Actes de colloque / Proceedings

Informatisation de l'amazighe

Théories, applications et perspectives

2ème **SITACAM 2011**

Symposium International sur le Traitement Automatique de la Culture Amazighe

Coordination
Ali Rachidi
Youssef Ait ouguengay

Informatisation de l'Amazighe : Théories, Applications et Perspectives

Actes / Proceedings

Informatisation de l'Amazighe : Théories, Applications et Perspectives

SITACAM'11

06 – 07 Mai 2011

Ecole Nationale de Commerce et de Gestion
Agadir – Maroc

En Partenariat avec

L'Institut Royal de la Culture Amazighe
Rabat – Maroc

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

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° 28

Titre	:	Traitement Automatique de la Culture Amazighe - Informatisation de l'Amazighe : Théorie, Applications et Perspectives
Coordination	:	Ali Rachidi et Youssef Ait ouguengay
Éditeur	:	Institut Royal de la Culture Amazighe
Couverture	:	Unité d'édition - CTDEC
Dépôt légal	:	2012 MO 1393
ISBN	:	978-9954-28-116-1
Imprimerie	:	El Maarif Al Jadida – Rabat
Copyright	:	© IRCAM

Comité d'Honneur

A. BENNANI
Président de l'Université Ibn Zohr
Agadir - Maroc

A. Boukous
Recteur de l'Institut Royal de la Culture Amazighe
Rabat – Maroc

Co-Présidents du Comité d'Organisation

A. Bendou
Directeur de l'Ecole Nationale de Commerce et de Gestion
Agadir – Maroc

A. Rachidi
Enseignant Chercheur à l'Ecole Nationale de Commerce et de gestion
Agadir – Maroc

Comité d'organisation

O. **Barakat**, ENCG, Agadir
F. **Bouchri**, ENCG, Agadir
C. **Cherkaoui**, ENCG, Agadir
Y. **Es Saady**, Faculté des Sciences, Agadir

Comité de Programme

Y. Ait Ouguengay, CEISIC, IRCAM, Rabat, Maroc
J. Antoine, LI, Université de Tours, France
A. Barakat, Faculté des lettres, Agadir, Maroc
M. Bellafkih, INPT Rabat Maroc
K. Bensoukas, Faculté des Lettres Rabat Maroc
E. Bouyakhf, Facultés des sciences, Rabat, Maroc
B. Chawki, ENSA, Agadir, Maroc
C. Cherkaoui, ENCG, Agadir, Maroc
D. Chiadmi, EMI, Rabat, Maroc
H. Douzi, Faculté des sciences, Agadir, Maroc
Y. El kirat, Faculté des Lettres, Rabat Maroc
M. Elyassa, Faculté des sciences, Agadir, Maroc
B. Erraha, ENSA, Agadir, Maroc
M. Hassoun, ENSSIB, France
M. Iazzi, Faculté des lettres, Agadir, Maroc

Mot Du Comité D'Organisation

Après le grand succès de la première édition du SITACAM organisé par l'ENCG d'Agadir- Maroc, les 12-13 Décembre 2009 (SITACAM'09) et qui avait rassemblé plus de 100 participants de différentes nationalités, le comité d'organisation organise la deuxième édition du Symposium les 06 – 07 mai 2011 sous le thème :

« Informatisation de l'Amazighe : Théories, Applications et Perspectives »

Outre le cadre scientifique visé par le comité d'organisation, un cadre très convivial est préparé confirmant ainsi la réputation de la ville d'Agadir comme terre d'accueil et d'hospitalité.

Nos remerciements s'adressent à tous les organismes qui ont soutenu l'organisation de ce symposium.

Il faut enfin rendre hommage aux membres du comité d'honneur, du comité d'organisation et du comité scientifique pour leur aide et leur dévouement.

Le Comité D'organisation

Mot Du comité de Programme

Le comité de programme de la deuxieme édition de Symposium International sur le Traitement Automatique de la Culture Amazighe (SITACAM'11), est très fier de présenter ces actes qui, nous l'espérons, intéresseront la communauté scientifique internationale en technologie de l'information et la linguistique.

