de l'alphabet. Du strict point de vue de l'enseignement à distance via un support électronique, la querelle graphique se pose en termes beaucoup moins tragiques puisque l'on peut commuter d'un système d'écriture à l'autre d'un simple clic de souris. Les échanges pourront s'opérer sans difficultés, y compris entre des partenaires qui n'utilisent pas la même graphie pour peu que les règles d'écriture soient les mêmes, c'est-à-dire que soit partagée l'analyse phonologique et morphosyntaxique de la langue. L'usage d'un même alphabet n'aboutit pas nécessairement à une écriture unique ; l'on s'en rend compte en consultant les ouvrages du dix-neuvième siècle et de la première moitié du vingtième écrit en caractères latins. Les sujets du baccalauréat qui en sont extraits sont totalement réécrits et pas seulement parce que les conventions adoptées pour transcrire les phonèmes qui n'existent pas dans l'alphabet latin diffèrent suivant les époques. Les coupures de la chaîne phonétique ne sont pas opérées aux mêmes endroits et certaines variations phonétiques, notées dans les anciens textes, ne le sont plus dans les notations à tendance phonologique d'aujourd'hui… C'est à ces règles orthographiques - indépendantes de l'alphabet employé- que nous avons consacré notre communication (Sadi, 1992) au colloque international de Ghardaïa de 1991. Bien qu'il subsiste encore des points de divergence sur ce sujet, l'on peut se féliciter de ce qu'aujourd'hui un large consensus s'est progressivement installé dans ce domaine.

On peut en fixer le point de départ au moment où les auteurs amazighophones ont commencé à écrire pour d'autres amazighophones. Ils se sont approprié l'écriture en optant pour une notation à dominante phonologique négligeant des nuances phonétiques que le lecteur amazighophone restitue de lui-même. Un tournant important dans cette évolution est marqué par la publication de l'ouvrage « Isefra » (Mouloud Mammeri, 1969) dont la notation ne distingue plus les consonnes occlusives des spirantes pourtant si fréquentes en kabyle.

Reste que l'aspect matériel a longtemps été une difficulté considérable. Aujourd'hui, le problème est réglé pour ce qui est de l'écriture fondée sur l'alphabet latin puisque tous les caractères et les marques diacritiques requises figurent dans Unicode et que les glyphes existent dans les polices standards, bien que nombre d'utilisateurs continuent encore à recours aux polices de caractères « bricolées ». La table des caractères accompagnés de leurs codes hexadécimaux et de leurs noms Unicode (qui ne sont pas traduits) reproduite en annexe le montre

clairement. L'avantage de la solution Unicode est qu'elle est prise en charge par les navigateurs récents et que l'on peut donc communiquer via l'internet sans avoir à installer au préalable sur la machine destinataire la même police utilisée pour écrire le texte émis.

Il faut encore préciser que, même si l'on n'utilise que les caractères latins, il convient de prendre quelques précautions dans le codage. Les caractères composés (consonne avec un point souscrit, une cédille ou bien un chevron,…) peuvent être codés de deux manières différentes : un code unique qui restitue le caractère composé et un double code, le premier pour la consonne et le second pour la marque diacritique combinée. Prenons un exemple : le caractère č « c avec un chevron » peut être obtenu à la fois à partir du seul code Unicode 010D mais également comme un caractère combiné c (dont le code est 0063) suivi du chevron combiné (dont le code est 030C). Cette propriété a déjà été signalée dans (Brugnatelli, 2002) et (Zenkouar, 2008). Les deux auteurs ont mis en garde contre les difficultés que le codage de ces signes pouvait engendrer pour certains moteurs de recherche lorsque l'on procède à la correction orthographique. Ajoutons, et c'est fondamental pour le point que nous traitons ici, que la même difficulté surgit lorsque l'on commute les systèmes de transcription. Pour contourner ces obstacles, nous proposons d'éviter le double codage de ces caractères et de n'utiliser que le code unique qui correspond au caractère composé.

Pour clore ces remarques, arrêtons-nous un instant sur les deux lettres grecques que la transcription de l'amazighe incorpore à l'alphabet latin, à savoir l'epsilon et le gamma. Apparues dans l'écriture de l'amazighe assez tardivement, ces deux lettres ont été adoptées sous l'influence des sémitisants qui les utilisaient déjà pour les langues sémitiques. Absentes des claviers latins, elles ont aussi posé les mêmes problèmes que les marques diacritiques. Ainsi dans certains documents, pour réaliser le gamma, l'on écrivait la lettre v et l'on ajoutait à la main une boucle pour obtenir ɤ. Est-ce cette pratique qui a conduit le consortium Unicode à introduire les caractères ɤ et Ɣ ?

Caractère	Code	Nom	Plage d'Unicode
Ɣ	0194	Latin capital letter gamma	Latin étendu-B
ɣ	0263	Latin small letter gamma	Extensions IPA

Nous proposons de retenir pour la notation à base latine de l'amazighe le gamma grec dans les versions minuscule γ (code 03B3, plage Grec et copte) et majuscule Γ (code 0393, plage Grec et copte).

Enfin, un problème particulier est posé par la lettre grecque epsilon ε (code 03B5, plage Grec et copte) dont la majuscule est réalisée par le signe E (code 0395, plage Grec et copte). La confusion au niveau graphique de la majuscule epsilon grecque avec le E majuscule latin (code 0045, plage Latin de base) est totale, même si les

codes d'Unicode ne sont pas identiques. Pour pallier ce problème, différents éditeurs ont utilisé, à la place de l'epsilon majuscule, le sigma majuscule grec Σ (code 03A3, plage Grec et copte) qui est aussi représenté dans la plage Latin étendu-B sous le code 01A9. C'est ce signe Σ qui a été adopté par exemple dans le livre « Inna yas Ccix Muḥend » (Mouloud Mammeri, 1989) pour noter la majuscule epsilon. Or tout en restant dans Unicode, il existe une meilleure solution en utilisant le E ouvert qui donne dans sa version minuscule ɛ (code 025B) et majuscule Ɛ (code 0190).

Précisons que tous ces signes sont disponibles dans les polices de caractères usuelles (Times new roman, Arial, Garamond, Gentium, Tahoma,…). Pour travailler dans un maximum de confort, le mieux est d'installer un clavier amazighe latin que l'on pourra personnaliser à sa guise, opération réalisable sous l'environnement Windows par exemple avec le logiciel MSKLC et pour les machines Apple sous environnement MacOS X on pourra se reporter au chapitre 5 du livre « Fontes et Codages » (Haralambous, 2004). Dans la création du clavier, l'on pourra retenir comme principe celui d'adopter la touche de l'accent circonflexe comme touche morte permettant d'associer une marque diacritique (quelle qu'elle soit) à une consonne. Le cas des consonnes pouvant recevoir deux marques diacritiques, comme par exemple la lettre t qui accepte le point souscrit et la cédille, doit être traité à part. Pour les lettres gamma et epsilon, on pourra encore les obtenir à partir de la touche morte associée à d'autres touches que l'on choisira. Par ce procédé, on parvient à un clavier très ergonomique qui permet d'écrire à la fois le français et l'amazighe sans avoir à changer de clavier. Ce procédé est celui proposé sur le site http://www.imyura.net/ (sauf pour le Ṭ qui est exclu du système de notation retenu par Imyura, le Ɛ qui est transcrit par Σ et le G avec chevron Ǧ noté par un G marqué avec le signe de brève, soit Ğ). Aujourd'hui, nous pouvons considérer que, grâce à Unicode, le problème d'écriture de l'amazighe en caractères latins avec un ordinateur et de transmission de textes sur la Toile est définitivement réglé.

5. Conclusion.

Retenons pour conclure que la méthode de travail utilisée, les instruments mis en œuvre dans le cadre de notre mission, le recours à l'EAD en particulier qui permet de faire collaborer des personnes ressources géographiquement éloignées les unes des autres, restent pertinents bien au-delà du cadre français et intéressent d'une manière générale tous ceux qui cherchent à enseigner une langue minoritaire en diaspora. Dans le cas de la langue amazighe, cette expérience peut profiter également aux actions conduites dans les pays d'Afrique du Nord où le manque d'encadrement en enseignants et la pénurie de documentation adaptée sont soulignés par tous. L'EAD ou la formation à distance ont un rôle à jouer, pas nécessairement en remplacement d'un enseignement présentiel, mais en complément, en appui de celui-ci. Cela vaut particulièrement pour l'Algérie qui ne

préconise pas un enseignement généralisé de l'amazighe et qui aura donc à connaître les difficultés liées à la dispersion du public amazighophone en zone arabophone. Enfin, la formation à distance peut aider à rompre l'isolement dont se plaignent trop d'enseignants de la langue amazighe.

Bibliographie

Agnaou, F. (2009). Vers une didactique de l'amazighe. Asinag, num.2, Rabat, p. 21-30.

Andries, P. (2008). Unicode 5.0 en pratique, Editions Dunod, Paris.

Boukhris, F. (2009). L'Université du possible : quelle place pour les études amazighes ? Asinag, num. 2, p. 31-44.

Boukous, A. (2004). La standardisation de l'amazighe : quelques prémisses. In Actes du séminaire organisé par le Centre de l'aménagement linguistique. Publications de l'Institut royal de la culture amazighe, Rabat, p. 11-22.

Brugnatelli, V. (2002). Tamazight et Unicode. La standardisation dans le domaine des ordinateurs. In Actes du Colloque international Tamazight face aux défis de la modernité, Juillet, Alger, p. 215-227.

Haralambous, Y. (2004). Utilisation d'Unicode (chap.5). In Fontes et codages, Editions O'Reilly France, p. 155-182

Mammeri, M. (1969). Isefra, Editions Maspéro, Paris.

Mammeri, M. (1989). Inna yas Ccix Muḥend, édition à compte d'auteur, Alger.

Sadi, H. (1990). Urar s tusnakt (mathématiques récréatives), Editions Asalu-ACB, Alger-Paris.

Sadi, H. (1992). Questions pour une orthographe de la langue courante. In Actes du colloque international de Ghardaïa du 20-21 avril 1991, Unité et diversité de tamazight, Alger, p. 96-114.

Zenkouar, L. (2004). L'écriture amazighe tifinaghe et Unicode. Etudes et Documents Berbères, num. 22, Paris, pp. 175-192.

Annexe

TABLE DES CARACTERES SPECIAUX UNICODES UTILISES POUR LA TRANSCRIPTION A BASE LATINE DE L'AMAZIGH.

SIGNE	CODE	NOM	PLAGE UNICODE
Č	010C	Latin capital letter c with caron	Latin étendu-A
č	010D	Latin small letter c with caron	Latin étendu-A
Ǧ	01E6	Latin capital letter g with caron	Latin étendu-B
ǧ	01E7	Latin small letter g with caron	Latin étendu-B
Ţ	0162	Latin capital letter t with cedilla	Latin étendu-A
ţ	0163	Latin small letter t with cedilla	Latin-étendu-A
Ḍ	1E0C	Latin capital letter d with dot below	Latin étendu additionnel
ḍ	1E0D	Latin small letter d with dot below	Latin étendu additionnel
Ḥ	1E24	Latin capital letter h with dot below	Latin étendu additionnel
ḥ	1E25	Latin small letter h with dot below	Latin étendu additionnel
Ṛ	1E5A	Latin capital letter r with dot below	Latin étendu additionnel
ṛ	1E5B	Latin small letter r with dot below	Latin étendu additionnel
Ṣ	1E62	Latin capital letter s with dot below	Latin étendu additionnel
ṣ	1E63	Latin small letter s with dot below	Latin étendu additionnel
Ṭ	1E6C	Latin capital letter t with dot below	Latin étendu additionnel
ṭ	1E6D	Latin small letter t with dot below	Latin étendu additionnel
Ẓ	1E92	Latin capital letter z with dot below	Latin étendu additionnel
ẓ	1E93	Latin small letter z with dot below	Latin étendu additionnel
◌̣	0323	Combining dot below	Marques diacritiques d'association
◌̌	030C	Combining caron	Marques diacritiques d'association
◌̧	0327	Combining cedilla	Marques diacritiques d'association
γ	03B3	Greek small letter gamma	Grec et copte
Γ	0393	Greek capital letter gamma	Grec et copte
Ɛ	0190	Latin capital letter open e	Latin étendu-B
ɛ	025B	Latin small letter open e	Extensions IPA

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

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

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

Mots clefs

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

1. Introduction

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

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

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

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

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

2. Base des systèmes d'écriture informatiques

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

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

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

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

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

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

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

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

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

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

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

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

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

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

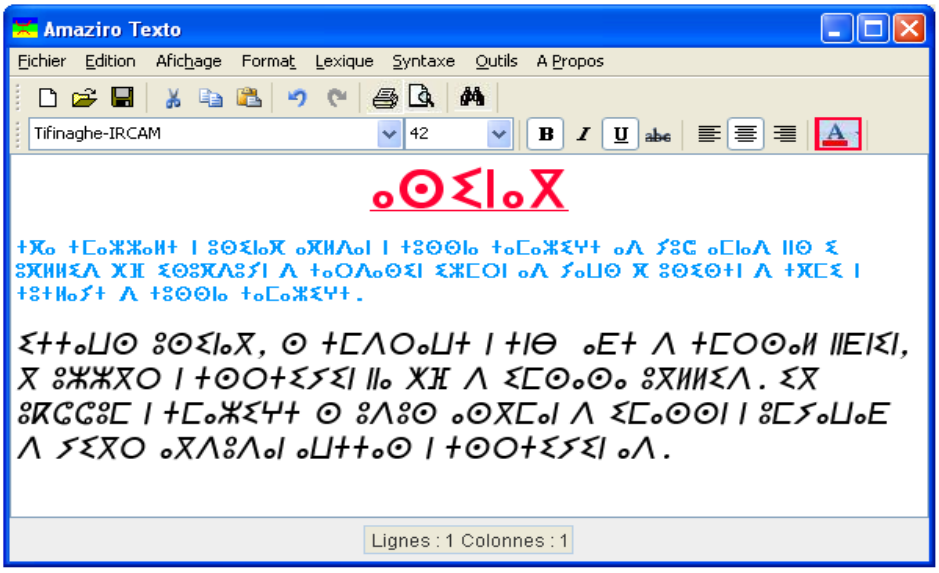


Figure 1 – Application Editeur de texte Amazighe

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

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

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

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

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

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

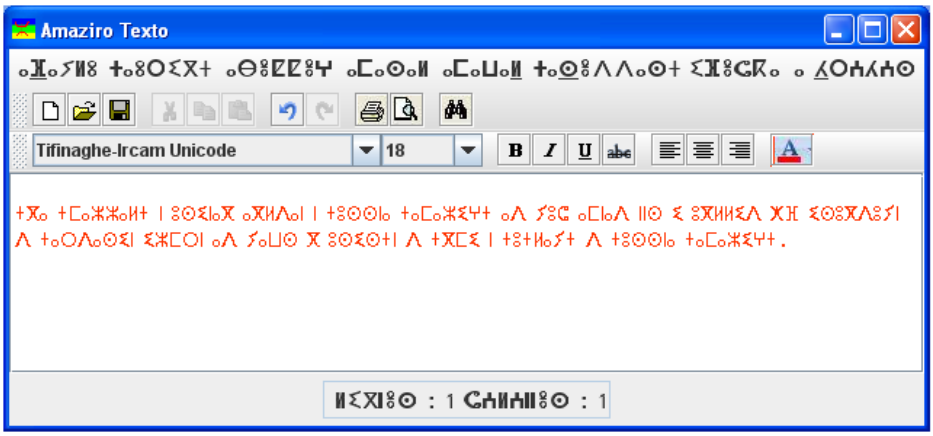


Figure 2 – Application Editeur de texte Amazighe avec l'interface en Tifinaghe

5. Conclusion et Perspectives

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

Références

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

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

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

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

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

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

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

Acabit : un outil d'extraction des termes complexes

S.BOULAKNADEL, B.Daille, D.Aboutajdine
Siham Boulaknadel, GSCM_LRIT Université Mohammed V, Agdal Rabat, Maroc
siham.boulaknadel@univ-nantes.fr
Beatrice Daille, LINA FRE CNRS 2729 Université de Nantes, Nantes, France, beatrice.daille@univ-nantes.fr
Driss Aboutajdine, GSCM_LRIT Université Mohammed V, Agdal Rabat, Maroc
aboutaj@fsr.ac.ma

Résumé : ACABIT est un outil de construction semi-automatique de banques terminologiques. Il permet d'une part de faciliter la tâche des experts en leurs proposant des candidats potentiels, et d'autre part d'imposer à l'aide des règles linguistiques un format morphosyntaxique aux termes (et aux experts), de manière à obtenir une liste uniforme et cohérente des termes de base du domaine.
Cet outil s'appuie conjointement sur des données linguistiques et sur des modèles statistiques. Une première sélection de candidats potentiels est extraite automatiquement d'un corpus étiqueté et lemmatisé, en utilisant les spécifications linguistiques exprimées en termes de structures morpho-syntaxiques. C'est ensuite sur ces candidats potentiels que s'applique le modèle statistique : le coefficient de vraisemblance. La tâche de ce dernier est de trier et de fournir une liste ordonnée de termes de base candidats, c'est-à-dire du plus au moins représentatif du domaine. ACABIT a traité des corpus en langue arabe, française, anglaise et japonaise. Or la partie application de cette contribution se focalisera sur la langue arabe.

1 Introduction

L'acquisition des termes est un domaine qui a fait l'objet de nombreux travaux de recherches ces vingt dernières années. Deux grandes directions ont été empruntées pour recenser automatiquement les termes : les modèles linguistiques (Bourigault 93), et les modèles statistiques (Smadja 93). Or, des nouvelles recherches entreprises au cours de la dernière décennie tendent à tirer profit de ces deux grandes approches pour proposer des méthodologies hybrides, qui ne sont ni purement linguistiques, ni purement statistiques. C'est sous cette dernière méthodologie que nous pourrons classer l'outil d'acquisition des termes Acabit, qui se compose de deux étapes :

- Un repérage linguistique des termes à l'aide de règles simples appliquées au corpus étiqueté,

- Un filtrage statistique des candidats termes retenus à l'étape précédente.

L'article est organisé de la manière suivante : la section 2 présente les données linguistiques de l'outil d'extraction des termes Acabit ensuite le filtrage statistique est détaillé à la section 3, l'application est décrite dans la section 4 et la conclusion dans la section 5 termine l'article.

2 DONNEES LINGUISTIQUES

La section suivante présente l'analyse linguistique qui consiste à extraire des structures morphosyntaxiques et leurs variantes en utilisant des règles morphologiques.

2.1 Les termes de base et leurs variations

A partir d'une étude linguistique effectuée sur une banque terminologique, il apparaît que les termes binaires où seules sont prises en compte les unités lexicales non fonctionnelles telles que les noms, les adjectifs (ou participes passés) et les adverbes séparés par des blancs dans l'écriture sont de loin les plus nombreux. L'approche statistique exigeant une bonne représentation du nombre d'échantillons, ACABIT se concentre sur l'extraction des termes de longueur 2, appelés « termes de base », et qui s'apparient à l'une des structures morpho-syntaxiques suivantes :

Nom Adj : instruction publique
Nom1 Prep Nom2 : principe d'égalité
Nom1 Prep Det Nom2 : apprentissage de la lecture
Nom1 Nom2 : apprenti lecteur
Nom1 à Vinf : savoir à enseigner

Cependant, les termes ne sont pas des unités lexicales figées et subissent des variations morphologiques et syntaxiques. Les variations ci-dessous sont prises en compte par la grammaire :
- Variations graphiques
- Variations flexionnelles
- Variations morpho-syntaxiques (variation de la préposition, présence ou non d'un déterminant)
- Variations syntaxiques (insertion de modifieurs, coordination).

2.2 Méthodologie d'extraction

Les termes binaires sont considérés comme des cooccurrences particulières qui possèdent les propriétés linguistiques ci-dessus : ils sont définies par rapport à leur structure morpho-syntaxique ; ils admettent des variantes.

Une grammaire locale permettant d'identifier les candidats termes et leurs variantes a été écrite en FLEX (librairie GNU sous UNIX ou LINUX). Une séquence morphosyntaxique reconnue par l'une des règles de grammaire constitue une occurrence d'un couple. Un couple est constitué de deux lemmes qui correspondent aux deux extrémités lexicales de la séquence ; par exemple, le couple (didactique, lecture) correspond aux séquences suivantes : didactique de la lecture, didactique déclarative de la lecture, didactique expérimentale de la lecture. Chaque séquence relevée est accompagnée de son schéma morphosyntaxique et de sa position dans le corpus (fichier, phrase).

3 FILTRAGE STATISTIQUE

ACABIT utilise dans un deuxième temps les résultats d'une évaluation de différentes mesures statistiques. Cette évaluation a permis de découvrir la meilleure mesure pour cette application, c'est-à-dire celle qui assigne un score élevée aux séquences les plus susceptibles de constituer des termes parmi la liste de candidats.

Les candidats termes sont alors triés selon le score statistique et le programme propose en sortie une liste ordonnée de couples.

4 APPLICATION

Les outils pour l'acquisition de termes sont assez bien décrits pour l'anglais ou le français, mais la question reste peu étudiée pour la langue arabe, qui se caractérise par sa flexion plus riche et son ordre des mots plus libre. Nous présentons l'outil d'extraction des termes arabes dans le domaine de l'environnement. Tout d'abord, nous décrivons notre corpus sur lequel a été mené l'étude, ensuite nous définissons les spécifications des termes complexes et leurs variations en langue arabe et nous soumissions cette liste des termes complexes à des mesures statistiques pour déterminer le potentiel terminologique de la séquence rencontrée.

4.1 Corpus

Nous avons décidé de construire un corpus à partir du web dans le domaine de l'environnement, restreint aux thématiques suivantes : la pollution, la purification de l'eau, la dégradation du sol, la préservation de la forêt, les catastrophes naturelles. Elles font l'objet d'une importante production langagière en arabe, comme l'atteste la présence de nombreux sites sur le web.

L'élaboration du corpus s'est déroulée en deux étapes : la récolte du web, et la normalisation des textes. Les étapes ont été réalisées par des locuteurs natifs.

4.1.1 Récolte du web

Pour la récolte des documents, nous avons effectué :

1) une recherche sur le web à l'aide du moteur http://www.google.com/intl/ar/ pour l'arabe.
2) une recherche interne sur des portails, notamment "Al-Khat Alakhdar" [20] et " Akhbar Albiae"[21] , en utilisant dans le cas échéant le moteur de recherche propre au site.

4.1.2 Normalisation

Pour chacun des documents sélectionnés, nous avons enregistré son url, et l'avons converti au format UNICODE sous type.txt.

4.1.3 Caractéristiques de la collection

Le tableau ci-dessous présente quelques indications des caractéristiques de notre corpus [AR-ENV].

Caractéristiques	Collection $[AR-ENV]$
Nombre de documents	1062
Nombre total de mots	475 148
Nombre total de mots différents	54 705

Table 1: Quelques données sur la collection de test $[AR-ENV]$

4.2 Méthodologie

Pour l'arabe, nous ne disposons pas de textes annotés. Alors, nous avons utilisé un analyseur de surface sur les textes bruts. La description des tâches effectuées pour l'arabe est présentée dans les sections qui suivent.

4.2.1 Partie linguistique

Pour l'identification des termes candidats, nous avons utilisé une plate-forme d'analyse de surface ASVM, complétée d'un analyseur morphologique de l'arabe (Diab *et al*. 04). L'ASVM est une adaptation du système anglais YamCha basé sur les Support Vector Machines. En outre, nous avons développé une grammaire de surface pour les termes complexes de l'arabe qui tient compte des principales structures nominales : l'accord en nombre, en genre entre le nom et l'adjectif, et les syntagmes à complément génitif. Ci-dessous des exemples tirés de notre corpus. Pour la description des termes arabes nous avons utilisé la transcription de Buckwalter[22].

[20] 1http://www.greenline.com.kw
[21] http://www.4eco.com

[22] http://www.qamus.org/transliteration.htm

Patron	Sous-patron	Terme	Traduction anglaise
N ADJ		Altlwv AlkmyAAy	chemical pollution
N1 N2		tlwv AlmAA	water pollution
	N1 b N2	Altlwv b AlrsAs	pollution with lead
N1 PREP N2	N1 l N2	AltErD l AlAmrAD	exposure to diseases
	N1 mn N2	Altxls mn AlnfAyAt	waste disposal

Table 2: Patrons syntaxiques

Variation des termes complexes

Nous présenterons, ci-dessous, en détail les variations que nous avons rencontrées pour les structures de l'arabe et un tableau récapitulatif (voir Table 3). Ces dernières sont basées sur la typologie proposée par (Daille 05).

- Variations flexionnelles :

Ces variations regroupent les différentes formes fléchies possibles pour un terme complexe. Les flexions concernent plus particulièrement la mise au pluriel du deuxième nom dans la structure N1 N2 (1) et la définitude (2).