Nous avons reçu plus de Cinquante articles de 4 pays différents. Ces articles ont été transmis aux membres du comité de programme pour subir une évaluation rigoureuse. Les papiers acceptés portent sur les thèmes majeurs : Reconnaissance de forme et plus précisément le caractère amazighe, Génie logiciel, Linguistique et Analyse lexicale et syntaxique de la langue amazighe.

Nous remercions sincèrement tous les membres du comité de programme pour leur excellent travail d'évaluation, de même tous les autres collègues qui ont aidé à l'expertise des papiers.

Nos remerciements à tous les membres du comité d'organisation et aux sponsors qui ont contribué au succès de SITACAM'11.

Le Comité De Programme

TABLE DES MATIERES

Session 1

Session 2

Session 3

Session 4

Session 5

Convertisseur pour la langue amazighe : script arabe - latin - tifinaghe

Ataa Allah Fadoua, Boulaknadel Siham

Centre d'Etude Informatique,
Systèmes d'Information et de Communication
Institut Royal de la Culture Amazighe
Avenue Allal El Fassi, Madinat Al Irfane, Rabat – Instituts ;
Adresse postale BP 2055, Hay Riad, Rabat
{ataaallah, boulaknadel}@ircam.ma

Résumé – Abstract

L'amazighe en tant que langue et culture a connu un processus de standardisation et d'intégration dans les nouvelles technologies de l'information et de la communication, qui est passé par plusieurs étapes : un codage spécifié par l'ASCII étendu, un codage propre dans le standard Unicode, l'élaboration des normes appropriées concernant la disposition du clavier amazigh et la création des polices de caractères tifinaghes. Dans ce contexte, le présent article a pour objectif de fournir à la langue un outil de conversion lui permettant de translittérer l'alphabet arabe et latin en caractère tifinaghe Unicode.

The Amazighe language and culture have experienced a process of standardization and integration in the information and communication technologies. This process has gone through several stages: character encoding specified by the extended ASCII, incorporation into the Unicode standard, implementation of a standard keyboard layout for Amazighe, and building new Tifinaghe fonts. In this context, we aim to provide the Amazighe language a conversion tool that could transliterate the Arabic and Latin scripts to Tifinaghe Unicode.

Keywords – Mots Clés

Convertisseur, translittérateur, Langue amazighe, Promotion, Diffusion.
Converter, transliterator, Amazighe language, Promotion, Dissemination.

Introduction

Depuis la création de l'Institut Royal de la Culture Amazighe (IRCAM), la langue amazighe jouie d'un statut institutionnel qui lui a permis d'avoir une graphie officielle et un codage propre dans le standard Unicode. Or dans le contexte de sauvegarder l'héritage littéraire de l'amazighe, et empruntant la politique entreprise par l'IRCAM visant la diffusion de la graphie tifinaghe sur tous les supports médiatiques et pédagogiques, il est intéressant de tirer profit de ce patrimoine culturel par l'élaboration d'outils informatiques permettant la réécriture des documents amazighes.

Cet article s'inscrit dans une démarche progressive qui vise à doter la langue amazighe d'applications capables de traiter de façon automatique les données, particulièrement les outils de conversion. Ainsi, nous avons procédé par le développement d'un outil de translittération qui consiste à convertir l'alphabet arabe en alphabet tifinaghe en partant de l'idée qu'il existe des documents en langue amazighe qui sont écrits en caractères arabes qui peut être exploitable, et en se basant sur l'outil de conversion développé au sein du Linguistic Data Consortium.

Dans la suite de cet article, nous exposons, dans la section 2, les caractéristiques de la langue amazighe à travers la description du système d'écriture et le codage employé. Ensuite dans la section 3, nous présentons notre proposition et une conception de l'outil développé en partant de l'existant. La section 4 est consacrée à la conclusion.

Caractéristique de la langue amazighe

L'amazighe est la langue autochtone de l'Afrique du Nord (Hachid, 200; Charles-André, 1978). Cette langue regroupe une trentaine de variétés dialectales présentes de la Méditerranée jusqu'au sud du Niger, et des Iles Canaries à la frontière ouest de l'Egypte. Particulièrement au Maroc, l'amazighe se répartit en trois grandes variétés régionales qui couvrent l'ensemble des régions montagneuses : le Tarifite au nord-est ; le Tamazighte au centre, le Moyen-Atlas et une partie du Haut-Atlas ; et le Tachelhite au sud et sud-ouest, le Haut-Atlas, l'Anti-Atlas et Sous.

Système d'écriture de la langue amazighe

Avant la standardisation de la langue amazighe, son écriture ainsi que sa modélisation informatique faisait appel à l'alphabet arabe ou l'alphabet latin enrichi par les caractères spéciaux empruntés de l'Alphabet Phonétique International (API). Or, dans l'objectif de fournir à la langue amazighe un système

alphabétique standard plus adéquat et utilisable pour tous les parlers amazighs actuels du Maroc, le Centre de l'Aménagement Linguistique (CAL) a préconisé les caractères tifinaghes, présentés dans le Tableau 1, comme alphabet de la langue amazighe (Ameur et al., 2004). Cet alphabet comporte :

- 27 consonnes dont : les labiales (ⵀ, ⵖ, ⴼ), les dentales (ⵜ, ⴷ, ⵟ, ⵠ, ⵍ, ⵏ, ⵇ, ⵅ), les alvéolaires (ⵙ, ⵣ, ⵥ, ⵛ), les palatales (ⵚ, ⵊ), les vélaires (ⴽ, ⵅ), les labiovélaires (ⴽⵯ, ⵅⵯ), les uvulaires (ⵇ, ⵅ, ⵖ), les pharyngales (ⵄ, ⵃ) et la laryngale (ⵀ);
- 2 semi-consonnes : ⵢ et ⵡ ;
- 4 voyelles : trois voyelles pleines ⴰ, ⵉ, ⵓ et la voyelle neutre (ou schwa) ⴻ qui a un statut assez particulier en phonologie amazighe.

Pour la ponctuation, les signes conventionnels de l'écriture latine sont employés : « » (espace), « . », « , », « ; », « : », « ? », « ! », « … », etc. Alors que pour les chiffres, les numéros arabes occidentaux (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) sont retenus.

ⵜⵉⴼⵉⵏⴰⵖ ⵏ ⵜⵣⵍⵎⴰⵢ			
	Tifinaghe	Correspondance latine	Correspondance arabe
ya	ⴰ	A	ا
yab	ⴱ	B	ب
yag	ⴳ	G	گ
yag^{w}	ⴳⵯ	g^{w}	گ‘
yad	ⴷ	D	د
yaḍ	ⴹ	**ḍ**	ض
yey	ⴻ	E	
yaf	ⴼ	F	ف
yak	ⴽ	K	ک
yak^{w}	ⴽⵯ	k^{w}	ک‘
yah	ⵀ	H	ه
yaḥ	ⵃ	**ḥ**	ح
yaɛ	ⵄ	E	ع
yax	ⵅ	X	خ
yaq	ⵇ	Q	ق
yi	ⵉ	I	ي
yaj	ⵊ	J	ج
yal	ⵍ	L	ل
yam	ⵎ	M	م
yan	ⵏ	N	ن
yu	ⵓ	U	و
yar	ⵔ	R	ر
yaṛ	ⵕ	**ṛ**	ڕ
yaɣ	ⵖ	Γ	غ
yas	ⵙ	S	س
yaṣ	ⵚ	**ṣ**	ص
yac	ⵛ	C	ش
yat	ⵜ	T	ت
yaṭ	ⵟ	**ṭ**	ط
yaw	ⵡ	W	ۏ
yay	ⵢ	Y	ێ
yaz	ⵣ	Z	ز
yaẓ	ⵥ	z	ژ

Tableau 1 : L’alphabet Tifinaghe préconisé par le CAL et consacré par l’IRCAM

Le codage du système amazighe

Dans le but d'introduire le système d'écriture amazighe préconisé par l'IRCAM, il a fallu en attendant l'amendement de l'Unicode trouver une solution rapide qui permettra la production des publications, particulièrement les manuels scolaires. Cette solution était le codage ASCII étendu représentée par la table de codage illustré à la Figure 1, où les caractères latins ont été substitués par leur correspondant en tifinaghe.