1. Nombre :

- tlwv AlmHyT (pollution de l'océan)
- tlwv AlmHyTAt (pollution des océans)

2. Définitude

Rappelons que la définitude est réalisée, elle aussi, par un morphème, préfixé (Al).

- Altlwv AlhwA'y (la pollution atmo-sphérique)
- tlwv hwA'y (pollution atmosphérique)

- Variations morphosyntaxiques et syntaxiques :

Les variations morphosyntaxiques affectent la structure interne du terme de base, et les mots qui le composent subissent des modifications relevant de la morphologie dérivationnelle.

1. Morphologie dérivationnelle : une variation conservant la synonymie est celle mettant en jeu un adjectif relationnel. Par exemple :

- b'r nfTy (puit pétrolier)
- b'r mn nfT (puit de pétrole)

Ce type d'adjectif existe dans beaucoup de langues mais il est plus au moins fréquent. L'arabe, comme l'anglais, tend à utiliser les adjectifs de relation avec une facilité que le français n'a pas encore égalée bien qu'il semble vouloir s'engager dans cette voie (Vinay & Darbelnet 66).

Les variantes syntaxiques modifient la structure interne de la structure du terme sans affecter les catégories grammaticales des mots pleins qui restent identiques.

Type	Sous-type	Terme	Variation
Modification	insertion	Altkwyn l ltrbp composition of the soil	Altkwyn Almstmr l ltrbp permanent composition of the soil
Modification	postposition	drjp AlHrArp degree of temperature	drjp AlHrArp AlEAlyp high degree of temperature
Coordination	expansion	tlwv Altrbp pollution of soil	tlwv AlmyAh w Altrbp pollution of soil and water
Coordination	tête	AlmkhAtr mn Altlwv Risks of pollution	AlmkhAtr w AlwqAyp mn Altlwv Risks and prevention of pollution

Table 3: Variation syntaxique

4.2.2 Partie statistique

La stratégie adoptée est l'extraction des séquences morphosyntaxiques comprenant deux unités lexicales. Ces séquences constituent une liste de candidats termes potentiels, qui sera soumise à diverses mesures statistiques. Ces mesures permettront de calculer le statut terminologique de la séquence rencontrée. Chaque mesure statistique repose sur un classement conceptuel des couples. Ce classement peut bien mettre plus en avant des expressions figées que des termes du domaine.

Notre objectif étant d'établir une liste des termes du domaine de l'environnement, il est essentiel de déterminer quelle mesure est la plus adaptée à l'extraction des termes. Nous avons décidé de comparer les valeurs obtenues pour chaque mesure à une liste de référence des termes du domaine. Cette évaluation s'effectue sur les 100 premiers couples extraits du corpus [AR-ENV]. Si le couple apparaît dans la liste de référence[23], il est considéré comme un bon candidat, sinon nous cherchons sa traduction compositionnelle en utilisant cette fois ci la banque terminologique Eurodicautom[24].

[23] www.fao.org/agrovoc

[24] http ://www.agris.be/fr/research/dico.html

4.3 Expérimentation

Nous avons évalué le système sur le corpus [AR-ENV]. Les termes extraits comprennent des termes complexes. Nous avons utilisé des valeurs relatives de précision (Kageura & Umino 96). Nous avons mesuré la performance du système sur les 100 termes initiaux de la liste des termes, classés par la IM^3 (Daille 94), T-score (Dunning 94), LLR (Dunning 94), FLR (Nakagawa & Mori 03). Les résultats généraux sont présentés dans le tableau ci dessous. Ainsi, dans l'ensemble des documents, un tri à l'aide de la LLR permet d'obtenir une bonne concentration des termes en tête de la liste de candidats termes. Les performances obtenues à l'aide de la LLR nous conduisent à conclure qu'il s'agit d'une mesure permettant de bien cerner le potentiel terminologique de certains des candidats termes recensés.

Type	P(%)
FLR	60%
T-score	57%
LLR	85%
IM^3	26%

Table 4: *Précision*

5 CONCLUSION

Dans cet article, nous avons présenté un outil pour l'extraction des termes, en combinant analyse linguistique avec les techniques de classement des candidats en fonction des différentes mesures statistiques. Nous avons défini les spécifications des termes complexes et leurs variations en langue arabe. Les résultats obtenus pour l'arabe sont semblables à celle des langues romaines (Ibekwe-SanJuan & Condamines 07). Il reste à évaluer cette approche sur différents domaines et application, comme l'emploi de notre approche pour le corpus amazigh.

References

[1] D. Bourigault. “An endogenous corpus-based method for structural noun phrase disambiguation”. In *Proceedings of the 6th Conference of the European Chapter of the Association of Computational Linguistics (EACL 93)*, Utrecht, 1993.

[2] F. Smadja. Xtract: An overview. Computers and the Humanities, 26:399-413, 1993.

[3] M. Diab, K. Hacioglu, et D. Jurafsky. Automatic tagging of arabic text: From raw text to base phrase chunks. In *Proceedings of HLT-NAACL 2004*, pages 149-152, Boston, 2004.

[4] B. Daille. Variations and application oriented terminology engineering. *International journal of theoretical and applied issues in specialized communication*, 11(1):181-197, 2005.

[5] J.P. Vinay and J. Darbelnet. Stylistique comparée du français et de l'anglais. Didier, Paris, 1966.

[6] K. Kageura, B. Umino. Methods of automatic term recognition: a review. Terminology, 3(2):259-289, 1996.

[7] B. Daille. *Approche mixte pour l'extraction de terminologie : statistiques lexicales et filtres linguistiques*. Unpublished PhD thesis, Université de Paris 7, France, 1994.

[8] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61-74, 1994.

[9] H. Nakagawa, T. Mori. Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2) :201-219, 2003.

[10] F. Ibekwe-SanJuan, M. T. Condamines, A. et Cabré Castellvi. Application-driven terminology engineering. Terminology, 2:1-17, 2007.

Conception et réalisation d'un système de recherche d'informations intégrant des connaissances sémantiques dans la phase d'indexation.

NAIMA TAZZITE(1), ABDELLAH YOUSFI(2), EL HOUSSINE BOUYAKHF(1)
(1) Faculté des sciences université Mohamed V Agdal, Rabat, Maroc
(2) Institut d'arabisation université Mohamed V Souissi, Rabat, Maroc

Résumé : La plupart des systèmes de recherche d'information (SRI) actuellement en service utilisent de simples mots clés pour indexer et rechercher une information dans un document. Ces mots clés sont des mots isolés, la plupart du temps réduits à une simple racine ou bien standardisés à l'aide d'une lemmatisation. Dans ces systèmes, un document est indexé par un ensemble de ces mots clés en négligeant la relation sémantique entre les mots. Dans cet article nous proposons d'intégrer l'un des niveaux de la sémantique (terminologie) dans un moteur de recherche pour la langue Arabe. Cette introduction est réalisée dans la phase d'indexation en utilisant des dictionnaires terminologiques des mots à rechercher.

Mots –clefs- keywords :

Sémantique, Recherche documentaire, dictionnaire terminologique, Okapi, Croft, Harman.

1. Introduction

La recherche documentaire (R.D) est un domaine très important dans le TAL, elle permet de faciliter l'accès à une information dans un flux de documents dans un temps réel. Plusieurs travaux ont été élaborés afin de mettre en œuvre un système de recherche d'information assez complet.

La plupart de ces travaux, en particulier pour la langue arabe, sont confrontés à des problèmes d'établir une correspondance entre l'information recherchée et l'ensemble des documents d'une collection. Parmi ces problèmes, il y'a les formulations différentes d'un même concept : Un document pertinent peut contenir des termes sémantiquement proches de ceux de la requête mais toutefois différents (synonymes, hyperonyme, terme ayant une forme morphologique différente, terminologie, etc.). Ce phénomène provoque une baisse du rappel de ces systèmes qui ne peuvent proposer à l'utilisateur certains documents pourtant intéressants. A ce problème vient s'ajouter celui de la polysémie des mots. L'ambiguïté qui en

découle est à l'origine d'une baisse de précision des systèmes puisqu'elle entraîne potentiellement la récupération de documents non pertinents.

Plusieurs travaux de recherche ont été réalisés pour voir l'influence de la sémantique sur les systèmes de R.D :

- El-Bèze a montré qu'il est possible d'utiliser les modèles de Markov cachés pour effectuer un étiquetage sémantique [1].
- Resnik a utilisé les liens qui existent dans WordNet entre les noms qui apparaissent dans une certaine fenêtre d'un texte pour déterminer le sens de ces noms [2].
- Towell et Voorhees ont montré que la polysémie a un effet de diminution de la précision des systèmes de R.D [3].
- Smeaton, Quigley et Gonzalo ont montré que la désambiguïsation sémantique et l'utilisation d'un thesaurus comme WordNet permettent d'augmenter les performances des systèmes de R.D [4], [5].
- Loupy a étudié l'influence de la sémantique sur les performances des systèmes de recherche documentaire [6].

Dans cet article nous nous intéressons à l'introduction d'un autre niveau de la sémantique, la terminologie des mots, dans un moteur de recherche pour la langue Arabe. Dans un travail précédent [7], [8] , nous avons traité ce problème en se basant sur l'approche d'extension du requête par des termes appartenant aux dictionnaires terminologiques du mot à rechercher. Les résultats obtenus sont très importants au niveau de la précision et du rappel, le seul inconvénient est que cette approche nécessite un temps élevé pour son exécution pendant la recherche.

Afin de remédier à ce problème, nous proposons dans cet article une autre approche qui s'appuie sur l'introduction de la sémantique dans la phase de l'indexation.

2. L'intérêt de la sémantique dans les moteurs de recherche

L'intérêt de l'introduction de la sémantique dans les moteurs de recherche est assez clair, elle permet de retrouver des résultats plus pertinents que les résultats retournés par une recherche qui ne prend pas en compte la sémantique. Pour voir ceci plus clairement, on donne l'exemple suivant :

2
يتكون الطب من عدة مجالات مهمة، نذكر منها: علم التشريح، علم الأوبئة وانتقال الفيروسات.
كما أنه يختص بدراسة الأمراض ومحاولة إيجاد أدوية فعالة.

1
يعتبر الطب من المجالات المهمة التي تحظى باهتمام الدولة، لكن رغم هذا الاهتمام يظل الطب يحتاج إلى المزيد من الدعم المادي.

Pour les moteurs de recherche actuels, si on demande la requête suivante : « الطب» en s'appuyant seulement sur la fréquence d'apparition de ce mot, le document 1 est plus pertinent que le document 2, or c'est l'inverse car le document 2 est plus riche de termes appartenant au domaine "الطب " (الأمراض، التشريح ، الأوبئة ، الفيروسات). Comme le montre cet exemple, il est très intéressant et nécessaire de développer un moteur de recherche qui prend en compte le niveau terminologique des mots pendant la phase de la recherche.

3. Rappel sur le moteur de recherche par extension de requête

Pour introduire la sémantique dans un moteur de recherche pour la langue arabe, nous avons construit des petits dictionnaires terminologiques de chaque terme (Tableau 1). Ainsi un terme $x_i \in V$ est associé au dictionnaire :

$$Dic(x_i) = \{ t_{i_1}, t_{i_2}, \ldots, t_{i_k} \}$$

Dans un travail précédent, cette introduction a été effectuée dans la phase de la recherche [7], c'est-à-dire avec l'extension de la requête par des termes appartenant au dictionnaire terminologique du mot recherché. Les résultats obtenus ont été très encourageants, et les pages retournées pour des requêtes données ont été très intéressantes.

L'inconvénient majeur de cette approche est le temps d'exécution de la phase de la recherche qui est supérieur au celui du moteur classique (sans extension) et ce fait résulte de l'extension de la requête avec d'autres termes. Pour remédier à ce problème nous proposons dans cet article d'introduire la sémantique dans la phase de l'indexation.

4. Introduction de la sémantique dans la phase de l'indexation

Dans notre nouvelle approche, on procède au calcul d'un poids d'un terme x_i en prenant en compte les poids de tous les éléments appartenant au dictionnaire $Dic(x_i)$ dans tous les documents.

On note par :

$$poid((x_i, t_{i_1}{}^{\theta}, t_{i_2}{}^{\theta}, ..., t_{i_k}{}^{\theta}), d) = \frac{F(x_i) + \theta \times \sum_{j=1}^{k} F(t_{i_j}, d)}{P_C(x_i) + \sum_{j=1}^{k} P_C(t_{i_j})} \quad (1)$$

Le poids du terme x_i avec les termes appartenant au dictionnaire d'ontologie $Dic(x_i)$ associé à x_i.

سياسة	قانون	اقتصاد	الفن	التاريخ	الطب
أحزاب	محامون	اقتصاديون	خزف	تراث	أدوية
نقابات	قوانين	تجارة	رسم	تقاليد	أمراض
انتخابات	قضاء	أرباح	شعر	فكر	مستشفايات
حكومة	قانونيون	خصخصة	أدب	مجتمعات	التشريح
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
سياسيون		ضرائب	نحت	مكتبات	عقاقير

Tableau 1: Un extrait des dictionnaires utilisés dans la phase

- θ est l'importance donnée aux termes $\{t_{i_1}, t_{i_2}, ..., t_{i_k}\}$ par rapport à x_i .

Dans ce travail, nous avons donné la même importance aux termes appartenant au dictionnaire $Dic(x_i)$ par rapport au terme x_i.

Afin d'évaluer notre modèle, nous avons utilisé les trois mesures de similarités : Harman [9], Croft et okapi.

5 Protocole expérimental

L'expérimentation a été effectuée sur un ensemble de documents composé d'un ensemble d'articles en arabe, le tableau 2 montre les caractéristiques de cette collection.

Nombre de documents dans la collection	Nombre de termes dans la collection	Taille moyenne du document
1000	116 600	150

Tableau 2 : Description de la collection de documents utilisée

L'expérimentation a été effectuée sur un nombre de requêtes numérotées à partir de 1.

5.1 Construction des dictionnaires d'ontologie :

Pour évaluer notre approche, nous avons construit à l'aide d'un expert en terminologie 20 dictionnaires terminologiques associés à 20 domaines. Chaque dictionnaire est constitué d'un certain nombre de termes qui caractérisent chaque domaine (tableau 1).

L'indexation des documents est effectuée par l'approche déjà citée dans le paragraphe précédent. Le processus d'indexation produit pour chaque document une liste de termes pondérés par la fonction *poids (.)*. Le processus de recherche consiste à associer à chaque document une mesure de similarité entre ce document et la requête recherchée.

6. Evaluation des performances de système (Avec analyse sémantique)

L'évaluation de cette nouvelle approche est effectuée par l'utilisation des trois mesures de similarité : Harman, Croft, et Okapi, et pour des différentes valeurs de θ. La mesure de l'efficacité du système est effectuée par les trois paramètres suivants : la précision, le rappel, et la courbe rappel-précision

6.1 Détermination du taux de mixité

Plusieurs tests ont été effectués sur les trois mesures de Harman, Croft et Okapi en faisant varier le taux de mixité θ. Les meilleurs résultats sont obtenus pour la mesure d'Okapi et pour $\theta=0,1$.

6.2 Comparaison des deux systèmes (sans et avec analyse sémantique)

L'apport de l'intégration des connaissances sémantiques dans la phase d'indexation est évalué au moyen de la comparaison entre les résultats obtenus sans et avec analyse sémantique.

Pour évaluer ces deux systèmes on a fait appel aux mesures suivantes : la précision, le rappel sur les n premiers documents trouvés (P(n) et R(n)) et la précision moyenne interpolée sur 11 point (IAP).
Le tableau 3 et la figure 1 indiquent les résultats obtenus en utilisant les différentes mesures précitées pour calculer les scores des documents. On remarque que le système avec analyse sémantique est globalement plus performant que le système utilisant la requête originale. Les meilleurs résultats sont obtenus avec la mesure d'Okapi. Elle permet notamment d'obtenir une amélioration absolue de 17% de la précision par rapport au système classique ($\theta=0$) pour les 15 premiers documents.

Il est également intéressant de noter que l'amélioration des résultats est distribuée sur tous les seuils des mesures (allant de 10 à 1000 documents). Cela signifie que l'amélioration n'est pas due à une simple réorganisation des documents en tête de liste mais bien à la découverte des documents pertinents qui n'auraient pas été ramenée par le SRI classique.

On voit donc que la mesure d'Okapi permet d'obtenir de meilleurs résultats pour tous les premiers documents rapportés, comparée aux mesures de Croft et de Harman. Tandis que la mesure de Croft obtient de meilleurs résultats pour un rappel élevé (mais la différence n'est pas très importante, que ce soit dans l'absolu ou dans le nombre de requêtes concernées).

Conclusion

L'étude des résultats précédents montre que l'utilisation de la mesure de Croft permet de retourner plus de documents pertinents que les deux autres systèmes. En revanche, Okapi est plus efficace pour la précision dans les premiers documents rapportés. Ce qui est l'objectif principal. Il semble donc préférable d'utiliser Okapi. Mais on peut aussi envisager une combinaison des deux mesures. En effet, il est tout à fait possible d'utiliser Croft pour récupérer les 1000 documents demandés et les réordonner ensuite avec Okapi.

Après ces expériences, la mesure d'Okapi semble donc la plus adapté à nos besoins.

6.3 Comparaison des deux systèmes (introduction de la connaissance sémantique dans la phase de recherche et dans la phase d'indexation)

La base terminologique décrite dans la partie précédente a été utilisée pour comparer les deux systèmes suivants : le premier sur lequel nous avons introduit les connaissances sémantiques dans la phase de recherche, et le second où nous avons intégré des connaissances sémantiques dans la phase d'indexation. Les résultats que nous avons obtenus sur cette collection (voir Figures 2 et Figure 3) montrent une augmentation systématique des performances mesurées par la précision moyenne sur les 11points de rappel et de la courbe rappel/précision.

L'introduction des connaissances sémantiques dans la phase de l'indexation a abouti d'une part à une réduction du temps moyen d'exécution (il est passé de 0.93s dans la phase de la recherche à 0.18s dans la phase de l'indexation) et d'autre part à une très nette amélioration de la précision, puisque la précision moyenne est de 70.35% (Tableau 4).

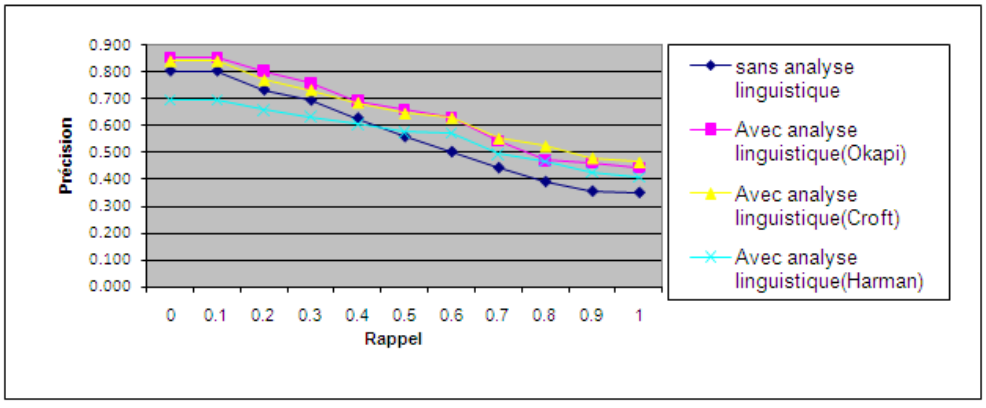
6.4 Interface graphique

Une première version de notre système a été développée en Java avec interface graphique (figure 4).

Suite à la formulation d'une requête de l'utilisateur, le moteur de recherche retourne sous forme d'une page HTML un ensemble de résultats constitués entre autres de l'URL du document identifié de son titre et d'extrait du document. La figure 4 montre l'exécution de la requête " الفن ".

	SANS Extension	Avec extension Okapi	Augmentatin (Okapi)	Avec extension Croft	Augmentatin (Croft)	Avec extension Harman	Augmentatin (Harman)
N=5	88.00%	94.00%	6.00%	94.00%	6.00%	82.00%	-6.00%
N=10	74.00%	88.00%	14.00%	87.00%	13.00%	69.00%	-5.00%
N=15	61.33%	78.00%	16.67%	76.00%	14.67%	60.00%	-1.33%
N=20	56.50%	71.00%	14.50%	69.00%	12.50%	55.50%	-1.00%
N=100	29.10%	46.40%	17.30%	46.50%	17.40%	41.30%	12.20%
N=1000	3.40%	5.21%	1.81%	5.70%	2.30%	4.72%	1.32%
Av_prec	39.30%	50.79%	11.49%	50.47%	11.17%	41.70%	2.40%
R=0.1	80.30%	85.40%	5.10%	84.00%	3.70%	69.60%	-10.70%
R=0.2	73.20%	80.27%	7.07%	77.00%	3.80%	65.80%	-7.40%
R=0.3	69.40%	75.73%	6.33%	73.30%	3.90%	63.40%	-6.00%
R=0.5	55.80%	65.97%	10.17%	64.60%	8.80%	57.73%	1.93%
R=1	35.10%	44.38%	9.28%	46.50%	11.40%	40.88%	5.78%
R(5)	7.70%	9.85%	2.15%	9.68%	1.98%	7.20%	-0.50%
R(10)	10.10%	17.62%	7.52%	17.17%	7.07%	8.69%	-1.41%
R(15)	11.70%	20.86%	9.15%	20.08%	8.37%	9.77%	-1.93%
R(20)	13.70%	23.09%	9.39%	22.24%	8.54%	11.01%	-2.69%
R(100)	25.50%	47.84%	22.34%	47.90%	22.40%	19.95%	-5.55%
R(1000)	27.90%	50.81%	22.91%	52.94%	25.04%	19.95%	-7.95%

Tableau 3: apport de la connaissance sémantique à la recherche documentaire



	Précision pour un rappel 0.10	Précision pour un rappel 0.2	Précision des 5 premiers documents	Précision des 10 premiers documents	Précision des 15 premiers documents	Précision moyenne	Temps Moyen d'exécution
Sans analyse sémantique	61.88%	53.75%	88%	74%	61.33%	41.70%	0.18s
Introduction des connaissances sémantiques dans la phase de recherche	77.75%	73.00%	89.23%	84.62%	76%	56.13%	0.93s
Introduction des connaissances sémantiques dans la phase de l'indexation	85.40	80.27	94%	88%	78%	70.35%	0.18s

Tableau4 : Intégration des connaissances sémantiques dans la phase d'indexation et la phase de recherche

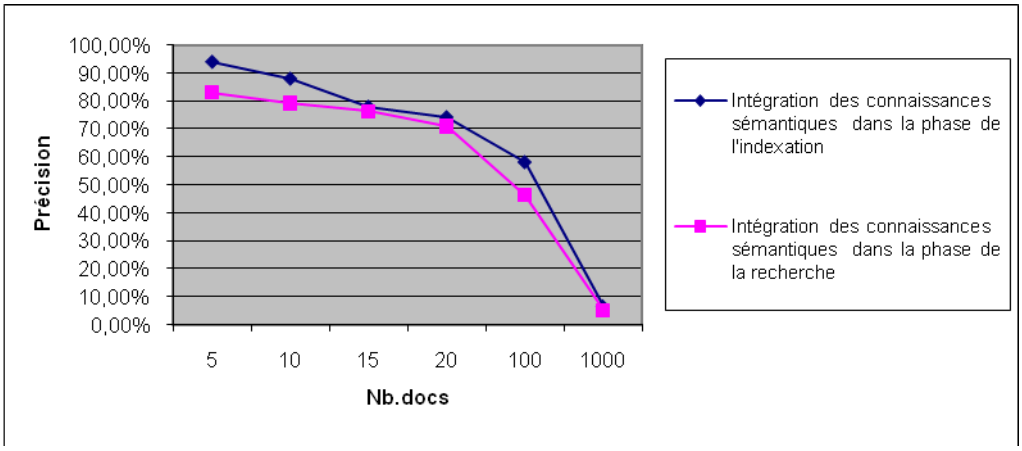


Figure2 : Intégration des connaissances sémantiques dans la phase d'indexation et la phase de recherche (courbe Précision /Nb.docs)

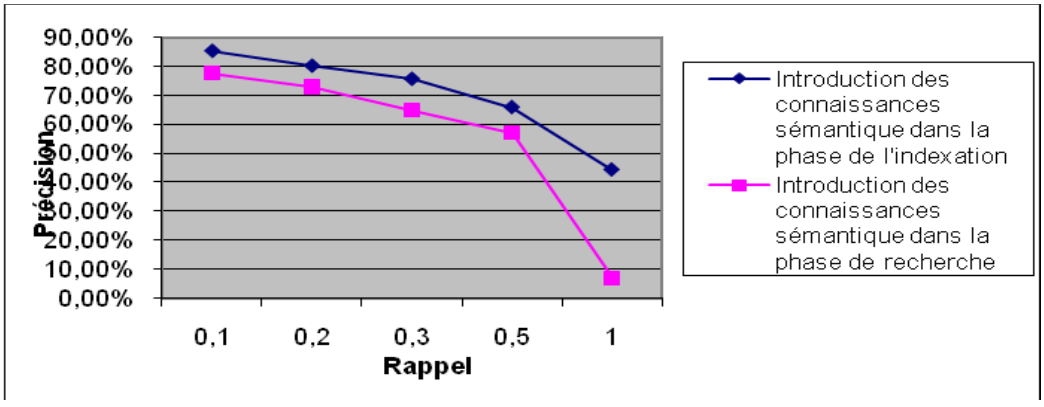


Figure3 : Intégration des connaissances sémantiques dans la phase d'indexation et la phase de recherche (courbe rappel- précision)

7 Conclusion et perspective

Nous avons présenté dans ce travail un moteur de recherche pour la langue arabe qui prend en compte l'un des niveaux de la sémantique (la terminologie des mots). Cette introduction est effectuée dans la phase de l'indexation. L'évaluation de notre système est réalisée par les trois mesures de similarités suivantes : Harman, Croft et Okapi. Les meilleurs résultats sont obtenus par la mesure d'Okapi et pour un taux de mixité égale à 0,1. L'introduction de la sémantique dans la phase d'indexation a considérablement réduit le temps de la recherche par rapport à l'ancienne méthode (introduction de la sémantique dans la phase de recherche) [7]. Ce temps est réduit de 75%. La performance de ce système en termes de précision et du rappel est un peu améliorée par rapport à l'ancienne méthode [7].