Figure 1. Codage du Tifinaghe en ASCII étendu

Or après l'amendement, le consortium Unicode a réservé aux caractères tifinaghes le bloc illustré par la Figure 2, qui précise les codes réservés à chacun des quatre sous-ensembles des caractères tifinaghes. Le premier sous-ensemble représente les lettres alphabétiques de base préconisées par l'IRCAM dont le nombre est de 33. L'Unicode ne code directement que 31 caractères et le caractère modificatif « ⵯ », qui permet de former les deux unités labiovélaires « ⵅⵯ » et « ⴽⵯ ». Le deuxième sous-ensemble contient les 8 caractères de la liste étendue, qui a été définie par l'IRCAM pour l'intérêt historique et scientifique. Le troisième sous-ensemble est formé de 4 lettres néo-tifinaghes utilisés fréquemment dans le reste du Maghreb. Et le quatrième sous-ensemble contient 11 lettres touarègues modernes dont l'usage est attesté (Andries, 2008).

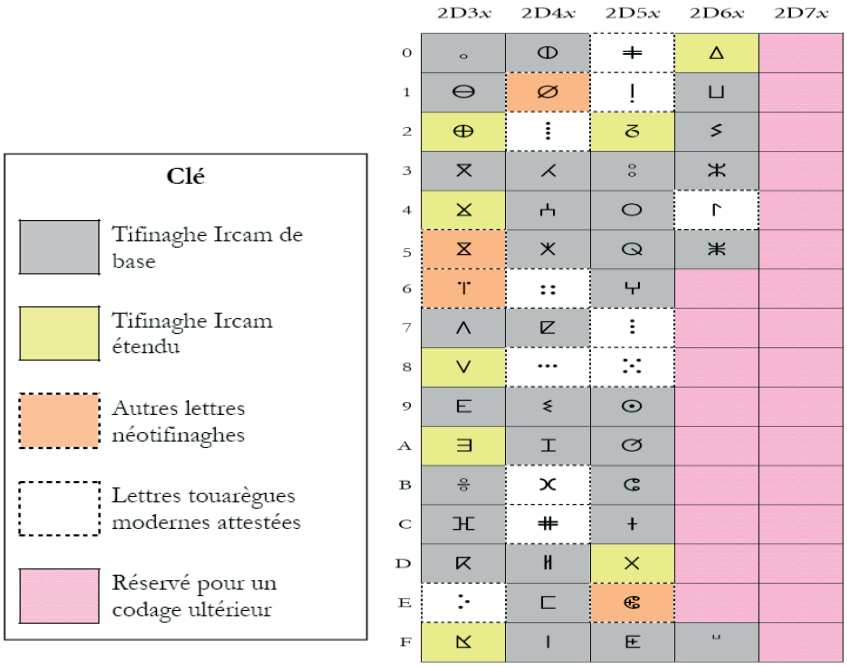


Figure 2. Codage du Tifinaghe en Unicode

Convertisseur pour la langue amazighe

Dès le codage des caractères tifinaghes, il a fallu préparer des outils pour convertir la représentation de ces caractères d'un codage à un autre. Dans ce cadre, le Centre d'Etudes Informatiques Système d'Information et de Communication (CEISIC) ainsi que le Linguistic Data Consortium (LDC) ont réalisé chacun de sa part un convertisseur permettant le passage entre le codage ASCII étendu et le codage Unicode. Alors que dans la perspective de promouvoir la langue amazighe et d'assurer la sauvegarde de son héritage littéraire, il apparaît intéressant de tirer profit des productions amazighes écrites en caractère arabe. Ainsi, nous avons proposé de compléter le convertisseur de LDC afin qu'il puisse traiter les données écrites en caractère arabe.

Convertisseur de LDC

Le convertisseur de LDC est un outil qui permet d'une part la conversion du caractère tifinaghe codé en ASCII étendu (IRCAM Latin) au caractère latin sous le codage Unicode (Unicode Latin), et d'autre part la conversion du caractère Unicode Latin au caractère tifinaghe codé en Unicode (Unicode Tifinaghe). En outre, il permet le passage direct du IRCAM Latin au Unicode Tifinaghe.

Proposition

Dans la perspective de doter la langue amazighe d'un outil de tranlittération des documents écrits en caractère arabe, nous avons proposé de compléter le convertisseur de LDC d'une manière à répondre à d'autres besoins qui pourraient être lucides dans l'avenir. Ainsi, nous avons opté pour ajouter d'autres fonctionnalités à cet outil afin qu'il assure la conversion entre les différents codes du caractère tifinaghe et la translittération dans les différents sens entre le caractère arabe, latin et tifinaghe.

Conversions existantes	Nouvelles conversions
IRCAM Latin → Unicode Latin	Unicode Tifinaghe → IRCAM Latin
Unicode Latin →Unicode Tifinaghe	Arabe → Unicode Tifinaghe
IRCAM Latin → Unicode Tifinaghe	Unicode Tifinaghe → Arabe
	Arabe → IRCAM Latin
	IRCAM Latin → Arabe

Tableau 2. Types de conversions assurées par le nouvel outil

Conception du convertisseur

L'outil développé assure des fonctionnalités qui consistent à convertir les différents codes du caractère tifinaghe et translittérer les textes amazighs écrits en caractère latin ou arabe au caractère tifinaghe et vice versa, comme il est illustré par la Figure 3.

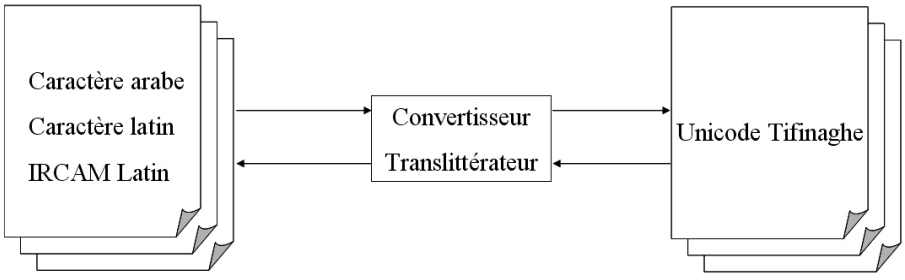


Figure 3. Conception de l'outil de conversion

Pour la translittération des documents écrits en arabe, nous nous sommes basées sur les correspondances utilisées dans le dictionnaire arabe – amazighe (Chafik, 1993), qui différent de celles indiquées dans le Tableau 1 par celles précisées dans le Tableau 3.

	Tifinaghe	Correspondance arabe
Yag	ⵅ	ڭ
yag^{w}	ⵅⵓ	ڭ
yak^{w}	ⴽⵓ	ڪ
Yi	ⵉ	ـئـ (au début d'un mot (
Yu	ⵓ	ؤ (au début d'un mot (
yaṛ	ⵕ	ر
Yay	ⵢ	) يau début d'un mot (
yaẓ	ⵥ	ڒ

Tableau 3. Correspondances distinctes établies par Mr Chafik

Au cours de notre implémentation, nous avons remarqué que les caractères { ڪ } ڒ ne figurent pas sur la table Unicode, ce qui nous a amené à remplacer « » par le caractères «ک ٗ» en s'inspirant de la proposition de Mr Chafik pour le caractère « ڭ» qui correspond au caractère tifinaghe « ⵅⵓ », et à substituer le caractère « » par «ڒ» vu leur ressemblance. Par ailleurs, en se basant sur les correspondances arabes du caractère tifinaghe pour translittérer les documents écrits en caractère tifinaghe vers celui en arabe, nous avons soulevé quelques anomalies à propos des caractères {ي ,ڕ ,ک ,گ} qui influence sur la cursivité de la langue arabe. Ce qui nous a amené à les remplacer par les propositions de Mr Chafik.

Tandis que pour la translittération du caractère latin au caractère tifinaghe, nous avons adopté les correspondances illustrées dans le Tableau 1.

Conclusion

Le présent article s'inscrit dans une stratégie progressive qui vise à encourager le développement d'outils de traitement automatique de la langue amazighe, particulièrement l'élaboration d'outils informatiques permettant la conversion des documents amazighes. Ainsi, nous avons entrepris de réaliser un outil de translittération qui consiste à convertir l'alphabet arabe et latin en alphabet tifinaghe en exploitant les documents en langue amazighe qui sont écrits en caractères arabes et latins, et en se basant sur l'outil de conversion développé au sein du Linguistic Data Consortium.