Dans le prochain travail, nous allons adopter cette approche à la langue amazighe. Ce travail consiste à mettre en place un système de recherche d'information en langue amazighe. Ainsi, nous allons d'une part présenter les spécificités de la langue amazighe du point de vue de la recherche d'information et d'autre part construire un corpus qui sera constitué de plusieurs documents amazighs. Nous allons aussi construire une dizaine de dictionnaires où chaque dictionnaire désigne un domaine spécifique.
Et comme vous avez déjà remarqué dans ce travail, nous avons supposé que tous les termes des dictionnaires ont la même pertinence. Dans le prochain travail, il est très important de calculer l'intérêt spécifique pour chaque terme d'un dictionnaire pour la langue amazighe. Et pour atteindre ce but nous allons utiliser des techniques de Data Mining. Ces techniques permettent d'extraire automatiquement des termes à partir des textes écrits en amazighe. Ces outils désignés pour l'acquisition des termes sont assez bien décrits pour l'anglais et pour le français.
Pour l'amazighe, nous allons adopter l'une des techniques de Data Mining appelée "Règles d'association" déjà utilisée pour les deux autres langues françaises [10] et Arabes [11].

Références

[1] M. El-Bèze; Les modèles de langage probabilistes : quelques domaines d'application; Habilitaion à Diriger desRecherches, LIPN (Université de Paris Nord) ; Paris ; 1993.

[2] P. Resnik ; Disambiguating noun grouping with respect to WordNet senses ;Actes de 3rd Workshop on Very Large Corpora ; M.I.T. ; 1995.

[3] G. Towel, E. Voorhees; Disambiguating highly ambiguous words; Special Issue on WSD, Computational Linguistics, Vol.24, No. 1; pp. 125-145; 1998.

[4] A. Smeaton, I. Quigley; Experiments on using semantic distances between words in image caption retrieval; Actes de SIGIR'96; 1996.

[5] J. Gonzalo, F. Verdejo, I. Chugur, J.Cigarran ; Indexing with WordNet synsets can improve text retrieval ; Acts the 17th International Conference On Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics - Workshop on Usage of WordNet for Natural Language Processing ;1998.

[6] Claude D. L., "Evaluation de l'apport de connaissances linguistique en désambiguisation sémantique et recherche documentaire". Thèse de doctorat, 2000.

[7] " Semantic Internet search engine with focus on Arabic language". Saudi computer Society in Riyadh from 25 to 28 March 2007.

[8] "Conception et réalisation d'un moteur de recherche arabe sur Internet". Colloque JETELA 2006, sous le thème: Traitement Automatique de la langue Arabe organisé par l'institut d'Etude et de Recherche pour l'Arabisation à Rabat du 5 à 7 juin 2006.

[9] Harman D., (1986) "An experimental study of factors important in document ranking". Actes de ACM, Conference on Research and development in Information Retrieval; Piste, Italie.

[10] Mohamed Hatem HADDAD: "Extraction et Impact des connaissances sur les performances des Systèmes de Recherche d'Information", thèse, 2002.

[11] "Integration of semantics in an Arabic search engine by using the association rules". est en cours d 'evaluation pour le journal " The International Arab Journal of Information Technology .IAJIT".

Interface de la recherche

- موقع وزارة التربية الوطنية والتعليم العالي وتكوين الأطر والبحث العلمي، 2005. http://genie.men.gov.ma
- وزارة التربية الوطنية والتعليم العالي وتكوين الأطر والبحث العلمي، المذكرة رقم 30 حول تنظيم تدريس اللغة الأمازيغية وتكوين أساتذتها، 2006.
- Principles of Teaching, Bloomsburg University, 2003. Url: http://teacherworld.com/potdale.html
- Captures images d'exemples de supports didactiques multimédia amazighes, développés par Abdellatif Hssaini.

- ❖ التكوين والتكوين المستمر لهيئة التدريس في مجال توظيف تكنولوجيا الإعلام والاتصال في الدرس اللغوي الأمازيغي، وتعزيز دور الإدارة والمراقبة التربويتين في التأطير في هذا المجال.
- ❖ إعداد بيئات افتراضية ومصوغات للتعليم والتكوين في اللغة والثقافة الأمازيغيتين، لفائدة التلاميذ واطر التربية الوطنية.
- ❖ تشجيع البحث العلمي وخلق شعب متخصصة في التكنولوجيا التربوية لتعليم اللغات في الجامعات والمعاهد العليا.
- ❖ تجهيز المؤسسات التعليمية بما يكفي من العتاد المعلوماتي والقاعات متعددة الوسائط، مع ربط بشبكة المعلومات العالمية.
- ❖ تشجيع مبادرات تطوير ملحقات البرامج بالأمازيغية Plugins ، وهذا نظرا لسهولة نشرها ودمجها مع برمجيات مختلفة، من قبيل برمجيات معالجات النصوص والمعاجم الالكترونية والكتب الرقمية وغيرها.
- ❖ تشجيع مبادرات تطوير البرمجيات الأمازيغية ذات القن المفتوح Open Source أو ذات إجازة Licence GPL.
- ❖ توعية إباء وأولياء التلاميذ بالأهمية التربوية والمعرفية لتكنولوجيا المعلومات والتواصل.
- ❖ تشجيع مواقع الانترنيت والمدونات التربوية الأمازيغية Blogs éducatifs، والموسوعات المجانية Les Wikis، بتوفير ملقمات تسكين مجانية Serveurs d'hébergement gratuits ، تكون في متناول المهتمين.

المصادر

- موقع وزارة التربية الوطنية والتعليم العالي وتكوين الاطر والبحث العلمي، 1999. http://www.men.gov.ma/dajc
- بشير عبد الرحيم الكلوب، التكنولوجيا في عملية التعلم والتعليم، دار الشروق للنشر والتوزيع، الطبعة الثانية، عمان الأردن، 1993، ص ص 20.36.
- رشيد لبيب، الأسس العامة للتدريس، دار النهضة العربية، بيروت، 1983، ص119.
- محمد زياد حمدان، وسائل تكنولوجيا التعليم، الطبعة الثانية، دار التربية الحديثة، 1980، ص21.
- حسين حمدي الطوبجي، وسائل الاتصال والتكنولوجيا، دار القلم، الكويت 1987، ص24.
- A De Corte et coll : Les fondements de l'action didactique, traduit du Néerlandais par V.V. Culsen, Paris, Editions Universitaires, 1990, P187.
- احمد حامد منصور، تكنولوجيا التعليم وتنمية القدرة على التفكير الابتكاري، منشورات ذات السلاسل، الطبعة الأولى، 1986، ص30.
- J. De Rosnay, Le macroscope, vers une vision globale, Edit Seuil (Coll Le point), 1977 , P261.

- ❖ غياب البنيات والمصوغات الخاصة بالتعليم والتكوين عن بعد في اللغة والثقافة الأمازيغيتين، خصوصا تلك التي توافق معيار "سكورم Scorm".
- ❖ قلة الباحثين والأطر التربوية المغربية المتمكنة من تقنيات تعليم ومعالجة اللغات بواسطة الحاسوب.
- ❖ نقص في تجهيز المؤسسات التعليمية ومؤسسات تكوين الأطر بالعتاد المعلوماتي، والربط بشبكة الانترنيت.
- ❖ قصور في دعم وتشجيع مبادرات التربويين الرائدة في مجال تطوير المحتويات الأمازيغية متعددة الوسائط، مما لا يساعد على تحيين الحاجيات وتحديد الأولويات في هذا المجال.
- ❖ محدودية هذه التكنولوجيا في تعليم بعض المهارات اللغوية، من قبيل التعبير الكتابي الإنشائي.
- ❖ المقاومة التي يبديها بعض المدرسين اتجاه العمل بالتقنيات الحديثة، وتشبثهم بالممارسة الصفية التقليدية في تطبيق المنهاج الدراسي للغة الأمازيغية، وهو شيء في نظرنا يرتبط إما بتصلب العقليات أو بهزالة الإمكانات المادية أو بغياب التكوين المستمر.
- ❖ قلة وعي آباء وأولياء التلاميذ بأهمية التكنولوجيا الحديثة في الرفع من حصيلة أبنائهم.

4. خاتمة واقتراحات

من خلال ما سبق يمكن أن نخلص إلى أن تكنولوجيا الإعلام والاتصال التربوية قد عرفت تطورا ملحوظا، تحولت عبره إلى عنصر أساسي في النسق التعليمي، فأصبحت لها قيمة حقيقية في المجال الديداكتيكي عموما وفي مجال تعليم وتعلم اللغات خاصة. غير أن واقع الممارسة التربوية بالمؤسسات التعليمية المغربية، وخصوصا خلال حصص مكون اللغة الأمازيغية، لا يعكس هذا التطور بما فيه الكفاية، مما يجعل المدرسة المغربية متخلفة عن مثيلتها من فرنسا أو كندا أو حتى من دول عربية كتونس والأردن، في مجال تعليم اللغات باستخدام التكنولوجيا الحديثة للإعلام والاتصال.

من ثم وجب علينا إن نقترح بعض الاقتراحات، مما سيسمح في نهاية المطاف بإدماج حقيقي وفعال لتكنولوجيا الإعلام والاتصال التربوية في الدرس اللغوي الأمازيغي، ومنها:

- ❖ تجاوز مشكل ندرة الموارد الرقمية التربوية الأمازيغية ب:
  - إشراك الأكاديميات الجهوية للتربية والتكوين والنيابات الإقليمية في إنتاج المضامين الرقمية بالأمازيغية، على اعتبار أن 15 بالمائة من المضامين التعليمية الوطنية يجب أن تكون ذات صبغة محلية، حسب الميثاق الوطني للتربية والتكوين.
  - العمل على إبرام شراكات مع القطاع الخاص، على اعتبار أن لهذا القطاع من الإمكانيات ما يؤهله لتطوير محتويات ديداكتيكية أمازيغية ملائمة للمنهاج الدراسي الأمازيغي .
- ❖ تشجيع مدرسي اللغة الأمازيغية على توظيف تكنولوجيا الإعلام والاتصال التربوية، منحهم هامشا من الحرية في استخدام هذه التكنولوجيا في ممارستهم البيداغوجية.

- ⵔⵉⵙⵏ ⵓⴰ ⵉⵙⵙⴽⴰⵙ ⵓⵉⵇⵉⵎ;
- ⵉⵙⵙⵔⵣⵖⵏ ⵓⵉⵇⵉⵎ.

ⵜⵉⵣⵓⵎⵉⵍⵉ	ⵜⵉⵣⵏⵓⵏ ⵏ ⵖⵙⵏⵎⵓⴰ ⴰ ⵖⵏⵎⵓⴰ	ⵉⵙⵔⵔⵖⵙⵓⵏ
1- ⵜⵓⵙⵜ	- ⵓⴰ ⵉⵙⵙⵎⵜⵉ ⵖⵙⵏⵎⵓⴰ ⵓⵙⵏⵜⵏ ⴰ ⵜⵍⵏⵔⵉⵎⵜ ⵜⵓⵎⵓⵜⵖⵜ. - ⵓⴰ ⵉⵎⵜⵉ ⵖⵙⵏⵎⵓⴰ ⵓⵙⵏⵜⵏ ⴰ ⵜⵍⵏⵔⵉⵎⵜ ⵜⵓⵎⵓⵜⵖⵜ. - ⵓⴰ ⵉⵙⵓⵓ ⵖⵙⵏⵎⵓⴰ ⵚⵜ ⵉⵙⵕⵙⵙⵜⵏ ⵉⵏⵎⵓⴰⵍ.	- ⵓⴰⵏⵉⵙⵙ ⵏ ⵖⵏⵎⵓⴰ
2 - ⵜⵉⵢⵓⵙ ⵙ ⵖⵜⵙⵙⵉ	- ⵓⴰ ⵉⵎⵎⵎ ⵖⵏⵎⵓⴰ ⵓⵉⵇⵉⵎ ; - ⵓⴰ ⵉⵙⵍⵉⵏⵎ ⵠ ⵖⵉⵇⵉⵎ ; - ⵓⴰ ⵔⵓⵎⵜⵉⵠⵖⵙⵓⵉⵍⵉⵏ ⵜⵉⵠⵓⵜⵓⵉⵏ ⵠ ⵖⵉⵇⵉⵎ.	- ⵜⵉⵜⵏⵍⵉⵜ ⴰ ⵖⵎⵙⵙⴽⴰⵙ ⵏ ⵜⴰⵓⵏ
3- ⵓⵙⵜⵓⵖ ⵏ ⵖⵉⵇⵉⵎ ⵜⵓⵙⴰⴰⵓⵜ ⵙ ⵜⵙⴰⴰⵓⵜ	- ⵓⴰ ⵔⵓⵍⵙ ⵖⵙⵏⵎⵓⴰ ⵉ ⵉⵏⵎⵓⴰⵍ ⵓⴰ ⵓⵜⵉⵍ ⵉⵍⵖⵓⵔ. - ⵓⴰ ⵉⵙⵍⵙⵓ ⵖⵙⵏⵎⵓⴰ ⵠⵜ ⵉⵍⵖⵓⵔ ⵏ ⵜⵙⴰⴰⵓⵜ ⵙ ⵔⵓⵜ ; - ⵓⴰ ⵔⵓⵙⵉ ⵠⵜ ⵜⵉⵜⵍⵉⵜ / ⵠⵜ ⵖⵙⵙⵙⴽⴰⵙ ; - ⵓⴰ ⵉⵙⵍⵙⵓ ⵠⵜ ⵔⵓ ⵖⵙ ⵉⵙⵙⵉⵍ ; - ⵓⴰ ⵉⵙⵓⵓ ⵠⵜ ⵉⵙⵍⵙⵓⵉⵜⵏ ; - ⵓⴰ ⵔⵓⵖⵙ ⵠ ⵖⵙⵙⵜⵜⵓⵖⵙ ⵏ ⵜⵙⴰⴰⵓⵜ.	- ⵉⵍⵏⵓⵜⵏ ⴰ ⵖⵖⵙⵙⴽⴰⵙ ⵏ ⵜⴰⴰⵓⵜ
4- ⵓⵙⵠⵎⵖⵏ ⵏ ⵖⵉⵇⵉⵎ	- ⵓⴰ ⵉⵙⵍⵙⵓ ⵖⵙⵏⵎⵓⴰ (ⵏⵛⴰ ⵓⵖⵙⵙⴽⴰⵙ) ⵠⵜ : - ⵖⵓⵏⵉ ? - ⵖⵉⵏⵎⵉ/ⵖⵓⵏⵠⵖ ? - ⵖⵓ++ⵓ ? ⵖⵓ ⵉⵎⵖⵙ ?ⵖⵓⵠ/ⵖⵓⵔⵎⵎⵉ ? - ⵓⴰ ⵔⵉⵍⵙ ⵉ ⵉⵏⵎⵓⴰⵍ ⵓⴰ ⵙⵙⵠⵎⵖⵏ ⵓⵉⵇⵉⵎ ; - ⵓⴰ ⵉⵙⵓⵓ ⵠⵜ ⵉⵙⵍⵙⵓⵉⵜⵏ ; - ⵓⴰ ⵉⵙⵙⵙⵓ ⵜⵉⵙⵙⵓⵏ ⵉ ⵙⵉⵇⵉⵎ.	- ⵜⵉⵏⵓⵏ / ⵓⵖⵙⵙⴽⴰⵙ

5.1.3 بعض معيقات إدماج تكنولوجيا الإعلام والاتصال التربوية في الدرس اللغوي الأمازيغي

يمكن أن نلخص أهم معيقات إدماج تكنولوجيا الإعلام والاتصال التربوية في دروس اللغة الأمازيغية في النقاط التالية:

- ندرة البرمجيات الرقمية الخاصة بتعليم اللغة الأمازيغية.
- نقص كبير في مواقع الانترنيت التي تعنى بتعليم اللغة الأمازيغية، وحتى وان وجدت فهي تعتمد على الارتجالية في تقديم المضامين التعليمية.

يقوم المتعلم بتقويم تعلماته باعتماد تقنيات مختلفة، مع إمكانية اللجوء إلى التغذية الراجعة ودعم وتعزيز المكتسبات في أي وقت.

تتيح تقنية الفيديو إمكانية محاكاة مواقف تواصلية حقيقية. يتعلق الأمر هنا بدعامة ديداكتيكية لدرس التعبير الشفهي.

4.3.3. نموذج جذاذة إعداد لدرس لغوي أمازيغي تدمج تكنولوجيا الإعلام والاتصال التربوية

هذا نموذج لجذاذة إعداد لدرس في اللغة الأمازيغية، تدمج تكنولوجيا الإعلام والاتصال التربوية في مادة القراءة بالمستوى الثالث ابتدائي :

ⴰⵎⴰⵣⵡⵍ ⵏ ⵜⵓⵔⵉⵜ ⵏ ⵜⵡⵓⵣ

2 – ⵣⵓⵍⵉⵡ ⵏ ⵜⵡⵓⵣ

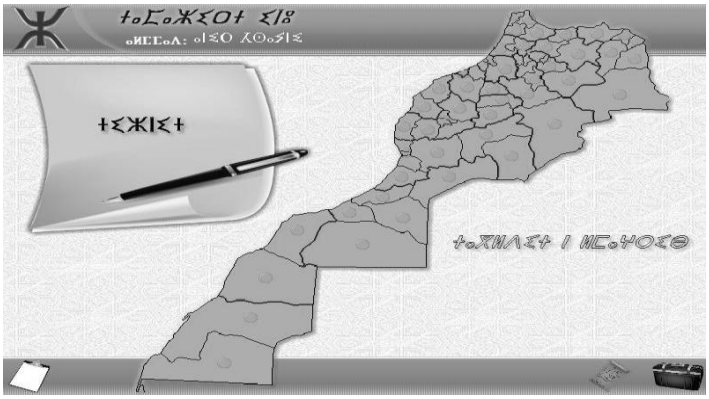
ⵜⴰⵙⴷⴰⵜ : ..... ⵜⵙⵓⵎⵜ : .....

ⴰⵙⵓⵙⵓⵙ ⵏ ⵙⵎⵙⴰⵡⴰⵏ : ..... ⴰⵣⵙⴰ : .....

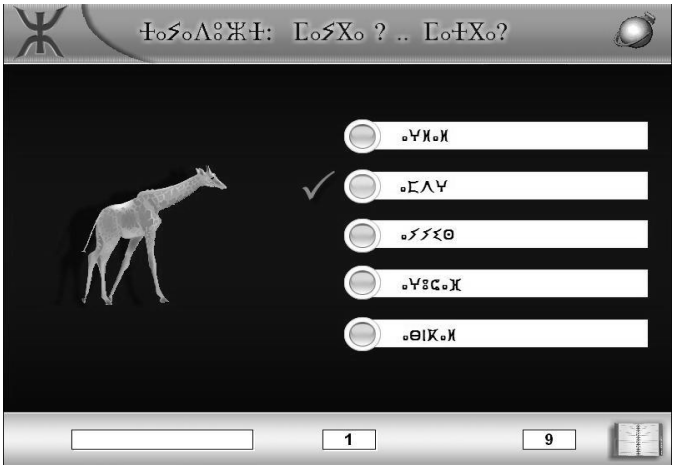
ⴰⵣⵓⵣⵉⵏ ⵏ ⵙⵏⵇⵣⵎ :.....

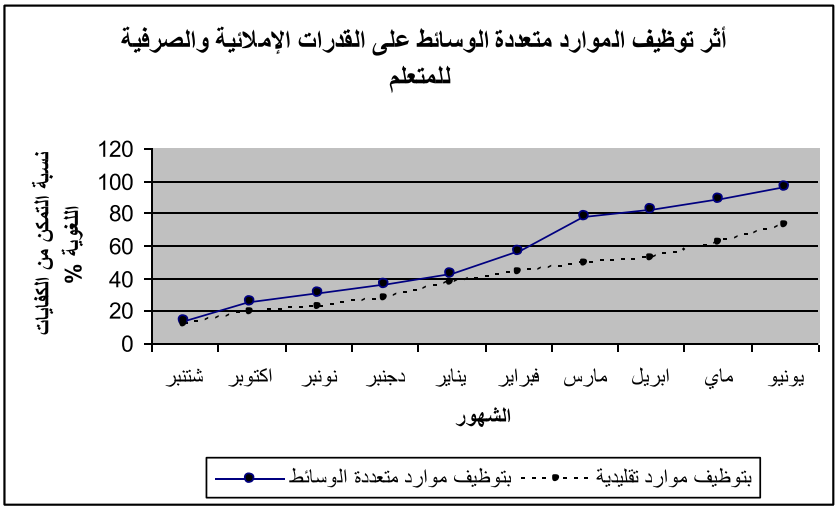
ⵅ ⵜⵣⵓⵎⵜ ⴰⴰ, ⴰⴰ ⵙⵣⵎⵎⴰⵙ ⵙⵏⵎⴰⴰ ⴰⴰ :

- ⵣⵓⵎⵙ ⵜⵣⵅⵙⵓⵣⵓⵣⵍ ⴰ ⵜⵙⴰⴰⴰⵙⵓⵣⵍ ⴰ ⵙⵏⵇⵣⵎ ;
- ⵙⵙⵙⵙⵣⵢ ⵜⴰⵓⵍⵅⵣⵎⵜ ⵜⴰⵎⴰⵜⵙⵜ ;
- ⵣⵙⵙⵙⵣⵢ ⵜⵣⵓⵍⵅⵣⵎⵣⵍ



بواسطة المعجم المصور ومقاطع الفيديو يكتسب المتعلم معجما أمازيغيا وظيفي





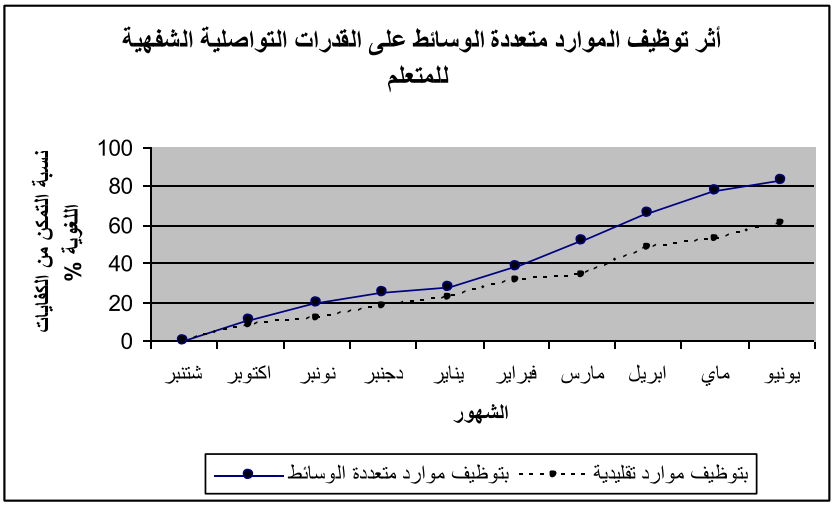
من الواضح أن الحاجة لتوظيف الإمكانيات التي تتيحها تكنولوجيا المعلومات والتواصل في الممارسة الصفية – وبالخصوص في الدرس اللغوي الأمازيغي– له مبرراته التربوية، والتي تزكيها النتائج الايجابية التي حققها العمل بها.