Références

Ameur M., Bouhjar A., Boukhris F., Boukous A., Boumalk A., Elmedlaoui M., Iazzi E., Souifi H. (2004), *Initiation à la langue amazighe*, Rabat, Maroc, IRCAM.

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

Chafik M. (1993), *المعجم العربي الأمازيغي*, El Maarif Aljadida, Rabat, Maroc.

Charles-André J. (1978), *Histoire de l'Afrique du nord des origines à la conquête arabe: Tunisie - Algérie – Maroc*, Editions Payot.

Hachid M. (2000), Les premiers berbères: entre méditerranée, Tassili et Nil, Provence, France, Edisud.

Amazighe Transliteration

M. EDDAHIBI[1], S.Mouhim[2], C.Cherkaoui[3], D.Mammass[4]

1,2,3,4 Laboratoire IRF-SIC, Faculté des sciences & ENCG
B.P.28/S – Agadir – Maroc.

eddahibi@yahoo.fr, mouhimsanaa@yahoo.fr, ccherkaoui@yahoo.fr, driss_mammass@yahoo.fr

Résumé – Abstract

La nécessité d'utiliser une translittération s'impose dans les technologies d'information et de communication (TIC),quandon ne dispose pasd'un environnement pour l'Amazighe. Cet article présente une translittération orthographiqueet une autre phonétique. La première va aider à surmonter les limitations des systèmes informatiques en attendant l'internationalisation des TIC.La deuxièmeest conçue pour une meilleur prononciation et compréhension des mots. L'application de translittération peut être utiliséeen ligne de commande ou à travers une interface.

The need to use a transliteration is a must in information and communication technology (ITC) without support toAmazighescript. This paper presents a system thatallows the use of both spelling-based and phonetic-based transliterations. The former will help to overcome technical limitations; the later can be used for words better understanding and pronunciation. The transliteration is designed to be used in command line or through a user interface.

Keywords – Mots Clés

Amazighe, translittération phonétique, translittération orthographique, Unicode, TIC.

Amazighe, phonetic-based transliteration, spelling-based transliteration, Unicode, ITC.

Introduction

Transliteration is a mapping fromone writing system to another. It associates alphabetical elements of a source language toalphabetical elements combinations of a target language. Machine transliteration process is used in several multilingual applications, like machine translation, text-to-speech synthesis and information retrieval. The transliteration operation is a need in human communication inenvironments without support for user's language or in heterogeneous environments.Heterogeneous communication environments refers to systems where one of the communicating users havedifficulties to input or display text.

Somedocument typesettingand text processing systems, like the well-known TEX system (Banouni, 2002), requires enteringtransliterated plain text. Transliteration can also be used to helpAmazighelanguage learners to better pronounce words.According to the context; the transliteration can be involved as an intermediate step or obtained as a finalresult especially in language learning process.The need of transliteration is in most cases due to the lack of Amazighe script support in ICT environments. That means technically, that the ICT environment didn't address issues related to Amazigheencoding systems.
There are two transliteration categories: phonetic-based transliterations and spelling-based transliterations. The mapping tables are different in the two categories. Even for the same category, transliteration tables are different due to cultural and geographical factors. For example, the letter "j" is pronounced differently in Spanish and French which are the two main foreign spoken languages in morocco.
The work released in this paper is a Romanization system that allows choosing the transliteration concept according to the context.

Transliteration and Amazighe language

Despite the fact that Amazighe is an ancient language, the Amazighe alphabet used is very recent (Rachidi, 2005). For this reason several valuable Amazighe documents were written using Arabic and French transliterations. Today, Amazighe documents are still transliterated. Especially, edocuments are romanized due to ICT environments limitations. Another reason is that the most of Amazighe speaking peoples don't know Amazighe alphabet in spite of government's efforts for teaching Amazighe.