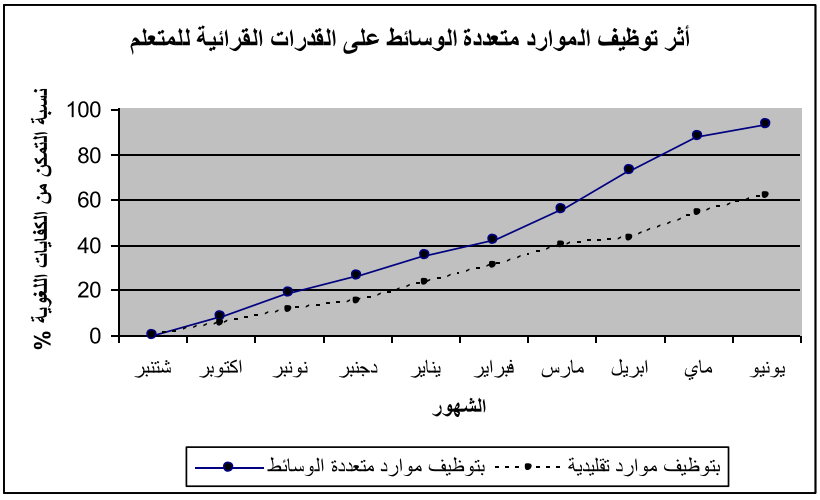
3.3.3. نماذج دعامات ديداكتيكية متعددة الوسائط

يتعرف المتعلم على أبجدية تيفيناغ بالنقر على الحروف وربط صورتها بنطقها.

ج) اثر توظيف تكنولوجيا الإعلام والاتصال في دروس اللغة الأمازيغية على المردودية التربوية

يتجلى لنا الأثر الايجابي لتوظيف تكنولوجيا الإعلام والاتصال في المكونات التعليمية لمادة اللغة الأمازيغية من خلال نتائج تجربة بسيطة قمنا بها على فوجين من متعلمي المستوى الأول لمدرسة ابتدائية: الفوج الأول يعتمد موارد متعددة الوسائط في أنشطته المدرسية (الخط المتصل في المبيان)، والفوج الثاني يعتمد موارد تقليدية (الخط المتقطع في المبيان)؛ وتهدف هذه التجربة إلى تقويم مدى تمكن كل من الفوجين من الكفايات اللغوية الأمازيغية خلال الموسم الدراسي وباعتماد الموارد الديداكتيكية المقترحة عليه، وجاءت النتائج كما هو واضح في المبيانات التالية:





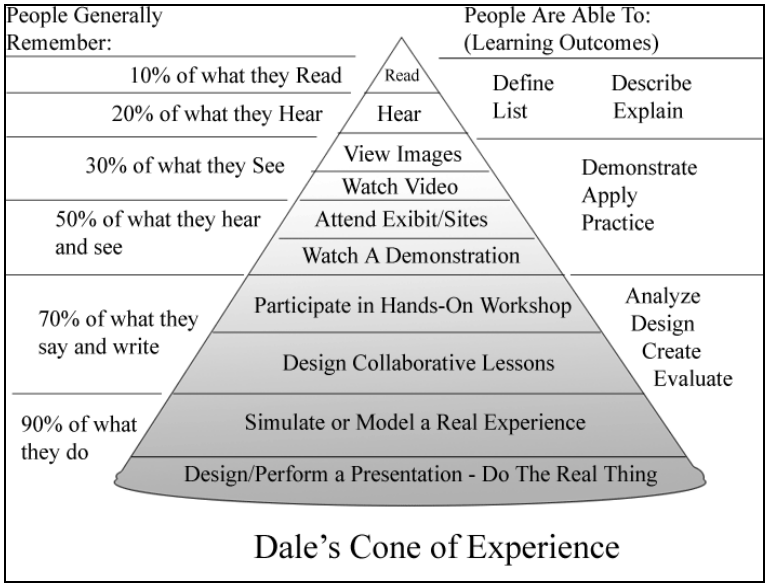
الرؤية، وهناك من يتعلم أكثر عن طريق الممارسة والتجربة، ومن المتعلمين من يستفيد أكثر بواسطة السماع.. ولكي تكون المردودية التربوية لمدرس اللغة الأمازيغية على مستوى محمود من الفعالية، ينبغي له تكييف الموارد الديداكتيكية التي يستخدمها مع الفوارق الفردية بين تلامذته، وهذا ما نسميه بالعمل بالبيداغوجيا الفارقية La pédagogie différenciées.

- ❖ تتجاوز إكراهات الزمان والمكان وتحل مشكل ندرة الموارد البشرية؛ فبنيات التعلم الالكتروني عن بعد Les plate-forme E-learning مثلا تحتوي على جميع محددات العملية التعليمية/التعلمية من مدرس ومتعلمين ومحتوى واليات للتقويم...الخ، وهذا دون حاجة لحضور فيزيقي للمدرس ودون التقيد بزمن مدرسي معين. وهذه الميزة يمكن توظيفها في أجرأة مصوغات أمازيغية خاصة بمناهج محو الأمية والتربية غير النظامية، والتكوين المستمر على سبيل المثال .
- ❖ تسهل على المدرس عملية التخطيط للدروس، ووضع السيناريوهات التعليمية؛ حيث تبيَّن أن الحاسوب أداة فعالة لتنظيم المعلومات المتعلقة بالبرامج التربوية الفردية للمتعلمين، فالبيانات التقويمية حول مواطن القوة والتعثر الفردية يمكن الاحتفاظ بها وتحيينها بسهولة في الحاسوب ومعالجتها واستثمارها بشكل ذكي في اتخاذ القرارات البيداغوجية الملائمة لكل متعلم، وكذا تعديل وتجديد السيناريوهات التعليمية بشكل دقيق ومضبوط.
- ❖ تشكل بيئة تعليمية ذات أدوات ومصادر معلومات متعددة (حواسيب، أقراص مدمجة تفاعلية، كتب رقمية، شبكة الانترنيت، فصول الافتراضية Classes virtuelles، سبورات بيضاء تفاعلية TBI ... وغيرها)، يتعامل معها المتعلم وتتيح له فرص اكتساب المهارات والخبرات وإثراء معارفه الثقافية واللغوية بسلاسة.
- ❖ تكسر الجمود في الجدول المدرسي التقليدي وذلك بتغيير مكان التعلم وأساليب التعليم ووسائله وتقنياته.
- ❖ تنمي لدى المتعلم روح التعاون والعمل التشاركي، وتساعده على الانفتاح على الآخر؛ فدراسة لغة وثقافتها لا يؤدي إلى الانفتاح على ثقافات وبيئات لغوية أخرى فحسب، بل قد يكسب المتعلم كفايات منهجية ضرورية لدراسة محتويات معينة (نصوص، خطابات، بيانات، صور، أفلام، مشهد مسرحي...الخ) .
- ❖ تنمية مهارات البحث والاستكشاف والتفكير وحل المشكلات لدى المتعلم .
- ❖ تزويد المتعلم بمهارات وأدوات تجعله قادرا على الاستفادة من التطورات المتسارعة في مجال تعليم اللغات باعتماد التكنولوجيا الحديثة للإعلام والاتصال.
- ❖ تتيح للهيئات التربوية إمكانية الإطلاع على المستجدات العلمية في الميدان البيداغوجي، مما يساهم بشكل ملحوظ في خفض تكاليف التكوين المستمر لهذه الهيئات.
- ❖ تعطي المتعلم الإحساس بالمساواة؛ بما أن أدوات الاتصال تتيح لكل متعلم فرصة الإدلاء برأيه في أي وقت ودون حرج، خلافا لقاعات الدرس التقليدية التي تحرمه من هذا الميزة إما لسبب سوء تنظيم فضاء الفصل الدراسي، أو ضعف صوت المتعلم نفسه، أو الخجل ، أو غيرها من الأسباب، لكن تكنولوجيا الإعلام في التعليم تتيح الفرصة كاملة للمتعلم لكي يبحث ويبني تعلماته ويبدع، ويعبر عن رأيه من خلال أدوات الاتصال من بريد إلكتروني ومنتديات النقاش وغرف الحوار وبرامج المحادثة الآنية وغيرها.

كما ينبغي أن نؤكد على أن توظيف هذه التكنولوجيات لا يعد تطورا في المنظومة التربوية المغربية فحسب، بل أصبح مطلبا راهنا لإشباع حاجة جميع مكونات المجتمع المغربي في رؤية هويتهم الثقافية واللغوية من خلال تقنيات وأدوات متجددة، يستعملونها ويتواصلون بها كل يوم.

أصبحت العملية التعلمية/التعليمية تشاركية بينه وبين المتعلم. وهذا في حد ذاته ييسر من الوقت والجهد الذي يبذله المدرس، ويمكنه من تطوير وتحسين أدائه كمربي.

- تتيح للمتعلم فترة تذكر أطول للمهارات اللغوية؛ فقد أثبت الباحثون في علم النفس المعرفي بأننا نتذكر وندرك عموما: 10 بالمئة مما نقرؤه و 20 بالمئة مما نسمعه و30 بالمئة مما نراه و50 بالمئة مما نراه ونسمعه و80 بالمئة مما نقوله و90 بالمئة مما نفعله. إن الطبيعة المرئية والمتحركة والتفاعلية لتكنولوجيا الإعلام والاتصال، تجعل منها الوسيلة التي تخدم الأهداف التعلمية أكثر من غيرها. (أنظر مخروط الخبرة ل"ديل Dale") [13]



مخروط الخبرة لديل Dale

- تيسر عملية الفهم وتساعد على تشكيل التفكير المنطقي لدى المتعلم؛ فالفهم هو قدرة الفرد على تمييز المدركات الحسية و تصنيفها وترتيبها بكيفية منطقية، فالفرد يتصل بالأشياء والمظاهر المختلفة عن طريق حواسه، و بالطبع لا يستطيع هذا الفرد أن يفهم المسميات أو الأشياء إلا إذا تم فهمها و التعرف عليها. وتتمظهر فعالية تكنولوجيا الإعلام والاتصال هنا في تمكين متعلم اللغة الأمازيغية في إدراك معنى بعض المضامين التعليمية، كالنصوص القرائية ونصوص التواصل الشفهي.
- تبث روح المبادرة في المتعلم والمدرس، وتشجع على الإبداع و التجديد.
- تأخذ بعين الاعتبار القدرات السيكولوجية والحسية لكل متعلم؛ فلكل متعلم خصوصيات تميزه عن زملائه في جماعة الفصل الدراسي، فقد أثبت دراسات علم النفس التربوي أن هناك اختلافات وفروق فردية بين المتعلمين في أساليبهم الإدراكية والمعرفية، إذ لكل متعلم أسلوبه الخاص في التعلم، ينظم به خبراته وإدراكاته؛ فمن المتعلمين من يتعلم بشكل أفضل عن طريق

---

[13] Principles of Teaching, Bloomsburg University, 2003. Url: http://teacherworld.com/potdale.html

2.3.3. تكنولوجيا الإعلام والاتصال كدعامة ديداكتيكية لتدريس اللغة الأمازيغية

أ) تكنولوجيا الإعلام والاتصال وبيداغوجيا التواصل

تتبني العملية التعليمية/التعلمية من منظور المقاربة التواصلية على أساس أن الفعل التعليمي/التعلمي هو مجموعة من عمليات التواصل والتفاعل؛ فاكتساب المهارات والكفايات – سواء كانت لغوية أو غيرها – لا يمكنه أن يتحقق إلا بين أفراد (متعلمون ومدرس) في موقف تعلمي معين، أو بين الفرد ومورد ديداكتيكي (حاسوب مثلا)، وبما أن العملية التعليمية/التعلمية تعتبر من تجليات مفهوم التواصل، فإنه بإمكاننا أن نتحدث عن مرسل وخطاب ومرسل إليه وقناة تواصل:

- المرسل؛ وهو مصدر الخطاب، وقد يكون إنسانا (المدرس) أو موردا ديداكتيكيا (برمجية، انترنيت، كتاب رقمي ...).
- الخطاب؛ ويمثله المحتوى موضوع التعلم. والمقصود به في مقامنا هذا المعارف اللغوية (القواعد والمتحكمات اللغوية مثلا) أو الأداء التواصلي (التعبير الشفهي أو الكتابي أو الإلقاء الشفهي المعبر مثلا) أو الاتجاهات والقيم (قيم التسامح والمواطنة مثلا).
- المرسل إليه؛ والمقصود به المتعلم الذي وُجه له الخطاب.
- القناة؛ نسميها كذلك بالوسيط وتتمثل في كل العناصر التي تسهل عملية التواصل بين المرسل والمخاطب، من لغة وطرق وتقنيات التدريس ووسائله. وهذه الأخيرة تحيلنا على الحديث عن أهمية تكنولوجيا الإعلام والتواصل كوسيط في نقل الخبرات والكفايات اللغوية إلى المتعلم.

ب) المزايا البيداغوجية للدعامات الديداكتيكية متعددة الوسائط في تعليم وتعلم اللغة الأمازيغية

ترتبط مزايا الدعامات الديداكتيكية متعددة الوسائط في اللغة الأمازيغية بمدى قدرتها على تحسين مخرجات الممارسة التربوية خلال درس اللغة الأمازيغية عموما. ويمكن تلخيص هذه المزايا فيما يلي:

- تشد وتحافظ على انتباه المتعلم وتخلق لديه الدافعية للتعلم La motivation؛ إن بث الدافعية للتعلم يعتبر مهمة أساسية في السيرورة التعليمية وتوجيه عملية التعلم. وهذه الدعامات تثير فضول المتعلم وتحفزه للتفاعل مع موضوع التعلم، الذي يصبح ذا أهمية بالنسبة للمتعلم. وتعتمد الدعامات الديداكتيكية متعددة الوسائط على عناصر الإثارة والتشويق، ونقصد بها: الصوت، الصورة، الألوان، الحركة (مثال: قاموس تفاعلي يحتوي على صور أشياء وكائنات مع تسمياتها النصية والصوتية)، وهذا على خلاف الدعامات الديداكتيكية التقليدية التي تتسم بالجمود في غالب الأحيان.
- مساعدة المتعلم على إدراك بعض المتحكمات اللغوية الأمازيغية، واكتسابه رصيدا وظيفيا أمازيغيا وتجنبه اللفظية؛ وهذا من خلال مواقف تواصلية حقيقية، والتي تمكن صياغتها على شكل مقاطع فيديو Séquences vidéo، مصحوبة بنصوص، تعرض أمام المتعلم الذي يسعى بدوره لمحاكاتها، مكتسبا بذلك كفاية لغوية تواصلية أمازيغية سليمة، زيادة على أن هذه التكنولوجيا قد تساهم في الحد من مشكل النطق والتأتأة لدى فئة من المتعلمين.
- تنمي في المتعلم روح الاستقلالية وتساعده على التعلم الذاتي L'autonomie؛ فتكنولوجيا الحاسوب تسمح للمتعلم ببناء معارفه بنفسه وبطريقة تفاعلية وفعالة كذلك، فالتعلم الذاتي أسلوب من أساليب التعلم المتطورة التي تمكن الفرد من أن يعلم نفسه بنفسه، وفقاً لقدراته الاستيعابية ولسرعته في التعلم ، وبما يتوافق مع ميوله واهتماماته. كما تمكنه هذه التكنولوجيا من تقويم مكتسباته ومهاراته اللغوية بشكل فوري، واللجوء إلى التغذية الراجعة Feed-back والتمارين الداعمة في أي حين وبشكل مستمر، دون التقيد بزمن مدرسي أو دراسي معينين. ودور المدرس – في هذا المنحى – قد تغير، فلم يعد قاصرا على نقل المعلومات والتلقين، بل

- ❖ محور المضامين الرقمية: في إطار هذا المحور ستعمل وزارة التربية الوطنية على توفير مضامين بيداغوجية تدعم المناهج الدراسية الوطنية، تعتمد على وسائط متعددة Multimédia (مثال: مضامين مثبتة على الحواسيب أو على أقراص مدمجة، بوابة تربوية متعددة الموارد على الانترنيت..). كما سيتم إحداث مختبر وطني لإعداد وتطوير المضامين البيداغوجية الرقمية، تناط به مهام الإنتاج والمصادقة على محتويات وطنية لدعم تعلمات التلاميذ ومساعدتهم على إنجاز مشاريع البحث عن الوثائق والمعلومات. [11]

3.3. اللغة الأمازيغية وتكنولوجيا الإعلام والاتصال التعليمية

1.3.3. منطلقات وأهداف تدريس اللغة الأمازيغية

يستمد إدماج اللغة الأمازيغية في المنظومة التعليمية مرجعيته الأساسية من خطاب العرش في 30 يوليو 2001، الذي أعلن فيه صاحب الجلالة عن إدراج الأمازيغية لأول مرة بالنسبة لتاريخ بلادنا في المنظومة التربوية الوطنية، ومن الخطاب الملكي السامي في أجدير بتاريخ 17 أكتوبر2001 حيث أكد جلالته، بمناسبة الإعلان عن الظهير الشريف المحدث للمعهد الملكي للثقافة الأمازيغية، أن "..الأمازيغية، التي تمتد جذورها في أعماق تاريخ الشعب المغربي، هي ملك لكل المغاربة بدون استثناء ..."، وأنها "...مكون أساسي للثقافة الوطنية، وتراث ثقافي زاخر شاهد على حضورها في كل معالم التاريخ والحضارة المغربية...".

وإذا رجعنا إلى النصوص المنظمة لتدريس مادة اللغة الأمازيغية، وبالخصوص منها الدليل المنظم لتدريسها ، والصادر عن وزارة التربية الوطنية في شتنبر 2006، فإننا سنجد أن تدريس هذه المادة مؤطر بتوجهات وأهداف محددة، يمكن تلخيصها في ما يلي:

- ❖ تمكين المتعلمين من إتقان اللغة الأمازيغية نطقا وقراءة وكتابة؛
- ❖ الانطلاق في وضع منهاج اللغة الأمازيغية من الفروع الأساسية للغة الأمازيغية، مع العمل بالتدرج على بناء لغة معيارية موحدة من خلال التركيز على البنيات اللغوية المشتركة بين هذه الفروع، وإعطائها الأولوية في وضع الكتب المدرسية والدلائل والملفات البيداغوجية والدعامات الديداكتيكية الأخرى، المكتوبة والسمعية والبصرية؛
- ❖ إغناء وتطوير الرصيد اللغوي الأمازيغي باعتماد الإبداع المعجمي، وتوظيف المعجم الأمازيغي المتداول في الدارجة المغربية، والانفتاح على الفروع الأمازيغية الأخرى المتداولة في مناطق أمازيغية خارج الوطن لإغناء المعجم الأمازيغي الوطني المشترك؛
- ❖ إخضاع عملية تعلم اللغة الأمازيغية لنظام التقويم المعتمد في باقي المواد. [12]

واعتمادا على الاختيار التربوي الذي يحكم باقي المناهج التعليمية، الذي يتبنى تنمية كفايات المتعلم كمرتكز أساسي، فإن تدريس اللغة الأمازيغية يستهدف بالأساس تنمية الكفايات التواصلية عند المتعلم (كفايات الإنصات، وكفايات التكلم، وكفايات القراءة، وكفايات الكتابة)، كما يستهدف تنمية الكفايات الإستراتيجية (أو كفايات تنمية الذات) والكفايات الثقافية (الرمزية منها والموسوعية) والكفايات المنهجية والكفايات التكنولوجية.

---

[11] موقع وزارة التربية الوطنية والتعليم العالي وتكوين الأطر والبحث العلمي، 2005.
http://genie.men.gov.ma

[12] وزارة التربية الوطنية والتعليم العالي وتكوين الأطر والبحث العلمي، المذكرة رقم 30 حول تنظيم تدريس اللغة الأمازيغية وتكوين أساتذتها، 2006.

في الثلاثينات، التلفزة المدرسية في الخمسينات، المعلوميات في السبعينات، تقنية الفيديو في الثمانينات، الموارد متعددة الوسائط في التسعينات..

3. تكنولوجيا الإعلام والاتصال لتعليم وتعلم اللغة الأمازيغية

1.3. تكنولوجيا الإعلام والاتصال في الميثاق الوطني للتربية والتكوين

أعطى الميثاق الوطني للتربية والتكوين الخطوط العريضة التي سيتم وفقها إدماج تكنولوجيا الإعلام والاتصال في النظام التعليمي المغربي. فنقرأ في الدعامة العاشرة من الميثاق ما يلي:

"سعيا لتحقيق التوظيف الأمثل للموارد التربوية ولجلب أكبر فائدة ممكنة من التكنولوجيات الحديثة، يتم الاعتماد على التكنولوجيات الجديدة للإعلام والاتصال وخاصة في مجال التكوين المستمر[...] ونظرا للأبعاد المستقبلية لهذه التكنولوجيات سيستمر استثمارها في المجالات الآتية، على سبيل المثال لا الحصر:

- معالجة بعض حالات صعوبة التمدرس و التكوين المستمر بالنظر لبعد المستهدفين وعزلتهم ؛
- الاستعانة بالتعليم عن بعد في مستوى الإعدادي والثانوي في المناطق المعزولة ؛
- السعي إلى تحقيق تكافؤ الفرص، بالاستفادة من مصادر المعلومات، وبنوك المعطيات، وشبكات التواصل مما يسهم، بأقل تكلفة، في حل مشكلة الندرة والتوزيع غير المتساوي للخزانات والوثائق المرجعية.

ومن هذا المنظور، ستعمل السلطات المكلفة بالتربية والتكوين، في إطار الشراكة مع الفعاليات ذات الخبرة، على التصور والإرساء السريعين لبرامج للتكوين عن بعد، وكذا على تجهيز المدارس بالتكنولوجيات الجديدة للإعلام والتواصل[...]". [(10)]

تبعا لهذه التوجهات العامة سيعطي جلالة الملك محمد السادس إشارة انطلاق البرنامج الوطني تعميم تكنولوجيا الإعلام والاتصال في ميدان التربية والتكوين " جيني Génie"، وكان ذلك في 15 من شتنبر 2005.

2.3. البرنامج الوطني لتعميم تكنولوجيا الإعلام والاتصال في ميدان التربية والتكوين "جيني Génie"

يشكل البرنامج الوطني لتعميم تكنولوجيا الإعلام والاتصال في ميدان التربية والتكوين "جيني Génie"، أحد التوجهات الإصلاحية الكبرى التي حددها الميثاق الوطني للتربية والتكوين، وحلقة أساسية على درب متابعة الإصلاح التعليمي الشامل. فقد تبنت الحكومة المغربية في مارس 2005 ، إستراتيجية تقضي بتعميم التكنولوجيات الحديثة للإعلام والاتصال في التعليم العمومي (المدارس الابتدائية، الثانويات الإعدادية، الثانويات التأهيلية، الجامعات)، بشكل يواكب البرامج المدرسية الوطنية. وتنبني هذه الإستراتيجية على ثلاثة محاور رئيسية :

- محور البنية التحتية: في إطاره سيتم توفير قاعات متعددة الوسائط المؤسسات التعليمية – تسمى بيئة العمل الرقمي– مجهزة بالعتاد المعلوماتي ومرتبطة بشبكة المعلومات العالمية "الانترنيت".
- محور التكوين: يهدف هذا المحور إلى تأهيل وتكوين الهيئة التربوية (هيئة التدريس والإدارة والمراقبة التربوية) والأطر الإدارية في مجالات استعمال تكنولوجيا الإعلام والاتصال.

[(10)] موقع وزارة التربية الوطنية والتعليم العالي وتكوين الاطر والبحث العلمي، 1999.
http://www.men.gov.ma/dajc

والذي يحدد مفهوم تكنولوجيا التعليم في "تطبيق المعرفة عن طريق التكنولوجيا بغرض رفع مستوى التعليم، أو هي استخدام الوسائل التكنولوجية في العملية التعليمية". في حين يعرفها احمد حامد منصور [8] بأنها "بمعناها الشامل تضم جميع الطرائق والأدوات والمواد والأجهزة والتنظيمات المستخدمة في نظام تعليمي بغرض تحقيق أهداف تعليمية محددة من قبل". ويضيف:"إن تكنولوجيا التعليم لا تعني مجرد استخدام الآلات والأجهزة الحديثة، ولكنها تعني في المقام الأول طريقة في التفكير لوضع منظومة تعليمية، أي أنها تأخذ بأسلوب المنظومات."

أما منظمة الأمم المتحدة للتربية والعلم والثقافة "اليونسكو"، فتعرف تكنولوجيا التعليم بكونها "منحنى نظامي لتصميم العملية التعليمية وتنفيذها وتقويمها كلها تبعًا لأهداف محددة نابعة من نتائج الأبحاث في مجال التعليم والاتصال البشري مستخدمة الموارد البشرية وغير البشرية من أجل إكساب التعليم مزيدًا من الفعالية (أو الوصول إلى تعلم أفضل وأكثر فعالية)."

من خلال هذه التعاريف يمكن أن نخلص إلى ما يلي:

- تأثر مفهوم تكنولوجيا التعليم بمفهوم التكنولوجيا ذاته بما تنطوي عليه من آلات مبرمجة وأجهزة تلفاز وحواسيب وغيرها.
- ضرورة الحذر من الخلط بين التكنولوجيا في التعليم التي يقصد بها توظيف بعض الوسائل التكنولوجية، كما هو الحال في الاستعانة بالموارد متعددة الوسائط التي تتيحها تكنولوجيا الإعلام والاتصال، وبين تكنولوجيا التعليم التي هي إطار نظري يتناول العملية التعليمية/التعلمية التي تعتمد على فلسفة النظم كمنهج وترتكز على الفلسفة التكنولوجية المتميزة بالضبط والتنظيم وتحديد المدخلات والمخرجات من أي عملية.
- إن تكنولوجيا التعليم في أرقى تجلياتها (الاستعانة بتكنولوجيا الإعلام والاتصال والموارد التربوية متعددة الوسائط...الخ) لا تعتبر نقيضا للطريقة التقليدية أو بديلا لها؛ فالتدريس بالكفايات L'enseignement par compétences أو التعليم المبرمج L'enseignement programmé مثلا كأشكال وتقنيات تعليمية تدخل في نطاق تكنولوجيا التعليم، لا تعني بالضرورة نهج طرائق فعالة؛ وحول هذا الموضوع يعلق دو روسناي J. De Rosnay [9] قائلا: "من دون مقاربة شاملة، تبقى المحاولات المختلفة لعصرنة التعليم محاولات محكوما عليها بالفشل...إن السمعي البصري لا يحقق أهميته البيداغوجية المباشرة إلا إذا كان التلميذ هو الذي يمارس الفعل المرتبط بما شاهده على الشاشة...".

من هذا المنطلق لا يمكن القول بأن ما توفره تكنولوجيا التعليم – بما فيها تكنولوجيا الإعلام والاتصال التعليمية – من تنظيم وضبط وهندسة بيداغوجية، يعتبر تحولا وتغييرا في الطريقة والمنهجية التعليميتين، ذلك أن محددات مفهومي الطريقة والمنهجية تبقى ثابتة لا تتغير.

3.2. تكنولوجيا الإعلام والاتصال التعليمية كجزء من تكنولوجيا التعليم

تكنولوجيا الإعلام والاتصال التعليمية جزء من تكنولوجيا التعليم، ويمكن أن نعرفها بأنها مجموع الأدوات والتقنيات والموارد الرقمية التي يمكن أن توظف في ميدان التربية والتكوين.

ويأتي إدماج هذه التكنولوجيا في الممارسة التربوية كمرحلة من المراحل التي سعت عبرها الأنظمة التعليمية لمواكبة التطور التكنولوجي السريع الذي عرفه العالم خلال القرن العشرين: الإذاعة المدرسية

---

[8] احمد حامد منصور، تكنولوجيا التعليم وتنمية القدرة على التفكير الابتكاري، منشورات ذات السلاسل، الطبعة الأولى، 1986، ص30.

[9] J. De Rosnay, Le macroscope, vers une vision globale, Edit Seuil (Coll Le point), 1977 , P261.

- **المرحلة الرابعة:** برز فيها مصطلح "وسائل الاتصال التعليمية"، وهذا بحكم ما عرفه علم الاتصال من تطور أثر على ميدان التربية والتعليم. فظهرت أطروحات تقول بأن أية علاقة إنسانية هي علاقة تواصلية وعملية التعليم بالضرورة تحكمها محددات التواصل. فلما كان المدرس مرسلا والمتعلم مستقبلا والمحتوى التعليمي هو الخطاب، فإن قناة التواصل Le canal de communication التي من خلالها تنتقل المعارف والمهارات من المدرس إلى المتعلم هي وسائل الاتصال التعليمية. وفي هذه المرحلة تم الانتقال من "مجرد الاهتمام بتوفير المواد التعليمية إلى الاهتمام بجوهر العملية وهو تحقيق التفاهم" [5]
- **المرحلة الخامسة:** في هذه المرحلة لم تعد فيها الوسائل مجرد أدوات تساعد على الإيضاح أو التواصل، بل تحولت وظيفتها وطبيعتها تحولا عميقا، فأخذت شكلا آخر يسمح بإدماج كلي لها ضمن الفعل التعليمي والتعلمي. إنها مرحلة تكنولوجيا التعليم التي أصبحت فيها الأداة التعليمية عنصرا أساسيا، تجاوزت فيها كونها مجرد أدوات، لتصير مكونا أساسيا ضمن أنظمة وأنساق systèmes، تحكم العملية التعليمية / التعلمية.

وتحيلنا هذه المرحلة الأخيرة على الحديث عن تعليم اللغة بمساعدة الحاسوب الذي ظهر في تلك المرحلة. فقد بدأ استخدام هذه الأداة في الستينيات من القرن العشرين بالولايات المتحدة الأمريكية. وكان أول مشروع لاستخدام الحاسوب في تعليم اللغات قد جرى بجامعة ستانفورد بولاية فلوريدا الأمريكية. وقد تأثرت المشروعات الأولى لتعليم اللغة بمساعدة الحاسوب بنظرية السلوك التي كانت تسيطر على تعليم اللغة واكتسابها، وهي النظرية التي تنظر إلى اكتساب اللغة على أنها علاقة محفز واستجابة، فكانت تطبيقات اللغة بمساعدة الحاسوب في الستينيات والسبعينيات جميعها تسمى التطبيقات السلوكية لتعليم اللغة بمساعدة الحاسوب. وكانت تلك التطبيقات ترتكز على التدريب، والتكرار.

عموما يمكن أن نقول بأن هذا التطور الحاصل في التسميات والوظائف لا يعني أبدا وجود قطيعة بين المراحل، كما لا يعني أن المدارس التعليمية قد تجاوزت المراحل الأولى في ممارساتها، إذ نجد اغلبها لا تزال تتموقع في المرحلة الأولى، مما يمكن اعتباره تخلفا عن التطور الذي يعرفه الميدان التربوي، كما هو الحال في بعض المؤسسات التعليمية المغربية.

2.2. مفهوم تكنولوجيا التعليم

تبعا للقراءات التي يمكن أن نقوم بها بهذا الخصوص، نستطيع أن نقول بأنه ليس هناك فهم واحد لمفهوم تكنولوجيا التعليم. فحسب De Corte وزملاؤه، هناك ثلاث دلالات متباينة لهذا المفهوم [6]:

أ) إدخال الأجهزة إلى ميدان التعليم.
ب) تنظيم التعليم بالاستعانة بالآلات.
ج) دلالة ترتبط بنظرية تعليمية تعتمد على تطبيق سيكولوجيا التعلم والمقاربة النسقية للعملية التعليمية/ التعلمية والتصورات السبرانية (النظرية الاعلاميائية).

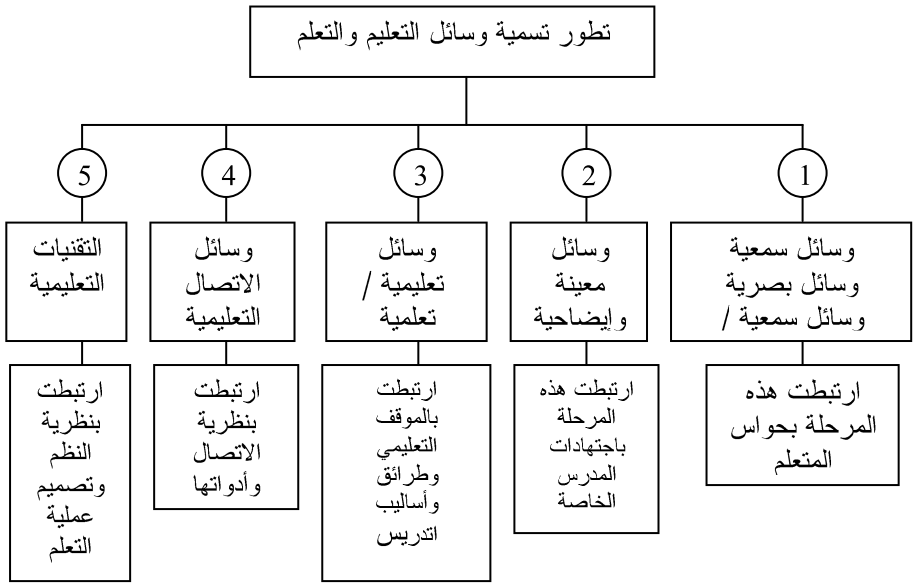
أما كلارك Clark فيعرف تكنولوجيا التعلم بما يلي: "هي الاستفادة من المخترعات والصناعات الحديثة في مجال التعليم". وهذا ما نجده أيضا تعريف شادويك Chadwick الذي أورده عبد الرحيم الكلوب [7]

---

[5] حسين حمدي الطوبجي، وسائل الاتصال والتكنولوجيا، دار القلم، الكويت 1987، ص24.

[6] A De Corte et coll : Les fondements de l'action didactique, traduit du Néerlandais par V.V. Culsen, Paris, Editions Universitaires, 1990, P187.

[7] بشير عبد الرحيم الكلوب، التكنولوجيا في عملية التعلم والتعليم، دار الشروق للنشر والتوزيع، الطبعة الثانية، عمان الأردن، 1993، ص36.



من خلال هذه الخطاطة نلاحظ أن:

- **المرحلة الأولى:** ارتبطت فيها التسمية بحواس المتعلم، حيث أن التربويين آنذاك يعتقدون بان عملية التعلم لا يمكنها أن تكون إلا انطلاقا من خبرة حسية " فكل مفاهيمنا عن العلاقات المختلفة مستقاة منها. وهذه المفاهيم تستخدم في تكوين مفاهيم أخرى عندما تتسع خبرتنا وتمتد" [3].

وهذه المرحلة التي اقترنت بالأساس بحاستي السمع والبصر، ويمكننا تقسيمها إلى ثلاث مراحل جزئية:

- مرحلة التعليم والتعلم المعتمد على المرئيات؛ حيث ساد استخدام الخرائط والصور الثابتة وذوات الأشياء، اعتقادا بأن عملية التعلم تتم بشكل أفضل عبر حاسة البصر.
- مرحلة التعليم والتعلم السمعي؛ بظهور أشرطة التسجيل الممغنطة وغيرها.
- في المرحلة الثالثة تمت المزاوجة بين ما هو مرئي وسمعي؛ وهذا ما مكن منه ظهور التلفاز والسينما.

- **المرحلة الثانية:** ساد فيها مصطلح المعينات التعليمية ووسائل الإيضاح؛ هذه الوسائل والمعينات التي يلجأ إليها المدرس أحيانا عندما يجد نفسه عاجزا عن شرح وتوضيح ما لا يستطيع توضيحه بالاعتماد على ذاته فقط، بمعنى أن استعمال هذه المعينات كان ثانويا في العملية التعليمية/التعلمية.
- **المرحلة الثالثة:** في هذه المرحلة تطورت التسمية لتتخذ مصطلح "الوسائل التعليمية/التعلمية"، وهذا بحكم الترابط الجدلي بين التعليم والتعلم كعمليتين متفاعلتين، وكذلك لكون تلك الوسائل لم تعد الغاية من استخدامها الإيضاح بقدر ما أصبح لها غاية أخرى هي خلق شروط ضرورية للتعلم وتحقيق الأهداف [3].

---

[3] رشيد لبيب، الأسس العامة للتدريس، دار النهضة العربية، بيروت، 1983، ص119.

[3] محمد زياد حمدان، وسائل تكنولوجيا التعليم، الطبعة الثانية، دار التربية الحديثة، 1980، ص21.

تكنولوجيا الإعلام والاتصال كدعامة ديداكتيكية لتعليم وتعلم اللغة الأمازيغية بالمدرسة المغربية

ذ.عبد اللطيف حسيني
مدرس اللغة الأمازيغية
abdellatif.hssaini@taalim.ma

1. مقدمة:

عرف الميدان التربوي في الآونة الأخيرة تطورا هائلا، ويرجع هذا التطور لمجموعة من العوامل الاجتماعية والاقتصادية والعلمية، الشيء الذي أثر بشكل جلي في الغايات المتوخاة من العملية التعليمية/التعلمية ومنهجياتها ووسائلها وأدواتها وأنماطها.

ولم تكن المملكة المغربية بمعزل عن هذا التطور، إذ عملت على إصلاح نظامها التعليمي بما يكفل تحسين جودته ويجعله قادرا على مواجهة التحديات الراهنة والمستقبلية للبلاد. ووضعت الدولة لذلك سنة 1999 ميثاقا أسمته: الميثاق الوطني للتربية والتكوين. وبموجب هذا الميثاق ستتم إعادة هيكلة أسلاك النظام التعليمي، وتحسين برامجه ومناهجه البيداغوجية. ولعل أهم المواضيع التي ركز عليها هذا الميثاق، موضوع إدماج التكنولوجيا الجديدة للإعلام والتواصل في الممارسة التربوية وموضوع تحسين تدريس اللغات، بما فيها تدريس اللغة الأمازيغية كمكون تعليمي أساسي. [1]

وسنتناول في مقالنا هذا تكنولوجيا الإعلام والاتصال من منظور بيداغوجي وديداكتيكي، وأهميتها في تعليم وتعلم اللغة الأمازيغية، في ضوء ما جاء به الميثاق الوطني للتربية والتكوين، وما استقيناه من ممارستنا الميدانية في قطاع التعليم المدرسي.

2. تكنولوجيا الإعلام والاتصال كوسيلة تعليمية

إن المهتم بميدان التربية والتكوين سيلاحظ أن تكنولوجيا الإعلام والاتصال التعليمية، والتي يصطلح عليها بالفرنسية: Les Technologies de l'Information et de la Communication pour l'Enseignement، تدخل في إطار الوسائل التعليمية. وقد تطورت تسميتها في تساوق مع تاريخ تطورها، وارتباطا مع الرؤى البيداغوجية والديداكتيكية التي وظفت من أجلها.

1.2. مفهوم الوسيلة التعليمية

يؤكد الكثير من البيداغوجيين أن تسمية الوسائل التعليمية مرت بخمس مراحل أساسية، كما تبين هذه الخطاطة التي أعدها عبد الرحيم الكلوب خبير اليونسكو في تقنيات التعليم (2) :

---

[1] موقع وزارة التربية الوطنية والتعليم العالي وتكوين الاطر والبحث العلمي، 1999.
http://www.men.gov.ma/dajc

(2) بشير عبد الرحيم الكلوب، التكنولوجيا في عملية التعلم والتعليم، دار الشروق للنشر والتوزيع، الطبعة الثانية، عمان الأردن، 1993، ص20.

Marquage des mots et collecte de leurs usages

Abdelkrim MOKHTARI
Département de Langue et de Littérature Françaises
Faculté des Lettres et des Sciences Humaines
Université Ibn Tofail, Kénitra, Maroc
abdelkrim_mokhtari@yahoo.fr

Résumé

Cet article se propose de présenter un module d'extraction des usages de mots à partir de textes. Ce module est en fait intégré dans un logiciel dédié au marquage des textes, Marqueur. Ce dernier s'appuie sur un dictionnaire incorporé pour décider quel lemme et quelle catégorie attribuer au mot rencontré. Les mots univoques sont par conséquent automatiquement marqués. Pour les cas ambigus, une confrontation entre le contexte présent et des contextes emmagasinés préalablement permet de trancher avec plus ou moins de succès. Enfin, une interface de marquage manuel est proposée pour les cas qui ont échappé au traitement automatique. L'extraction des usages est utilisée pour des besoins ponctuels, mais elle est également conçue pour permettre l'alimentation du dictionnaire, enrichissant de la sorte les marques du dictionnaire.

Mots clés

Marquage des textes, étiquetage, extraction, dictionnaire électronique, usage, recherche d'information, concordance

Abstract

This article purposes to present a unit of extracting word usages from texts. Such unit is as a matter of fact integrated into a software program devoted to text tagging, *Marqueur*. The latter is based on an incorporated dictionary in order to determine what lemma and what category to be attributed to the word encountered. Univocal words are as a result automatically tagged. As for ambiguous cases, a confrontation of the current context with previously stored contexts allows a more or less successful settlement. Finally, a manual tagging interface is proposed for cases that resisted automatic processing. Usage extraction is resorted to for specific needs, but is also designed to take part in feeding the dictionary, improving thus the latter's tagging.

Keywords
Tagging, extraction, electronic dictionary, use, usage, text retrieval, concordance

1. Introduction

En Traitement Automatique des Langues, il est d'une grande utilité de disposer, en plus d'informations absolues, d'informations contextuelles sur les unités linguistiques rencontrées dans un texte à traiter, informations qui pourraient relayer les règles utilisées. Ces indications contextuelles s'obtiennent manuellement ou automatiquement et nécessitent des outils à même d'opérer le rattachement de l'information au mot en contexte. Le marquage des textes (ou étiquetage), qui concerne généralement la catégorie grammaticale et le lemme, peut être exploité en soi dans des applications comme il peut servir d'étape intermédiaire dans la préparation d'autres sources d'information.
Le marquage des textes peut être mis au service d'applications relevant de disciplines diverses (recherche documentaire, lexicographie, terminologie, analyse de contenu, analyse du discours etc.). Il trouve en particulier sa place dans la linguistique dite de corpus. Comme l'expliquent Habert B. et al. (1997, pages 7 et suivantes), des recherches sont faites dans le cadre d'une linguistique descriptive qu'il faut opposer à la linguistique fondée sur des modèles et sur l'intuition. La linguistique du corpus repose sur l'observation empirique de textes préalablement étiquetés. Les outils de marquage sont donc les bienvenus pour préparer ces corpus, sachant que des applications spécifiques pourront exploiter de tels matériaux. Marqueur[25] est un outil qui s'inscrit dans cette mouvance.
Nous proposons ici de présenter un des modules de ce logiciel. Ce module a été rajouté récemment et consiste à permettre l'extraction des fragments de texte entourant les occurrences du mot recherché. L'utilisateur aura la latitude de retenir ceux qui dénotent un usage particulier du mot en question.
A ce stade, il peut sembler que les deux opérations (marquage et extraction), qui relèvent d'activités opposées, n'ont pas de liens entre elles. Mais nous avons intégré dans notre projet le moyen d'emmagasiner dans le dictionnaire les usages retenus, de manière telle que les deux activités se rejoignent et qu'en conséquence, les usages deviennent eux-mêmes des marques.
Dans ce qui suit, nous rappelons le fonctionnement de Marqueur, de manière brève puisque déjà exposé ailleurs (Mokhtari, 2005), mais suffisamment pour que l'exposé sur le module d'extraction des usages, qui est donné dans la seconde section, soit intelligible. Nous exposerons ensuite la démarche suivie pour

[1]Nous avons développé le logiciel Marqueur en collaboration avec Michael Mepham, dans le cadre de notre thèse Ph. D. que nous avons soutenue en 1998. Nous l'avons amélioré depuis et l'avons présenté, entre autres, aux Septièmes Journées Scientifiques du réseau de chercheurs Lexicologie, Terminologie Traduction, Bruxelles, 2005.

l'extraction et l'exploitation des usages avant de donner quelques perspectives et une conclusion.

2. Un marquage enrichi

Une unité linguistique peut se décliner sous plusieurs formes. Seule l'une d'entre elles est retenue pour incarner l'entité, le mot-entrée, le type ; on parle alors de lemme. Ainsi, toutes les flexions d'un mot comme *petit*, à savoir *petite*, *petits*, *petites* sont des formes du lemme *petit*. Et toutes les formes de conjugaison du verbe *avoir* ne sont que des manifestations du lemme *avoir*. Bien évidemment, des ambigüités existent, et nous ne sommes pas toujours, en tant qu'usagers de la langue, conscients par exemple qu'une des formes du lemme *avoir* croise une des formes du lemme *avion* (il s'agit de *avions*). Il peut paraître naturel que les mots *discours*, *faudrait, pu,* soient lus comme des formes associées aux lemmes respectifs *discours* (nom), *falloir* (verbe), *pouvoir* (verbe) plutôt qu'aux lemmes *discourir*, *faillir*[26], *paître*[27], également possibles.

La catégorie grammaticale elle-même peut conduire à deux lemmes différents. C'est pourquoi, dans notre conception du système de marquage et d'extraction de données linguistiques, nous avons regroupé les deux dimensions dans un seul concept : lemme-catégorie. La catégorie grammaticale et le lemme sont assemblés dans un même trait distinctif.

Le logiciel Marqueur est un outil qui permet d'annoter les textes. Une large partie du corpus est traitée automatiquement, le reste est traité interactivement par l'utilisateur. Un dictionnaire incorporé au logiciel relève toutes les unités du lexique et fournit l'information sur le caractère ambigu ou non de l'unité traitée. Les formes linguistiques qui sont univoques, ne pouvant recevoir qu'un seul lemme-catégorie, sont traitées automatiquement. Celles qui sont attachées à plus d'un lemme-catégorie passent par une étape, également automatique du traitement, qui consiste à vérifier si le dictionnaire dispose de contextes lexicaux ou grammaticaux susceptibles de résoudre le cas en question. Ces contextes sont préalablement enregistrés dans le dictionnaire. Il aurait été souhaitable d'intégrer

---

[26] Le Grand Robert, version électronique, apporte la remarque suivante : « REM. Dans l'usage actuel, faillir n'est plus guère employé qu'au passé simple (dans la langue écrite), au participe passé, et surtout à l'infinitif et aux temps composés. Les autres formes, et en particulier le présent de l'indicatif, le futur et le conditionnel (dont les 3e pers. du sing. sont communes à faillir et à falloir : il faut, il faudra, il faudrait) étaient déjà vieillies au xixe s. C'est à cette époque que sont entrées dans l'usage les formes je faillirai, je faillirais au futur et au conditionnel (→ cit. 1). La tendance à conjuguer faillir tout entier sur finir, déjà notée par Littré (« … quelques grammairiens disent que ce verbe, dans le sens de faire faillite, se conjugue régulièrement sur finir : Quand un négociant faillit…; s'il faillissait… »), se recommande aujourd'hui de l'autorité de bons écrivains (cf. Grevisse, 701, 20). » C'est nous qui soulignons.

[27] La forme *pu* peut être associée au lemme *pouvoir* ou bien au lemme *paître*. Mais l'usage ne retient pas cette dernière association.

dans le système les règles morphologiques et syntaxiques, mais cet effort n'a pas été fait.
Pour tous les cas qui auraient échappé au traitement automatique, nous avons prévu une dernière étape. Elle est dite interactive ou manuelle. L'utilisateur parcourt les cas récalcitrants et les traite de manière séparée[28].
L'interface (voir Figure 2, page 114) permet de contrôler la progression du traitement et de vérifier les différentes marques.

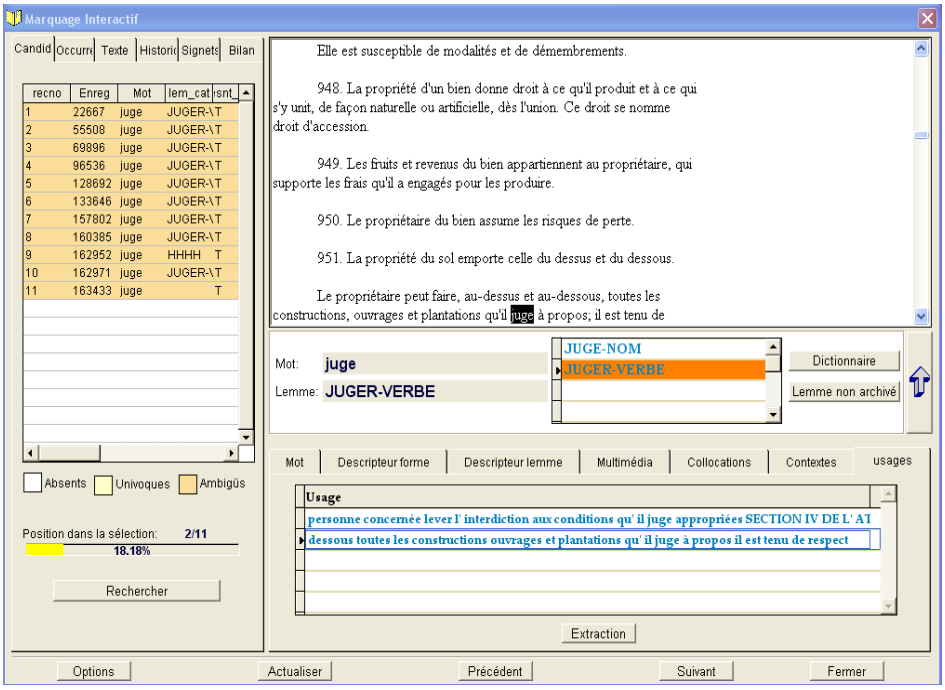


Figure 2 Interface pour marquage des mots.

La partie gauche de l'interface permet d'aller vers une fenêtre où on peut sélectionner les mots et les occurrences qu'on souhaite traiter, mais d'autres onglets proposent différentes informations sur l'avancement du traitement, le mode de traitement etc.
La partie droite comprend le texte objet du traitement, le mot courant, ses différents lemmes potentiels, le lemme retenu, et les autres marques rattachées au lemme, y compris les usages. L'accès au dictionnaire est assuré au niveau de cette interface.
Une procédure de sélection assez avancée (voir Figure 3, page 115) permet à l'utilisateur de sélectionner finement les cas à traiter ou à vérifier. L'utilisateur a ainsi le contrôle sur ce qui a été fait de manière opaque par le logiciel, et sur ce

[28] S'il y a plusieurs occurrences, le programme les traite automatiquement, soit immédiatement, soit de manière différée, selon les options retenues par l'utilisateur.

qu'il a lui-même décidé. Vu sous cet angle, le logiciel est un outil qui permet de préparer avec le maximum de certitude des textes témoins, dénommés *golden texts*, Voir Brill, E. (1995)).

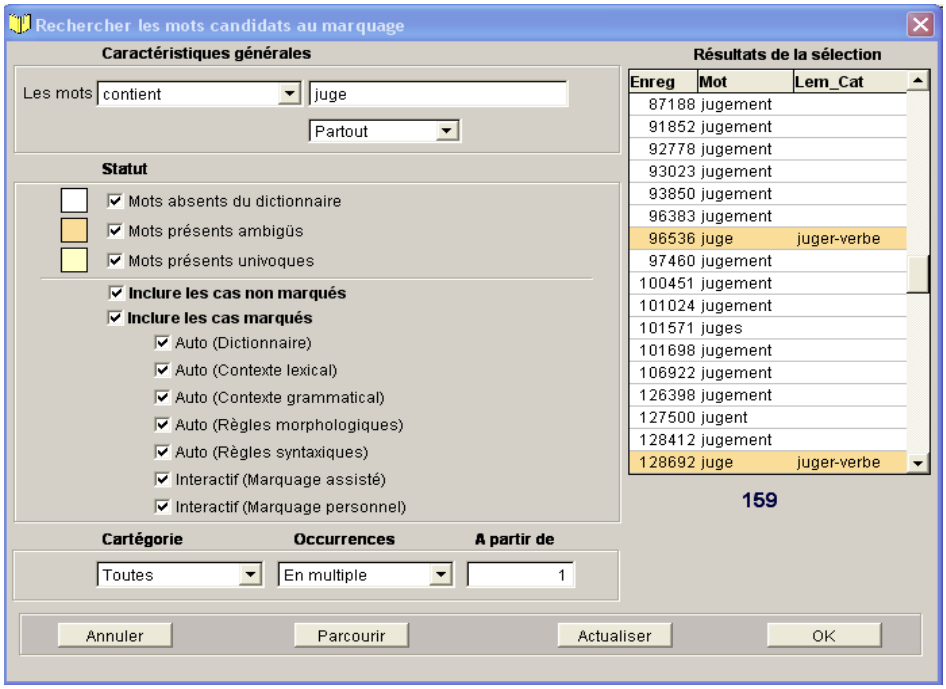


Figure 3 Interface permettant la sélection des mots à traiter interactivement.

Mais l'interface permet aussi de tirer parti de ces marques dans la mesure où il est possible de procéder à une recherche d'unités obéissant à certains critères. Ainsi, en cochant la case *Mots absents du dictionnaire* et en décochant toutes les autres cases, peut-on isoler les mots présents dans le texte mais absents du dictionnaire implanté dans le logiciel. Les mots *inédits* d'un texte de spécialité seraient isolés d'un seul trait, et montrés ensuite dans leurs contextes respectifs. D'un texte du moyen âge, ressortirait le lexique qui lui est propre, pour autant que ce lexique n'a pas été préalablement intégré dans le dictionnaire. Autre exemple : on peut facilement dégager les mots se terminant par *er* et qui sont en même temps des noms.

3. Les usages

Nous utilisons le terme *usage* au sens courant. C'est-à-dire, comme le mentionne le Petit Robert, « Le fait d'employer les éléments du langage, de les réaliser dans le discours; manière dont ils sont employés (opposé à mention) ». Plus spécifiquement, le dictionnaire explique que l'usage est la « Mise en œuvre de

l'ensemble des éléments du langage par la parole; expression verbale de la pensée dans un milieu et un temps donnés ».
Naturellement, le logiciel ne pourrait pas relever l'usage pris dans ce sens (l'acte d'employer), mais le résultat qui en découle. En cela, nous rejoignons l'acception donnée quand on parle de dictionnaire des usages, renvoyant au lexique utilisé. Nous renvoyons au lexique, mais nous renvoyons en même temps au voisinage de ce lexique, le cotexte.
L'utilisateur du programme pourrait s'intéresser à certains usages en partant de distinctions que les linguistes spécifient mais dont les frontières ne sont pas étanches. Selon le dictionnaire de linguistique (Jean Dubois et al. 1973 : 68), on distingue, par rapport à la norme, ce qui relève de l'autorité et ce qui relève du bon usage[29]. Pour invoquer une autorité, on doit référer à certaines personnalités comme des grammairiens notoires, ou des institutions (on peut citer l'Académie Française). Quant au bon usage, Grevisse cite dans son ouvrage qui porte le même nom, un nombre imposant d'auteurs. Il s'agit néanmoins de listes plus ou moins fermées. S'invite dans le débat sur le bon usage, l'appartenance de l'énoncé à une époque ou à plusieurs époques, les archaïsmes par exemple sont exclus du bon usage. (Raison pour laquelle nous comptons intégrer dans la base des usages, l'information sur la source du texte et son époque).
Par ailleurs, la notion de bon usage ou d'autorité, est à mettre en équation avec la notion d'acceptabilité, où il est question, en plus de la grammaticalité de l'énoncé, de règles contextuelles comprenant, entre autres, les « propriétés psychologiques du sujet » (Jean Dubois et al. 1973 : 5), sans oublier, à notre sens, d'autres paramètres dont la pragmatique et les théories de la communication rendent compte. Quand il ne se contente pas de sa propre intuition, le chercheur fait des enquêtes et collecte des données textuelles. L'outil d'extraction que nous proposons apporte une aide dans ce sens.
Mais l'utilisateur du programme peut trouver de l'intérêt non pas dans le bon usage, mais dans ce qui le transgresse. Les énoncés incorrects ou relâchés auront l'attention du grammairien, pendant qu'un chercheur en stylistique pourra être interpelé par des énoncés impurs ou vulgaires.
Un autre type d'énoncés peut recueillir l'attention du chercheur, c'est le discours littéraire insolite et libre, lorsqu'il n'est pas régenté par les règles strictes de la grammaire. Il peut incarner une source de création et donner vie à la langue. Les énoncés rebelles au bon usage seraient à noter autant que le discours dont ils s'écartent, et il est intéressant de noter qu'André Goosse, responsable de la refonte du bon usage, ait intégré, dans son livre, un auteur comme Lautréamont et donné place à des textes que Grevisse lui-même n'avait pas retenus. De manière générale,

[29] Lire la note 3 dans le présent article, en particulier les passages soulignés, pour prendre la mesure d'un commentaire où le Grand Robert mêle une appréciation de l'usage à une recommandation de l'autorité, à propos du verbe *faillir*.

on constate, à lire la préface du Bon usage (Ed. 1994), une ouverture sur des niveaux de langue, des registres, auparavant écartés.
Dans une autre perspective, en langue anglaise, on retrouve deux termes distincts : *use* et *usage*. Widdowson distingue en effet entre *usage* et *use*. Reprenons à Kirstin Malmkjaer, Ed. (1991:458) un extrait de l'auteur : *usage* est défini comme étant « *that aspect which makes evident the extent to which the language user demonstrates his knowledge of linguistics rules* »[30] alors que *use* est vu comme « *another aspect of performance : that which makes evident the extent to which the language user demonstrates his ability to use his knowledge of linguistic rules for effective communication* »[31]. En favorisant ce second concept, Widdowson aurait contribué au développement de l'approche communicative dans l'enseignement des langues (Cf Kirstin Malmkjaer, Ed. (1991:458).
Cela étant dit, il en découle que l'intérêt de la recherche des usages repose sur la perspective dans laquelle se place le chercheur. Voici par exemple, organisées en niveaux d'intervention du programme, quelques questions concrètes que nous nous posons, et pour lesquelles l'outil d'extraction peut apporter une aide :

1- Extraction d'usages relatifs à une unité linguistique ou une combinaison d'unités.

Nous voulons vérifier l'usage du mot gageure / gageüre dans des textes contemporains.

2- Extraction d'usages relatifs à de petites unités linguistiques comme des graphèmes isolés ou combinés, des syllabes ou combinaisons de syllabes, ou des caractères[32]

Nous voulons vérifier quels sont les mots qui contiennent deux voyelles contigües comme dans *idéal* ou *aérien*, cités par André Goosse comme étant des cas de hiatus autrefois répréhensible aux yeux de certains puristes, malgré le caractère poétique des deux exemples cités.

3- Extraction d'usages relatifs à une expression ou à un syntagme : le champ est large, puisqu'à ce niveau c'est toute la grammaire, dans sa dimension syntagmatique, dans ses aléas de construction, qui est mise en jeu.

Nous voulons vérifier la tournure *continuer de* par rapport à la tournure *continuer à*.

---

[30] Passage que nous traduisons par : cet aspect qui montre à quel point le locuteur donne la preuve de sa connaissance des règles de la langue.
[31] Passage que nous traduisons par : un autre aspect de la performance : celui-là même qui montre à quel point le locuteur donne la preuve de sa capacité à utiliser sa connaissance des règles de la langue dans une situation de communication effective.
[32] La perspective de celui qui cherche des caractères peut être différente de celle adoptée par celui intéressé par la recherche des graphèmes. Par ailleurs, sur le plan de la représentation informatique, le caractère ne se confond pas toujours avec un octet. Bien que cela dépasse l'objet de cet article, il nous semble intéressant de disposer d'un outil qui pointerait les caractères en termes de nombre d'octets. Ce qui permettrait d'isoler certains types de caractères.

Nous voulons aussi voir attestée la tournure *Elle l'obligea d'admettre* (plutôt que *à admettre*). Et bien d'autres tours dont la littérature regorge et que le Bon usage a choisi d'épingler avec le signe « ° », dès lors qu'ils sont catalogués comme inusités, et que les linguistes vont affubler du signe « ? », pour peu qu'ils suspectent en eux un trait d'agrammaticalité et du signe « * » lorsque l'énoncé sort du cercle de l'admissibilité.

L'extraction des usages s'apparente à ce que l'on appelle habituellement la recherche des concordances (les logiciels traitant de cette tâche sont dits des concordanciers[33]). Développer de tels outils peut paraître de nos jours comme un défi mineur. Néanmoins, la recherche de concordances se heurte à de multiples questions tant au niveau des formes recherchées qu'au niveau de la nature des mots recherchés, de l'information dont ils sont assortis. Nous avons, quant à nous, axé notre réflexion sur les possibilités de retrouver les occurrences d'un mot, en tenant compte du cotexte, mais nous avons surtout projeté d'utiliser les usages retenus comme une brique dans un processus global qui consiste à marquer les mots et à en emmagasiner les marques pour une utilisation ultérieure. L'intérêt ici est donc double : (1) élargir la gamme des marques apportées aux mots du texte et (2) les rendre disponibles dans une base de données.

Pour le premier aspect, au-delà des marques de base, à savoir le duo lemme-catégorie, nous avons prévu dans le programme la possibilité d'ajouter d'autres marques, en l'occurrence celles incarnées par l'emploi des mots dans le contexte. Ces marques sont bien entendu ouvertes dans la mesure où les usages sont le fruit d'une liberté, même relative, du locuteur, à la différence du cas des catégories grammaticales qui constituent un paradigme restreint.

Pour le second aspect, le cumul des usages des mots collectés alimentera progressivement le dictionnaire électronique. Lequel dictionnaire peut servir à différentes recherches, sachant qu'il se présente sous forme de base de données, laquelle étant d'autant plus malléable que les outils qui la manipulent sont actuellement d'une grande puissance pour retrouver ou filtrer l'information.

Dans ce qui suit, nous évoquerons dans une première section les aspects relatifs à l'extraction des usages, et dans une seconde section, nous parlerons de l'exploitation possible des données extraites.

A. Extraction

Le texte du corpus[34] fait partie d'un projet. Celui-ci répertorie les dictionnaires, et les tables à utiliser. Il est ouvert dans le menu principal. On peut appeler l'interface d'extraction des usages soit depuis l'environnement de marquage, soit de manière autonome depuis le menu principal. Dans les deux cas, on aboutit à l'interface suivante (Voir Figure 4, page 119).

[33] Voir le site http://linguistlist.org/

[34] Nous avons choisi de parler de texte du corpus plutôt que texte cible afin d'éviter une ambigüité possible avec l'usage qui en est fait dans le domaine de la traduction.

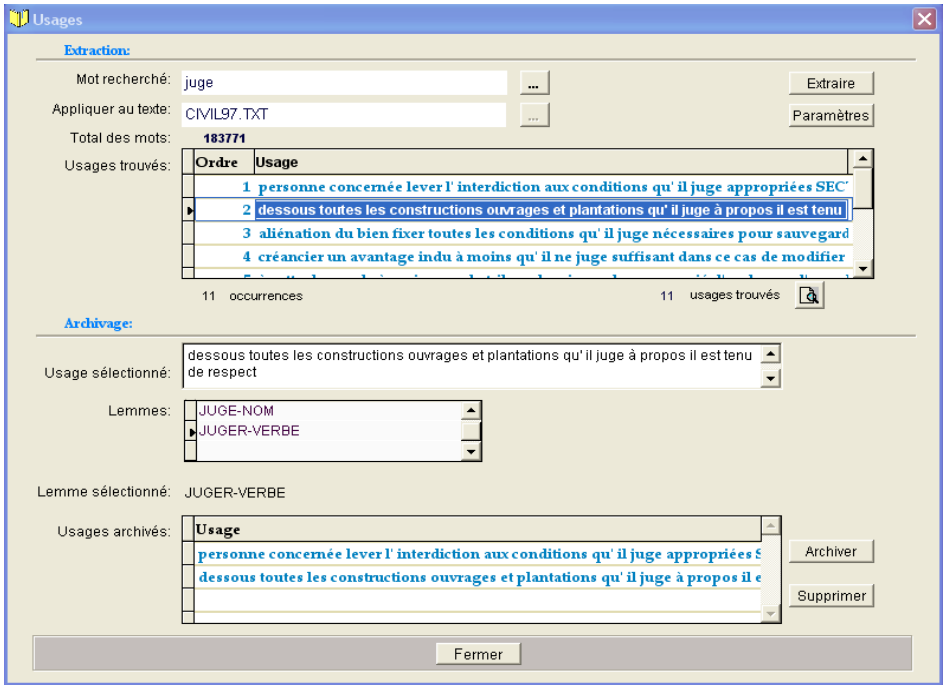


Figure 4 Interface pour extraction des usages.

Quand l'interface est appelée depuis l'environnement du marquage, l'extraction s'applique au mot courant (le mot sur lequel travaille l'utilisateur lors d'une session de marquage). Quand elle est appelée de manière autonome, l'extraction s'applique au mot choisi par l'utilisateur à travers une interface de recherche.
Une fois le mot choisi, l'utilisateur, à moins de vouloir travailler avec les valeurs par défaut, doit déterminer les paramètres d'extraction des usages. Une interface appropriée s'affiche à la demande et permet de sélectionner les valeurs qui correspondent le mieux à ses objectifs.
Ces paramètres sont répartis en paramètres fondamentaux et paramètres complémentaires, comme l'illustre la Figure 5, page 121.

1- Paramètres fondamentaux :

Ces paramètres sont de type morphologique et lexical. Ils interviennent à deux niveaux. Le premier niveau concerne la forme du mot. Le second niveau exploite les marques associées aux mots, quand le texte a déjà été étiqueté. En effet, sachant qu'un mot, qui se présente à nous comme forme, est en même temps assorti de marques donnant entre autres le ou les lemme(s), nous avons jugé intéressant de tirer parti de ces marques et de permettre ainsi de retrouver exclusivement les occurrences dont un lemme particulier est mis en œuvre dans le contexte. Tirer parti des autres marques est aussi envisageable mais cela pourrait faire l'objet d'un développement ultérieur.

Ces nuances, qui concernent le mot et son contexte, sont déclinées dans l'interface des paramètres de la façon suivante :

a. Rechercher un mot quel que soit son contexte (2 cas de figure)

- Retenir l'usage de la forme sélectionnée et seulement cette forme. Les variantes de cette forme seront ignorées.
- Retenir l'usage de toutes les formes associées à un des lemmes (l'utilisateur choisit le lemme qui l'intéresse dans la palette fournie par le programme, l'usage retenu sera celui de toutes les formes associées à ce lemme).

b. Rechercher un mot en tenant compte du contexte.

- L'utilisateur peut spécifier des contraintes sur la forme et sur le lemme du mot, spécifier des contraintes sur le voisinage de ce mot (le cotexte). L'utilisateur devrait dans ce cas exprimer ces contraintes sous forme d'expression rationnelle[35].

2- Paramètres complémentaires :

- Longueur de l'extrait en termes de nombre de caractères. (l'utilisateur qui souhaite manipuler ces données dans une base de données serait intéressé par la dimension du champ à utiliser pour recueillir les usages).
- Longueur de l'extrait en termes de nombre de mots. (l'utilisateur pourrait être intéressé, en fonction de ses hypothèses de travail, par le nombre de mots).
- Tenir compte ou non de la casse du mot recherché. Exemple :

[35] L'expression rationnelle, en anglais *regular expression* est un ensemble de codes utilisés pour spécifier des paramètres de recherche sur un texte. Certains langages comme PERL intègrent de manière native cet outil, d'autres comme xBase dans Visual Foxpro nécessitent un codage supplémentaire. Par ailleurs, citons le programme Linguistica développé par Jacques Ladouceur et présenté à l'Université Ibn Tofail, Kénitra, en 2000. Ce programme présente à l'utilisateur une interface qui tire profit des expressions rationnelles.

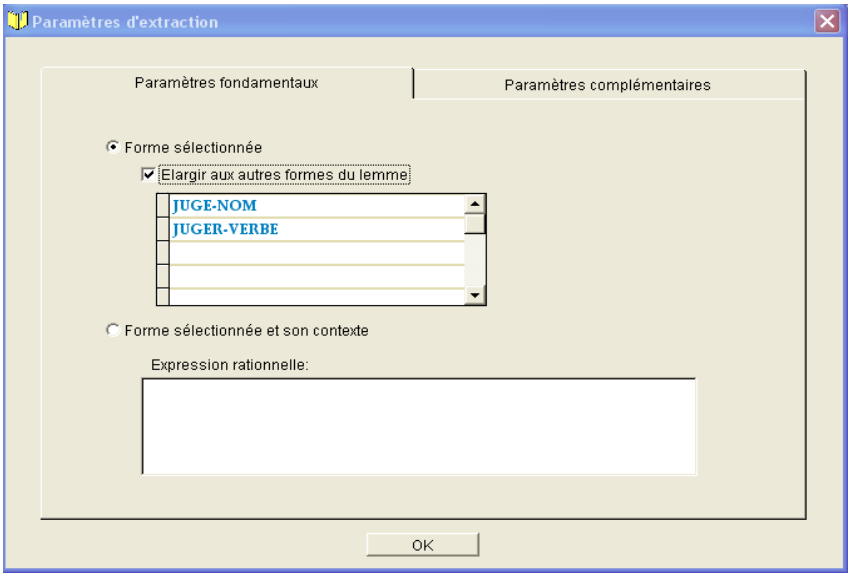
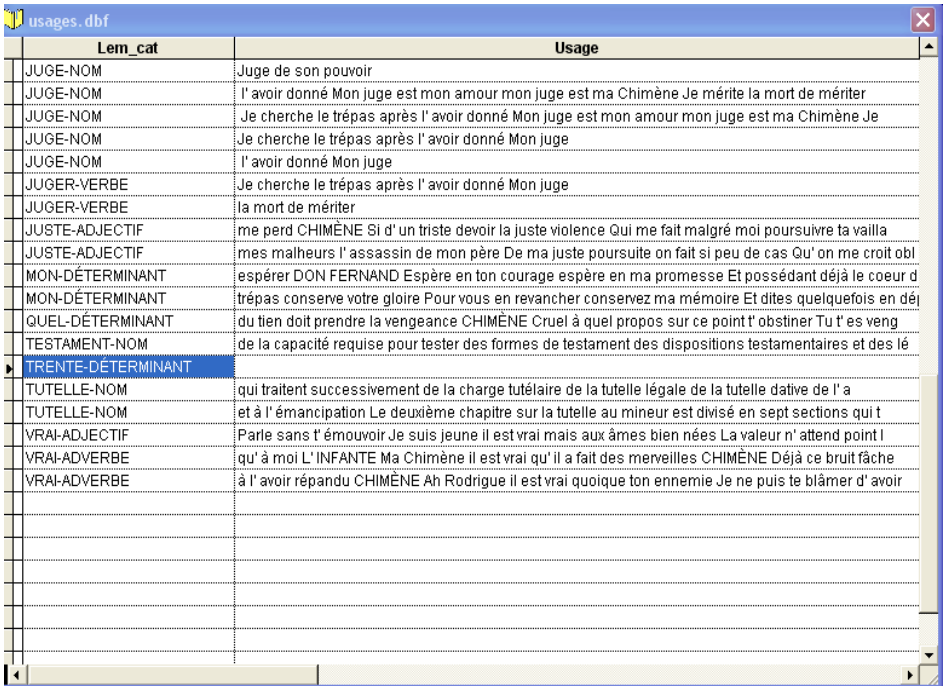


Figure 5 Interface pour les paramètres d'extraction

Le programme ne décide pas de lui-même du bon ou du mauvais usage. C'est un outil qui permet d'extraire du texte des énoncés, en fonction de paramètres précis, et de les disposer dans des listes où l'utilisateur va exercer son expertise pour faire le tri et retenir ce qui serait le plus conforme à ses objectifs.
Mentionnons par ailleurs que l'utilisateur peut ouvrir simultanément deux fichiers texte et procéder à l'extraction. Les usages retenus seront ceux obéissant aux paramètres sélectionnés, mais la vérification se fera pour les deux textes.

B. Exploitation

Les usages extraits sont enregistrés dans le dictionnaire. Cela implique que la consultation du dictionnaire peut apporter pour chaque entrée l'information sur les usages, provenant potentiellement de plusieurs sources.
Le dictionnaire, organisé sous forme de base de données, peut être l'objet d'applications diverses. C'est une base lexicale, structurée selon les principes des bases des données relationnelles. Le fichier qui répertorie les usages peut néanmoins être utilisé de manière autonome, indépendamment des fichiers auxquels il est lié, comme on peut le voir à la Figure 6, page 122. Il est en outre envisageable de convertir cette base en données texte brut ou d'autres formats.
Marqueur permet de marquer les mots d'un texte, les usages étant considérés comme une marque parmi d'autres. Lorsqu'on cherche un mot dans un texte ouvert dans Marqueur, le dictionnaire met à notre disposition toutes les marques qu'il trouve vis-à-vis du lemme dont le mot est assorti.
Le dictionnaire est le dépôt des usages provenant de différentes sources. La cohérence et la signification de ces données brutes appartiennent au chercheur qui choisit son corpus et retient ses critères.



usages.dbf

Lem_cat	Usage
JUGE-NOM	Juge de son pouvoir
JUGE-NOM	l' avoir donné Mon juge est mon amour mon juge est ma Chimène Je mérite la mort de mériter
JUGE-NOM	Je cherche le trépas après l' avoir donné Mon juge est mon amour mon juge est ma Chimène Je
JUGE-NOM	Je cherche le trépas après l' avoir donné Mon juge
JUGE-NOM	l' avoir donné Mon juge
JUGER-VERBE	Je cherche le trépas après l' avoir donné Mon juge
JUGER-VERBE	la mort de mériter
JUSTE-ADJECTIF	me perd CHIMÈNE Si d' un triste devoir la juste violence Qui me fait malgré moi poursuivre ta vailla
JUSTE-ADJECTIF	mes malheurs l' assassin de mon père De ma juste poursuite on fait si peu de cas Qu' on me croit obl
MON-DÉTERMINANT	espérer DON FERNAND Espère en ton courage espère en ma promesse Et possédant déjà le coeur d
MON-DÉTERMINANT	trépas conserve votre gloire Pour vous en revancher conservez ma mémoire Et dites quelquefois en dé
QUEL-DÉTERMINANT	du tien doit prendre la vengeance CHIMÈNE Cruel à quel propos sur ce point t' obstiner Tu t' es veng
TESTAMENT-NOM	de la capacité requise pour tester des formes de testament des dispositions testamentaires et des lé
TRENTE-DÉTERMINANT	
TUTELLE-NOM	qui traitent successivement de la charge tutélaire de la tutelle légale de la tutelle dative de l' a
TUTELLE-NOM	et à l' émancipation Le deuxième chapitre sur la tutelle au mineur est divisé en sept sections qui t
VRAI-ADJECTIF	Parle sans t' émouvoir Je suis jeune il est vrai mais aux âmes bien nées La valeur n' attend point l
VRAI-ADVERBE	qu' à moi L' INFANTE Ma Chimène il est vrai qu' il a fait des merveilles CHIMÈNE Déjà ce bruit fâche
VRAI-ADVERBE	à l' avoir répandu CHIMÈNE Ah Rodrigue il est vrai quoique ton ennemie Je ne puis te blâmer d' avoir

Figure 6 Le fichier des usages est relié à la base des données mais il peut être utilisé de manière autonome.

Conclusion

Nous avons présenté un module d'extraction des usages des mots, intégré à un logiciel de marquage. Utilisé en soi, l'outil répond à des besoins ponctuels tels que le repérage d'occurrences, en tant que formes ou en tant que lemmes, en tenant compte de différentes contraintes relatives au cotexte exigées par l'utilisateur.

Utilisé en association avec la partie marquage des mots du texte, le module d'extraction des usages permet de retenir les résultats en les appariant avec les lemmes répertoriés dans le dictionnaire.

Cette deuxième démarche est au service de tout chercheur en langues qui, dans une perspective de linguistique du corpus, aurait à traiter d'un bon usage ou de ses écarts. Mais le programme est ouvert à d'autres perspectives de recherche. Cette ouverture est à la fois sa force et sa faiblesse. Etant encore en développement, les paramètres relatifs à l'usage étant nombreux, il s'agira d'en faire l'inventaire, d'en hiérarchiser la pertinence, avant de les implémenter.

Nous n'avons pas cherché à placer dans le programme une capacité quelconque à évaluer automatiquement l'usage, comme d'autres ont prévu d'analyser le style. Nous avons simplement placé une recherche quelque peu triviale dans un cadre complexe, celui de l'usage, bon ou mauvais. Le mot d'ordre d'André Goosse est d'observer la langue. Le chercheur a tout à gagner à se placer en observateur de la

production d'un scripteur (voire d'un locuteur) témoin de la langue. Tant mieux si un outil flexible vient à son secours !

Bibliographie

Brill, Eric (1995): *« Transformation-Based Error-Driven Learning and Natural Language Processing: A case Study in Part-of-Speech Tagging »,* dans **Computational Linguistics**, volume 21, numéro 4, p. 543-565.

Dubois Jean, Giacomo Mathée, Guespin Louis, Marcellesi Christiane, Mercellesi Jean Baptiste, Mével Jean-Pierre, (1973) : **Dictionnaire de linguistique**, Larousse.

Ghiglione, R., Landre, Bromberg, M., Molette, P., (1998) **: L'analyse automatique des contenus**, Dunod, 154 p.

Grevisse, (1993), **le Bon usage, Grammaire française**, 13ème édition (refondue par André Goosse), Duculot.

Habert B., Nazarenko A., Salem A., (1997) **: Les linguistiques du corpus**, Armand Colin, 240 p.

Malmkjaer Kirstin, Ed. (1991) : **The Linguistics Encyclopedia**, Routledge, London and New York.

Martin W., Heymans R. et Platteau F., (1988): *« Dilemma, an automatic lemmatizer » dans* Willy Martin (éd.) **Computational Linguistics at the University of Antwerp,** Antwerp papers in Linguistics n° 56.

Mepham Mepham, 1991: *« Efficient Pre-processing for the Creation of Large-scale Full-text Data Bases »* dans **Actes du Colloque RIAO 91**.

Mokhtari Abdelkrim, 2005 : *Marqueur, un logiciel de marquage semi-automatique de textes*, in **Actualité scientifique, Actes des septièmes Journées scientifiques du réseau de chercheurs Lexicologie, Terminologie Traduction**. Ouvrage consacré au thème *Mots, Textes et Contextes*, Edition des Archives Contemporaines, Bruxelles.

Mokhtari Abdelkrim, (1994) : *« Collocateur : Première version d'un logiciel de reconnaissance semi-automatique d'expressions »* dans **Actes des 8èmes Journées de Linguistique,** Ciral, Université Laval, Québec.

Mokhtari Abdelkrim, (1998) : **Cohésion lexicale et automatisation : application à la désambiguïsation lexicale**. Thèse Ph.D., Université Laval, Québec, Canada.

Robert, Paul, (1994), remanié et amplifié par Rey-Debove (Josette) et Rey (Alain), **le Nouveau Petit Robert,** dictionnaire alphabétique et analogique de la langue française, Dictionnaires le Robert, Paris.

Vers un dictionnaire Web de la langue Amazighe

El Mehdi IAZZI[1], Mohamed OUTAHAJALA[2]
[1]mehdiiazzi@gmail.com
[2]outahajala@ircam.ma

Résumé

Cet article est axé sur le contenu linguistique et les aspects informatiques d'un projet en cours de réalisation à l'Institut Royal de la Culture Amazighe concernant l'élaboration d'une base de données (du lexique, la grammaire et des textes) électronique de l'amazighe marocain. L'objectif est de présenter les aspects majeurs des constituants de cette application (lexique, grammaire et textes), ainsi que son architecture, principalement le dictionnaire, et le traitement des variations dans le cadre de la norme linguistique élaborée au Maroc. Nous axerons également notre intervention sur les outils informatiques investis et l'architecture du programme développé.
Afin d'optimiser l'utilisation de la base de données lexicale, cette dernière est conçue de manière à fournir dès le départ (au niveau de l'encodage dont la conception est extensible et modulaire) toutes les informations jugées nécessaires, et ce pour chaque entrée lexicale (entrée normée, prononciation(s), catégorie et sous-catégories grammaticales, nature dérivationnelle, aires géolinguistiques, origine si emprunt ou néologisme, domaines, variations flexionnelles ou supplétives, ...). Ces informations permettront également d'interroger sous différents angles la base de données : relever des ontologies du lexique par domaine (lexique de l'agriculture, lexique de l'artisanat, etc.), des familles dérivationnelles (les mots dérivés de la même base ou racine), etc.

Introduction

L'informatique est devenue un outil d'enseignement dont on ne peut nier l'importance, ainsi il a fait son apparition dans le domaine d'apprentissage et notamment dans celui des dictionnaires. C'est dans ce cadre que s'inscrit le projet base de données de l'amazighe. Cette base de données linguistiques sera publiée sur Web, afin de faciliter l'accès pour différents types d'usagers.

L'objectif de ce projet est d'accompagner l'intégration de l'amazighe dans le système éducatif et les médias marocains en mettant à la disposition des formateurs et des usagers de la langue amazighe une base de données amazighes qui répond à leurs besoins.
Dans un souci de clarté, nous avons décomposé cet article en quatre parties qui seront détaillées comme suit :

- une première partie présentera le contexte et les résultats attendus ;
- une seconde partie développera les besoins linguistiques du dictionnaire qui sont conformes au standard marocain tel qu'il est enseigné dans les écoles du royaume ;
- une troisième partie traitera de la conception de la solution informatique et de sa réalisation.

1. Le contexte de réalisation et les résultats attendus

Dans cette partie, il s'agit de présenter le processus de la standardisation de la langue amazighe ainsi que le contexte de réalisation et les résultats attendus d'une base de données amazighe.

1.1 Standardisation de l'amazighe

Au Maroc, la standardisation de la langue amazighe ne peut se réaliser qu'en adoptant une stratégie réaliste qui prend en considération la variation et la diversité linguistiques. Elle est basée sur un système neutralisant, au niveau graphique, essentiellement les divergences phonétiques entre les parlers dans les limites géopolitiques de l'Etat. Les différences les plus importantes sont dues aux processus de spirantisation des occlusives et les assimilations résultant de la rencontre de certains phonèmes. L'écrit permettra de neutraliser ces faits. Depuis la création de l'institut royal de la culture amazighe (dorénavant IRCAM), un travail important a été réalisé dans le domaine de la normalisation linguistique, de la codification de la graphie tifinaghe dans Unicode et de l'enseignement de l'amazighe au niveau du primaire.

1.2 Contexte et résultats attendus

La lexicographie amazighe publiée jusqu'ici consiste en un ensemble d'ouvrages consacrés aux vocabulaires (glossaires et lexiques, voire même quelques dictionnaires) ou un ensemble de travaux de descriptions grammaticales ou encore des recueils de textes. Ces publications ont les caractéristiques suivantes:

- elles portent sur quelques parlers circonscrits géographiquement. Certaines variantes sont représentées, d'autres moins, voire même non étudiées. Il n'existe pas de travail général regroupant les données de tous les parlers. Les chercheurs qui travaillent dans ce domaine doivent consulter plusieurs publications.
- elles sont éparses et inaccessibles dans la plupart des cas. Certaines publications remontent au XIXe siècle et au début du XXe. Les quelques exemplaires existants ne sont consultables que dans des bibliothèques spécialisées, généralement en France.

- Il n'existe pas de document général regroupant de manière systématique les données de tous les parlers (phonétique, sémantique, morphologique, phraséologique, etc.) (*v*. cas de Chafik).

Pour pallier ses insuffisances et contribuer efficacement à la normalisation de l'amazighe, l'IRCAM œuvre à développer un dictionnaire amazighe en ligne et à le mettre à la disposition des enseignants de l'amazighe et des créateurs, ainsi que les chercheurs et les étudiants.
Les matériaux de ce dictionnaire proviennent, d'une part, des sources documentaires existantes comme les glossaires, les lexiques, les textes et les descriptions grammaticales publiés, et, d'autre part, d'enquêtes et de vérifications sur le terrain. Les sources écrites et publiés feront donc l'objet d'un dépouillement systématique conjugué à des vérifications et à des compléments. Ce travail constituera l'ossature principale de la base de données. Cette œuvre se poursuivra à travers un programme de recueil de l'ensemble lexical des différents usages de l'amazighe et de textes oraux inédits.

2. Besoins linguistiques du dictionnaire

Pour rendre compte des structures du lexique (structures formelles et sémantiques, convergences et divergences intradialectales et interdialectales) et répondre aux besoins des usagers, nous avons retenu les aspects suivants:

- L'entrée lexicale normée selon l'alphabet adopté avec lien vers les règles orthographiques (aoriste/impératif 2ème personne du singulier pour le verbe, singulier de base pour le nom, forme non supplétive pour la préposition, etc.).
- La/les réalisation(s) phonétique(s) régionale(s) [spirantisation, affrication, rhotacisme, effacement, allongement vocalique compensatoire, etc.].
- La racine de l'entrée pour faciliter le regroupement des familles de mots.
- La catégorie grammaticale et les variations morphologiques ou supplétives: Nom commun / nom propre/ nom de parenté …etc. pour le nom ; genre/ nombre/ état pour le nom et l'adjectif ; transitivité - valence et conjugaison pour le verbe ; variations supplétives pour la préposition ; valeurs du déterminant/ de la préposition /de l'adverbe/ de la conjonction, etc.
- Si le mot est dérivé, la nature de la dérivation est fournie (causatif, réciproque, réfléchi, passif etc. pour le verbe; et nom d'action, nom d'agent, nom de lieu, nom d'instrument, nom d'état) ainsi que la base de la dérivation.
- Les indications géolinguistiques: il s'agit de préciser les zones dialectales et les parlers où le mot est attesté (e.g. tarifit - Aït Iznassen, Ikebdanen, Tamsaman, Igueznayen, etc.-; tamazight – Aït Ndhir, Zemmour, Aït Mguild, Izayanen, etc.-; tachelhit – Ihahan, Ida Outanane, etc.).
- Sources de la lexie: document publié (précisions sur la référence: auteur,

date d'édition, page) ou collecte (donner informateur / lexicographe).

- Si le mot est un emprunt, les précisions sur la langue source sont également données (phénicien, hébreu, latin, arabe, français, espagnol, etc.) et le mot d'origine.
- En cas de néologisme, la ou les sources et la référence complète sont précisées.
- Le ou les sens de l'entrée en amazighe et la ou les variantes régionales de l'amazighe où chaque sens est attesté (avec indications sur la ou les source(s): document ou informateur/ lexicographe).
- Le ou les synonymes en amazighe au niveau intradialectal et interdialectal, avec un lien vers chaque synonyme.
- Le ou les homonymes au niveau intradialectal et interdialectal, avec un lien vers chaque homonyme.
- Le ou les antonymes au niveau intradialectal et interdialectal, avec un lien vers chaque antonyme.
- Les équivalents en arabe, en français, en anglais, etc.
- Le ou les domaines d'usage (agriculture, anatomie, architecture, structures sociales, etc.).
- Le ou les doublets de l'entrée avec des précisions sur les variantes régionales où chaque doublet est attesté ainsi qu'un lien vers l'entrée de chaque doublet.
- Contextes phraséologiques et locutions figées.
- Illustrations iconiques : photos, croquis, etc. (principalement pour les arts, les métiers, les techniques, les éléments de l'environnement naturel comme la faune et la flore).

3. Conception et réalisation de la solution informatique:

Pour réaliser cette application, nous avons utilisé le langage de modélisation UML. Les acteurs de l'application sont: le lexicographe, le modérateur, l'administrateur et l'anonyme.

a. Diagramme des cas d'utilisation

Ce diagramme montre les acteurs de l'application et ses cas d'utilisation envisageables dans sa version actuelle. Comme le montre la figure 1, l'utilisateur anonyme est le seul acteur qui ne s'authentifie pas, il peut faire des recherches simples ou bien avancées (en faisant des recherche par mot, par domaine, par racine, par parler…etc.).
Le lexicographe peut gérer les lexies qu'il a crée (ajouter, supprimer ou éditer des lexies qui n'ont pas été encore validées par le modérateur), le lexicographe remplit plusieurs formulaires représentant la morphologie ou bien la forme supplétive, le ou les sens du mot, le ou les doublets…etc. Il peut également commenter les lexies

créées par les autres lexicographes, dans le but d'améliorer les définitions octroyées aux entrées lexicales.
Le modérateur, quant à lui, valide les lexies créées par les lexicographes. Il peut également modifier ou augmenter les entrées saisies, en mettant à jour la morphologie, le ou les sens…etc.
Enfin, l'administrateur gère les informations sur les utilisateurs. Le système de son coté vérifie les informations rentrées et valide ou bien refuse les informations saisies.

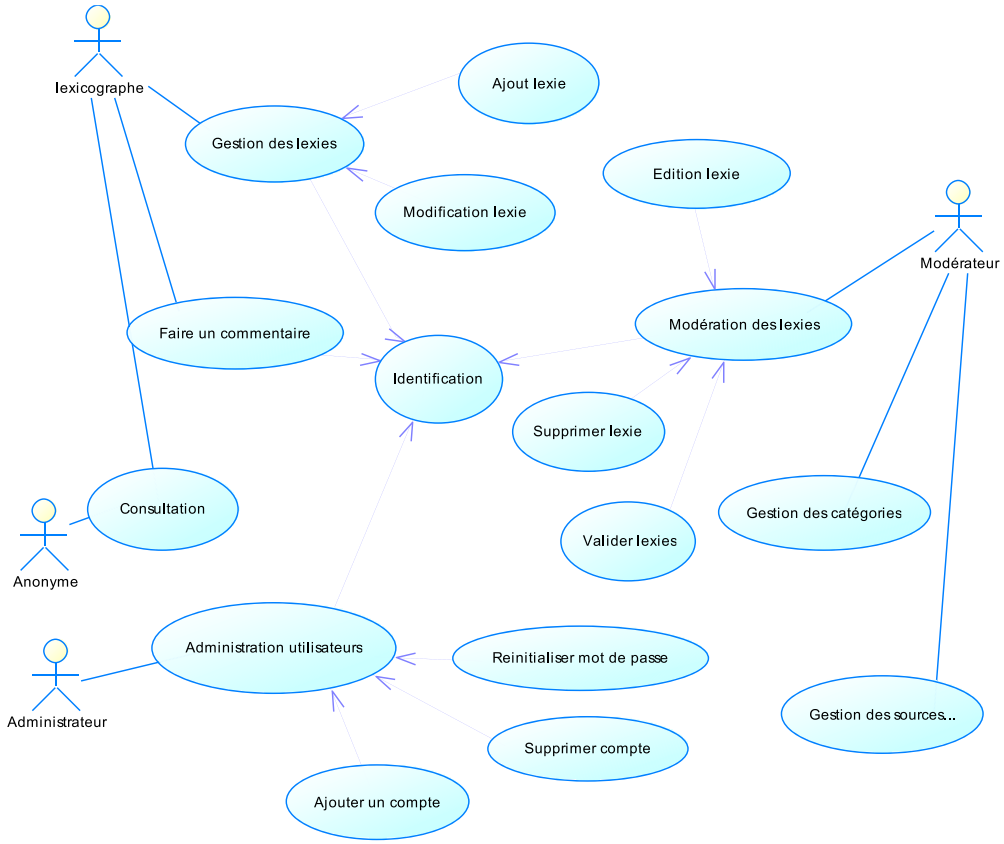


Figure 7: Cas d'utilisation

b. Diagramme des classes de l'application:

Chaque classe du diagramme ci-dessous a des attributs et des méthodes, par exemple la classe lexie a pour attributs par exemple: lexie, racine, source_lexie, emprunt, langue d'emprunt, néologisme, source_neologisme…etc. et comme méthode ajout_lexie, maj_lexie pour mettre à jour une lexie et supprimer_lexie pour la suppression d'une lexie. La classe lexie est associée à plusieurs autres

classes comme le montre la figure 2, par exemple la classe utilisateur, la classe morphologie.etc.

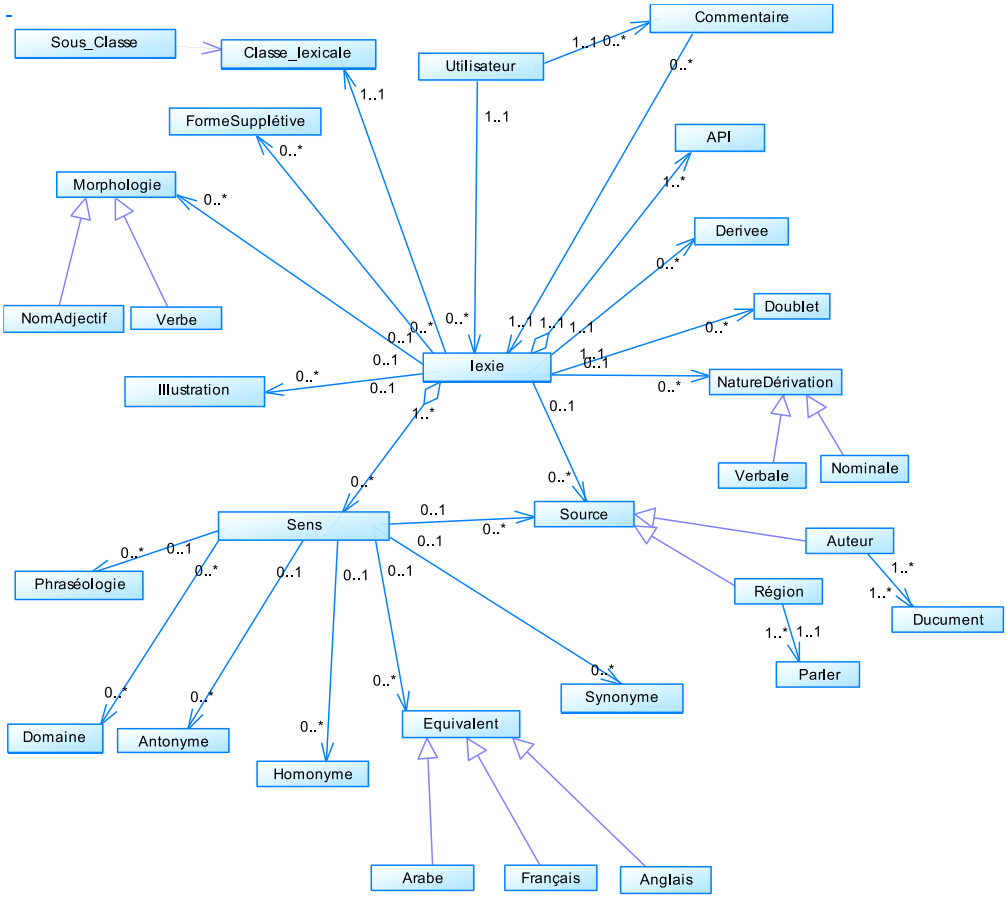


Figure 8: Diagramme de classe

Afin de réaliser cette application, nous avons utilisé la technologie .net et le serveur de base de données SQL Server 2000 pour la persistance des classes de cette application. Elle tourne actuellement sur un serveur IIS 5.0 qui tourne sous Windows 2000. Cette application et en phase de validation et elle est installée actuellement en deux versions une pour la collecte et la consultation des entrées lexicales de l'IRCAM et l'autre version pour la collecte et la consultation des entrées lexicales relatives au projet "Modes de production et de transmission de la culture dans les sociétés berbères" du programme FSP-Maghreb.

Figure 3: instance de l'application pour les entrées lexicales de l'IRCAM

Figure 4: instance de l'application pour le projet du programme de FSP-Maghreb

Comme le montre les figures 3 et 4, les deux projets ont des interfaces différentes avec un même schéma de la base de données. Le projet FSP-Maghreb contient un peu plus de deux mille entrées lexicales.

Références

- Ameur, M., Bouhjar, A., Boukhris, F. Boukouss, A., Boumalk, A., Elmedlaoui, M., Iazzi, E., Souifi, H. (2006a), Initiation à la langue Amazighe. Publications de l'IRCAM.
- Ameur, M., Bouhjar, A., Boukhris, F. Boukouss, A., Boumalk, A., Elmedlaoui, M., Iazzi, E. (2006b) Graphie et orthographe de l'Amazighe. Publications de l'IRCAM.
- Boukhris, F. Boumalk, A. El moujahid, E., Souifi, H. (2008). La nouvelle grammaire de l'Amazighe. Publications de l'IRCAM.
- Iazzi, E.M. et Outahajala, M. (2008). Amazighe Data Base, Marrakech LREC08.

Demain, encore plus de tifinaghes sur Internet

Patrick Andries
Conseils Hapax, Québec, Canada
Membre du consortium Unicode
patrick@hapax.qc.ca

Résumé. Lors de cette communication, nous nous pencherons sur des aspects de l'Internet où les tifinaghes sont actuellement peu présents : les adresses de domaine internet, les adresses de courriel, les identifiants XML et les polices téléchargeables par les navigateurs.
In this paper, we will review areas in Internet where tifinagh characters are currently underrepresented: Internet domain names, email addresses, XML identifiers and dynamically downloadable browser fonts.

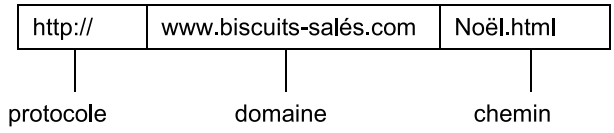
Mots-clés. Unicode, tifinaghe, ISO 10646, amazighe, informatique, réseaux, jeux de caractères, normalisation, polices, XML, noms de domaine, IDN, NDI, courriel, messagerie, touareg, arabe, navigateurs, internet.

1. Introduction

Il existe aujourd'hui encore plusieurs domaines d'Internet dans lesquelles il est difficile d'utiliser les tifinaghes, voire impossible. Nous allons nous pencher sur certains de ceux-ci et expliquer les évolutions les plus récentes dans ces domaines et les difficultés rencontrées qui limitent l'utilisation des tifinaghes dans ces applications.

2. Noms de domaine internationalisés (NDI)

Il est désormais permis d'avoir des adresses 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 divise 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), elle est régie par les lois canadiennes. Les domaines de tête génériques *(.com*, *.org*, *.biz)* sont administrés aux États-Unis en vertu des lois américaines.

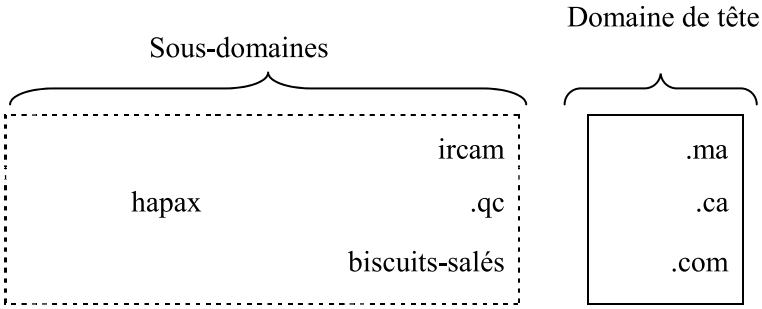


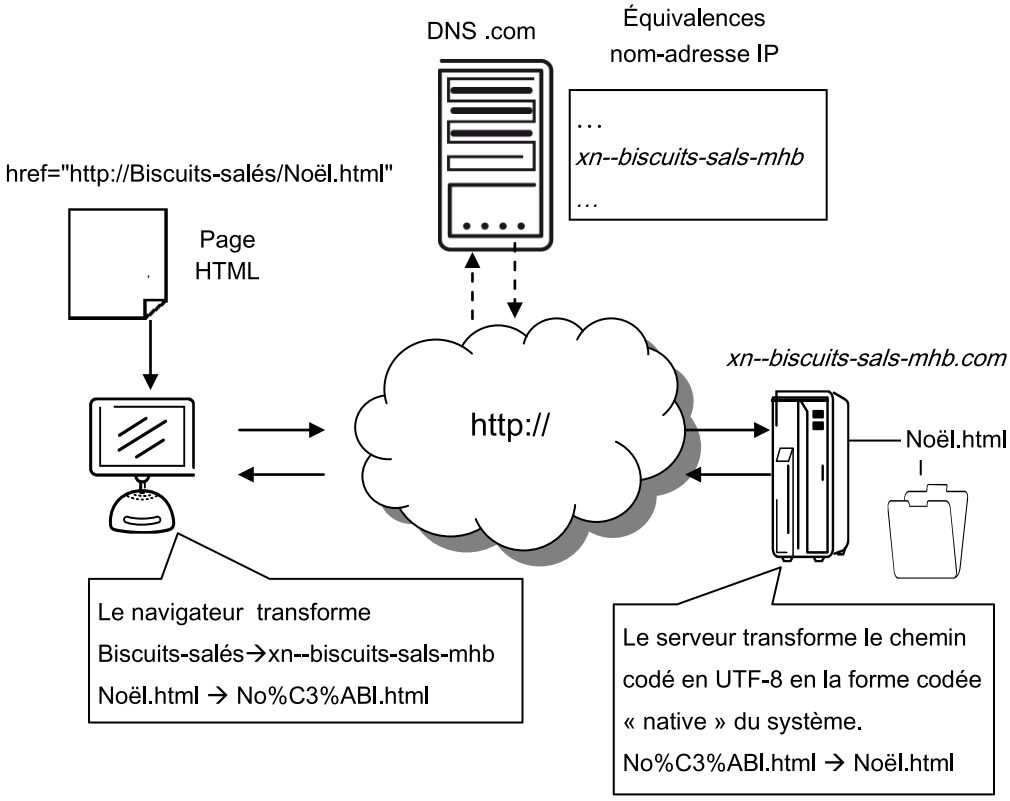
Figure 9.Parties d'un nom de domaine

Les navigateurs modernes transformeront de manière transparente les adresses de domaine comme *biscuits-salés* en des adresses de domaine compatibles avec l'infrastructure actuelle prévue pour des adresses ASCII. L'utilisateur ne se rendra habituellement pas compte des transformations effectuées en coulisse par le protocole appelé *punycode*. Dans notre cas, ce protocole transforme `biscuits-salés` en `xn-biscuits-sals-mhb` afin de pouvoir réutiliser toute l'infrastructure internet actuelle qui ne permet que l'utilisation d'ASCII. C'est sous ce nom étrange que sera enregistré le nom de domaine internationalisé auprès de l'autorité responsable pour l'attribution des sous-domaines pour un domaine de tête particulier (l'ACEI au Canada par exemple). Et c'est sous cette forme que le nom sera diffusé aux différents serveurs de noms de domaine (les DNS) qui effectueront le passage entre le nom d'un domaine et son adresse IP qui n'est composée que de chiffres. Lors d'une demande de chargement d'une page HTML, seuls les

navigateurs verront donc le nom de domaine internationalisé, tout le reste de l'infrastructure ne verra que de l'ASCII, sous la forme *punycode* si le nom de domaine contenait des caractères non ASCII. C'est ce qu'illustre la figure 2.

Figure 2.Transformation des IRI

Pour le codage du chemin de la ressource sur le serveur (à savoir `Noël.html`), le mieux est d'adopter le RFC 3987 qui préconise l'utilisation d'UTF-8 comme



codage des chemins avant leur transformation en notation `%hh` et leur transmission sur le réseau : `Noël.html` devient donc `No%C3%ABl.html`.

La prise en charge des NDI peut augmenter le risque de parodie ou d'hameçonnage (« phishing ») par l'entremise de caractères similaires à ceux auxquels s'attend l'utilisateur, mais dans un autre système d'écriture. Une adresse ainsi contrefaite pourrait, par exemple, utiliser un « o » cyrillique (U+043E) à la place d'un « o » latin (U+006F) ! (Ces deux caractères ont le plus souvent une apparence identique.)La stratégie la plus souvent employée pour détecter ces adresses contrefaites consiste pour les navigateurs à ne pas permettre l'emploi dans une

même partie du nom de domaine délimitée par un « . » de caractères provenant d'écritures différentes (pas de caractères latins et cyrilliques pour reprendre notre exemple dans un même segment). Les autorités d'enregistrement de noms mettent en œuvre des règles similaires pour éviter tout risque d'usurpation : elles sont souvent plus draconiennes encore que les navigateurs. C'est ainsi que l'autorité suisse d'enregistrement n'acceptera que les caractères latins utilisés par ses langues nationales au sein des noms de domaine qu'elle octroie. Toutefois, comme on l'a vu, ces noms de domaine pourront comprendre des accents et la cédille.

Malheureusement, en 2008, seul un sous-ensemble des caractères d'Unicode peut être utilisé dans les noms de domaine internationalisés. Ils doivent tous faire partie du répertoire d'Unicode 3.2, or ce répertoire ne comprend aucun caractère tifinaghe puisque cette écriture n'a été incluse que dans la version 4.1 d'Unicode.

Toutefois, une nouvelle version de la norme qui régit les caractères permis dans les NDI est en préparation (IDNA2008) elle devrait permettre d'utiliser tous les caractères tifinaghes d'Unicode.

3. Adresses de courriel internationalisées

Il existe un domaine voisin des noms de domaine internationalisés (NDI) où, pour l'instant, la prise en charge des caractères Unicode est incomplète. Il s'agit de celui des adresses de courriel internationalisées.

C'est ainsi que s'il est possible – pour autant que vos programmes prennent en charge les NDI – d'avoir des adresses de courriel de ce type :

pandries@hapax.qc.ca

hsu.wang@环球门邮.info

Il n'est pas possible à l'heure actuelle d'avoir des adresses où les caractères à la gauche de l'arrobe (@) sont des caractères accentués ou appartenant à une écriture non latine comme dans les exemples ci-dessous :

françois.côté@biscuits-salés.com

测试@环球门邮.info

En effet, les adresses de courriel internationalisées ne sont pas compatibles avec le standard SMTP de base. Ainsi, la commande RCPT TO qui sert à préciser le destinataire ne permet que des caractères ASCII. Il faut donc étendre SMTP pour permettre l'utilisation de UTF-8 dans les commandes MAIL FROM et RCPT TO. Cette extension est l'objet du standard expérimental RFC 5336, elle se signale dans le protocole SMTP par l'envoi de l'option UTF8SMTP, dont l'usage permet de vérifier que le serveur de transmission de courriels avec lequel correspond votre logiciel comprend bien la nouvelle norme. Si ce n'est pas le cas, l'échange s'effectuera à l'aide d'adresses en ASCII.

Mais voyons à quoi ressemble un échange entre un client (votre logiciel de messagerie) et un serveur messagerie. Le client (C) appelle le serveur (S) dans le dialogue ci-dessous :

```
S: 220 amy.cooptel.qc.ca ESMTP
C: EHLO courriel.orange.be
S: 250-amy.cooptel.qc.ca
 250-ENHANCEDSTATUSCODES
250-8BITMIME
250-UTF8SMTP
C: MAIL FROM:<françois.côté@biscuits-salés.com> ALT-
   ADDRESS=fcote@biscuits-au-sel.com
S: 250ok
C: RCPT TO: <测试@环球卡邮.info>
S: 250 ok
```

Dans la première ligne, le serveur s'annonce et donne son adresse (amy.cooptel.qc.ca). Ensuite, à la ligne suivante, le client salue le serveur (EHLO) et fournit son nom de domaine (courriel.orange.be). Plus loin, le client précise qu'un message de François Côté doit être envoyé à 测试@环球卡邮.info.

Notez que les deux adresses comprennent des caractères nonASCII, à gauche et à droite de l'arrobe (@). Le MAIL FROM : contient deux adresses de l'expéditeur : une internationalisée et une autre ASCII de repli qui se conforme au protocole SMTP de base. L'adresse du destinataire peut également comprendre un paramètre ALT-ADDRESS pour préciser une adresse patrimoniale ne contenant que des caractères ASCII.

Le RFC 5336 précise qu'un agent de transfert de courriers peut annoncer qu'il accepte UTF8SMTP signalant ainsi à son correspondant que celui-ci peut lui envoyer des adresses UTF-8. C'est ce qui se produit ci-dessus : le serveur envoie un « 250-UTF8SMTP » au client pour l'informer que ce dernier peut envoyer des adresses nonASCII.

Le RFC 5336 définit égalementla stratégie que doit adopter un serveur lorsqu'il tente de transmettre un message internationalisé à un ancien agent de transfert (un autre serveur) qui ne gère pas cette extension. Plusieurs solutions de repli sont envisagées et permises, la plus fréquente étant probablement le « déclassement » précisé dans le RFC 5504. Ce déclassement vers l'ASCII utilisera, entre autres, le paramètre ALT-ADDRESS précisé ci-dessus dans le seul cas de l'expéditeur.

4. Identificateurs XML

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

```
<?xml version="1.0" encoding="UTF-8"?>
<texte xml:lang="ber" lang="ber">
   <h1>ⵜⴰⵎⴰⵣⵉⵖ ⵜⴰⵎⴰⵣⵉⵖ</h1>
   <p>ⵉ ⵡⵜⵜⵡⵉⵏ ⵅⵓⵙ ⵙⴰⵢ, ⵉⵓⵏⵖ ⵜⵜⵡⵜⵜⵓ ⵜⵙⵉⵉ ⵉⵏⵜ ⵙⵅⵜⵓⵏ: ⵙⵏⵙ ⵉⵅ ⴰⴰ
   ⵙⴰⴰ ⵖⵓ ⵙⵇⵣⵣⵖⵓ, ⵙⵎⵎⵜ ⵢⵜⵉ ⵜⵏⵜ ⵏ ⵙⵜ ⵜⵜⴰⴰⵓⵜ !</p>
</texte>
```

Il existe toutefois une restriction dans les caractères admis dans les noms d'éléments et d'attributs ainsi que les valeurs d'attributs des fichiers XML 1.0. En effet, XML 1.0 ne permet pour ces noms que des caractères appartenant au répertoire d'Unicode 2.0. Ceci signifie que ces noms ne peuvent contenir des caractères tifinaghes, ou d'autres provenant d'écritures comme l'éthiopien également introduit 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.

Ceci signifie qu'en XHTML ceci n'est pas permis :

```
<a href="#□ □ □" name="□ □ □" id="□ □ □">attribut éthiopien</a>
```

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

```
<□ □ □>...</□ □ □>
```

Pour dépasser cette limite imposée par Unicode 2.0, il faut que vos documents soient conformes à XML 1.1 ou XML 1.0 (5^{e} édition).

5. Polices de navigateur téléchargeables dynamiquement

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. Cela est un peu moins vrai dans la future version de MS Windows, Windows 7, un nouveau système d'exploitation qui comprendra d'office une police Ebrima qui permet d'afficher le tifinaghe.

Toutefois, même après l'introduction de Windows 7 prévue pour la mi-2009, le problème demeurera entier, car il existera 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 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.

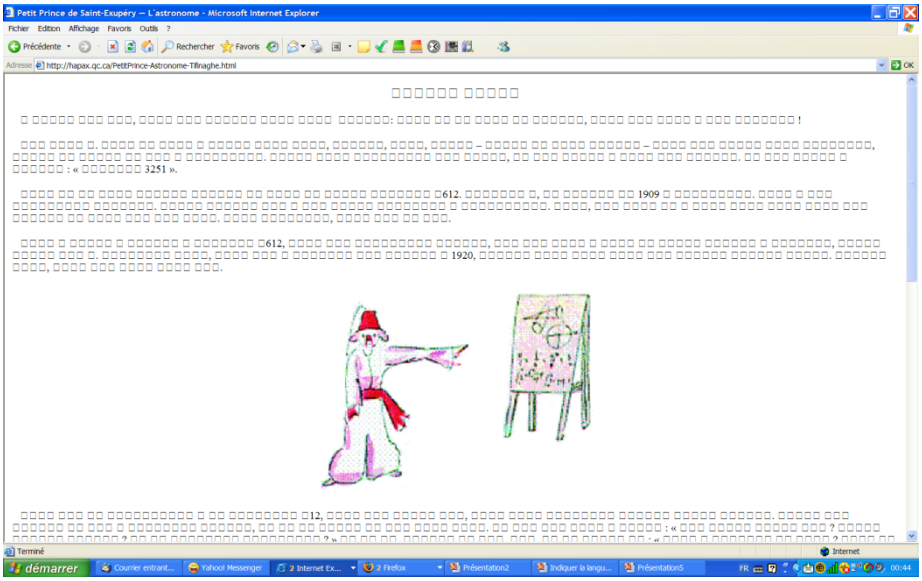


Figure 3.Police manquante

Il existe plusieurs remèdes possibles. 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.

5.1 Police téléchargeable avec Internet Explorer

Les polices téléchargeables ne sont pas une nouveauté : Internet Explorer 4 le permettait déjà en 1997. Comme nous le verrons, ci-dessous, ce qui a changé c'est le fait que d'autres navigateurs Internet offrent désormais cette fonctionnalité. Mais revenons à Internet Explorer.

Microsoft offre un programme qui se nomme WEFT et qui permet d'analyser, d'une part, une page HTML et, d'autre part, une police TrueType incorporable. C'est-à-dire une police dont l'auteur permet qu'elle accompagne un document.

Sur la base de ces deux sources, WEFT produit une police téléchargeable (dont le suffixe est .eot) et un extrait de feuille de style CSS qui fait référence à cette police .eot. L'illustration ci-dessous décrit graphiquement ce processus.

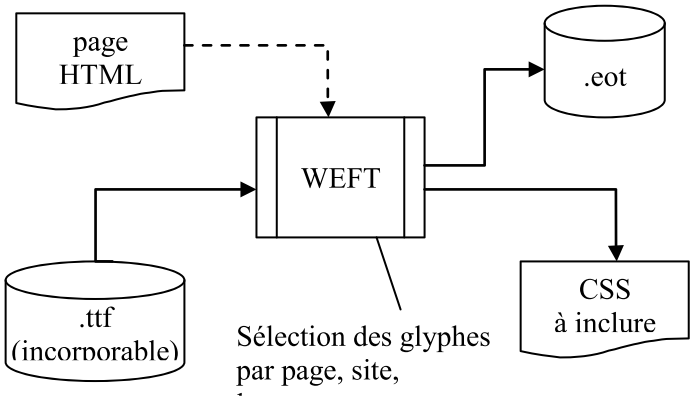


Figure 4.WEFT de Microsoft

Le fichier .eot ne doit pas nécessairement comprendre tous les glyphes de la police TrueType fournie en entrée, c'est le plus souvent inutile et inefficace. WEFT permet ainsi de sélectionner les seuls glyphes nécessaires à l'affichage d'une page HTML ou d'un ensemble de pages.

La police `.eot` est chiffrée et idéalement ne comprend donc que le sous-ensemble de glyphes nécessaire. Pour des raisons de sécurité, le fichier `.eot` produit est lié à une liste de serveurs explicites et ne peut être utilisés qu'à partir de ceux-ci. Le plus gros désavantage lié à ce mécanisme est cependant le fait que ces fichiers ne fonctionnent que sur Windows. L'extrait de code ci-dessous représente le code CSS que WEFT produit et qui doit être inclus dans les fichiers HTML fournis en entrée pour qu'ils invoquent la police `.eot` stockée.

```
<style type="text/css">

        @font-face {
            font-family: "Ajoure-eot";
            font-style:  normal;
            font-weight: normal;
            src: url(http://hapax.qc.ca/polices/AJOURE0.eot);
          }

          body { font-family: "Ajoure-eot";}
```

```
</style>
```

Une fois la page HTML de la figure 3 modifiée pour inclure ce code CSS, le texte tifinaghe apparaîtra correctement dans Internet Explorer pour peu que la police `.eot` mentionnée dans la partie `src :url()` du `@font-face` soit bien présentée à l'adresse mentionnée.

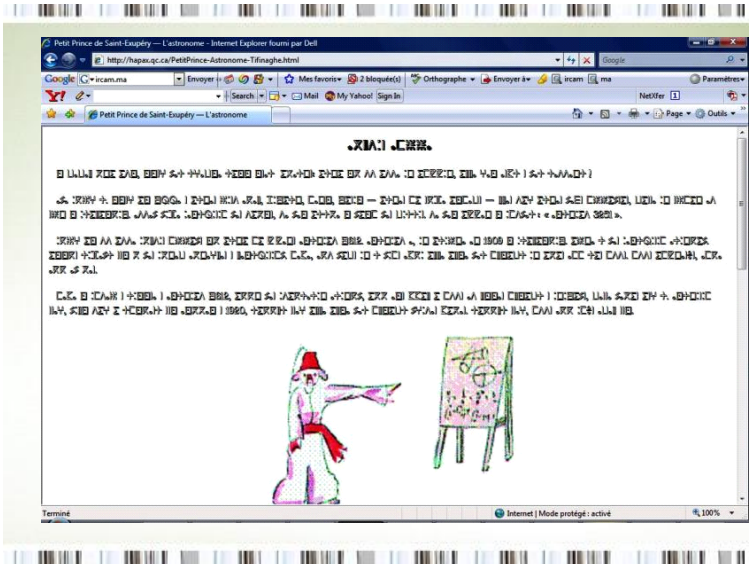


Figure 5.*Affichage de la police tifinaghe avec MS Explorer*

5.3 Les autres navigateurs

Malheureusement, les autres navigateurs internet comme Firefox, Opera, Safari ou Chrome ne comprennent pas les fichiers `.eot` dont le format est propriétaire.

Toutefois, `@font-face` permet également de préciser le chargement d'une police TrueType (.ttf) ou OpenType (.otf). Les dernières versions de Firefox 3.1 (Windows) et Safari 3.1 (sur Mac et Windows) prennent en charge `@font-face`.

Il est dès lors possible d'écrire une feuille de style qui fonctionne sur plusieurs navigateurs grâce à du code comme celui que l'on retrouve ci-dessous :

```
<style type="text/css" >
   @font-face {
font-family: "Ajoure-ttf";
      font-style:  normal;
      font-weight: normal;
      src: url(http://hapax.qc.ca/polices/ajoure.ttf);
}
```

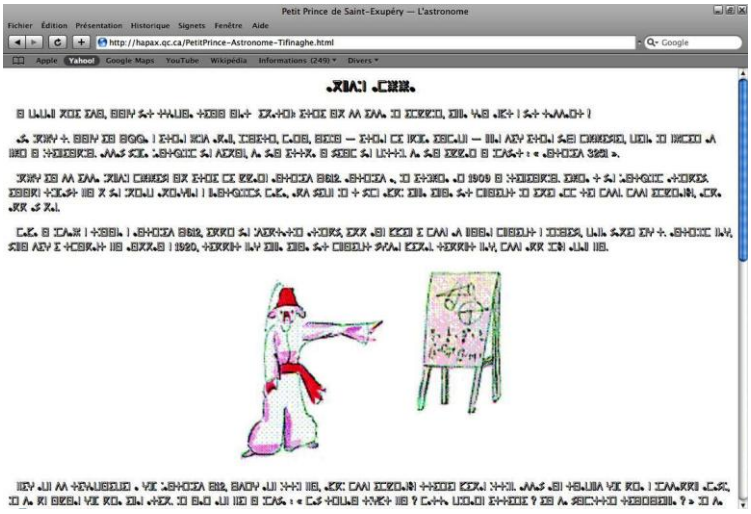
```
   @font-face {
     font-family: "Ajoure-eot";
      font-style:  normal;
      font-weight: normal;
src: url(http://hapax.qc.ca/polices/AJOURE0.eot);
   }

   body {font-family: "Ajoure-ttf","Ajoure-eot";}
   h1 {font-size: 16pt; text-align:center; space-before: 12pt;
   space-after: 12pt;}
p   {text-indent:9pt;}

</style>
```

Dans un premier temps, on trouve deux `@font-face`. L'un pour les fureteurs internet qui prennent en charge le chargement des polices TrueType, l'autre pour MS Explorer qui met en œuvre les polices « .eot ».

Ensuite, il suffit de préciser en regard du sélecteur du corps complet du document (`body {…}`), les deux polices téléchargeables équivalentes, l'une sous le format `.ttf` et l'autre sous le format `.eot`. Le navigateur choisira la première police disponible où il trouvera les glyphes nécessaires. Dans le cas de Safari ce sera dans la police `.ttf`, pour ce qui est de Microsoft Explorer ce sera dans la police `.eot`.

Grâce à ce code CSS et après avoir hébergés les polices correspondantes à l'adresse référencée par leur `src :url()` correspondant, on peut désormais afficher la même page HTML tifinaghe dans Safari. Et cela, sans que l'utilisateur n'ait besoin d'installer au préalable la bonne police tifinaghe.

Figure 6.*Affichage de la police tifinaghe avec Safari*



5.3 À plus long terme…

Pour l'instant, les polices TrueType sont envoyées d'un bloc alors que les polices `.eot` peuvent n'envoyer que les glyphes nécessaires à l'affichage de la page HTML correspondante, allégeant d'autant la transmission de la police incorporée.

Le W3C considère actuellement un mécanisme ouvert de même type que `.eot`, il devrait être valable pour toutes les plateformes, il sera associé à des pages d'un site, pourra produire des sous-ensembles de glyphes sélectionnés et comprimer ces polices téléchargeables.

6. Conclusion

Depuis 2005 et la normalisation du tifinaghe 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. Il reste malheureusement quelques obstacles à leur emploi généralisé comme nous l'avons vu dans le cas des noms de domaine internationalisés, en XML, dans les adresses courriels et dans l'affichage de pages sur Internet quand une police tifinaghe manque, ce qui est encore souvent le cas.

Toutefois, dans chacun de ces domaines, les choses avancent et des solutions techniques se dessinent pour que l'utilisation des tifinaghes dans tous ces domaines se fasse désormais sans entraves et cela principalement grâce à l'inclusion de cette écriture dans Unicode.

7. Remerciements

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 cette communication 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.

8. Bibliographie

Andries, P. (2008*). Unicode 5.0 en pratique*,Dunod éditions, Paris.

Bortzmeyer, S. (2008), *RFC 5336: SMTP extension for internationalized email address*, disponible à <http://www.bortzmeyer.org/5336.html>.

Yao, J. et Mao W. (2008), *RFC 5336, SMTP Extension for Internationalized Email Addresses*, disponible à < http://www.ietf.org/rfc/rfc5336

يطرح إدماج اللغات في منظومة التكنلوجيات الحديثة عدة تحديات نظرا لطبيعة مجال المعلوميات من جهة، والخصائص الصرفية والنحوية للغة من جهة أخرى.

وفي سياق اللغة الأمازيغية، تم فتح ورشات مهمة في ميادين البحث، خاصة البحوث التي تتعرض لنظام كتابة الأمازيغية، تيفيناغ، وكذا معالجتها في البرمجيات وأنظمة التطوير. ويتطرق الكتاب الذي بين أيدينا إلى مختلف الإشكالات التي تم طرحها ونقاشها خلال الندوة الدولية TICAM'08 في نسختها الثالثة والتي تمحورت حول "وضعية وآفاق الأمازيغية في تكنولوجيات الإعلام". وتسعى هذه المقالات إلى التعريف بالمجهودات المبذولة من طرف الباحثين لحل إشكالات حوسبة الأمازيغية.

L'intégration des langues dans les technologies de l'information comporte des défis inhérents aux contraintes des technologies informatiques, et aux spécificités morphosyntaxiques de la langue. Dans le cas de la langue amazighe, un chantier de recherche important traite l'adaptation de son système d'écriture Tifinaghe et sa mise en forme par rapport aux logiciels et plateformes de développement.

Le présent ouvrage rassemble les travaux qui ont été exposés pendant la 3éme Edition du colloque TICAM'08 sur le statut et les opportunités de l'amazighe dans les technologies de l'information. Il met en exergue les efforts d'intégration de l'amazighe dans les nouvelles technologies de l'information émergentes et prometteuses.

20 DH