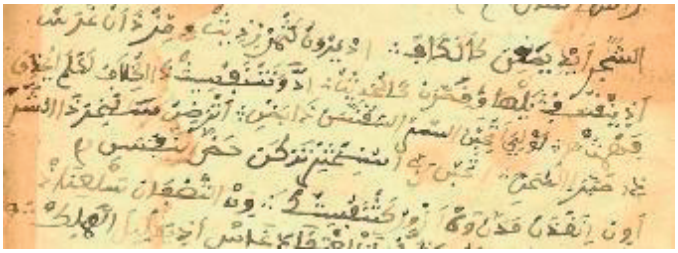


Figure 1: Amazighe text transliterated to Arabic

As long as, several ICT systems lack support for Amazighe script, efforts to overcome all limitations aren't sufficient to avoid the use of transliterations.All the non-Latin-based script languages suffer from these problems, because multilingualism isn't among the main goals of ICT systems. Multilingual support is in most cases added after application implementation by using plug-ins or add-ons.

Encoding problems

Encoding system is a table of characters that can be rendered using font's glyphs or processed using a processing system (Haralambous, 2007). There are twotypes of characters: linguistic and logical characters. Linguistic character is a unit of information that corresponds to a grapheme. It represents letters, numerical digits, and punctuation marks. Currently, Amazighe alphabet is encoded using Unicode character set.

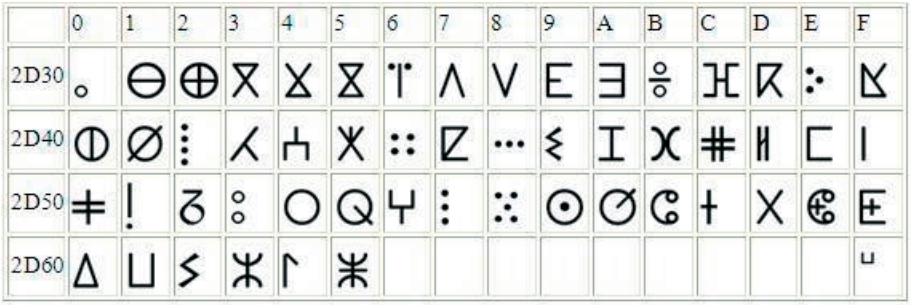


Figure 2: Unicode Planfor IRCAM Tifinagh characters

To support Amazighe, ICT systems should support Unicode. Text rendering engines that supportsAmazigheshould properly displayUnicode characters. To display text, rendering engines should convert characters to glyphs that are displayed on the screen. In the absence of the appropriate font, the text is displayed as empty boxes or isn't displayed at all. In some cases, characters aren't recognized and consequently replaced by question marks. When the rendering engine default character encoding isn't Unicode, characters are confused with other

characters.Some document composition tools replace Unicode characters by their XML character entity.

Original text	Display problems
ⴰⴼⵓⵙ ⵓⵔ ⵉⵜⵜⵓⵔⵓⵏ ⴰⵟⴰⵔ ⴰⴳ ⵢⴰⵎⵓ	□□□□ □□ □□□□□□□ □□□□ □□ □□□□
	No text displayed
	???? ?? ??????? ???? ?? ????
	ÿþ 0-<-S-Y- S-T- I-\-\-S-T-S-O- 0-_-0-T- 0-3- b-0-N-S-
	ⴰⴼⵓⵙⵓⵔⵉⵜⵜⵓⵔⵓⵏⴰⵟⴰⵔⴰⴳⵢⴰⵎⵓ

Figure 3: Amazighe text display problems

In some cases,in spite of the support of Unicode, it isn't the best solution; say for example, in Short Message Service (SMS).The text communication service component of phone imposes text size limitation.
Currently, in GSM environments, an SMS message can contain up to 140 bytes of text data. When text message contains characters encoded using the GSM 7-bit character set (Whistler,2009), 160 7-bit characters are compressed into 140 bytes to produce the usual 160 character limit.
*SMS Max Size = 140*8 bits = 160*7 bits*
If the sent message contains characters that are not part of the GSM 7-bit character set, such as Amazighe, then the text needs to be encoded in Unicode UCS-2 format where each message is limited to 70 16-bits characters.
*SMS Max Size = 70*16 bits = 160*7 bits*
Using a transliteratedGSM 7-bitstext is more efficient in this case than Unicode original text.

Transliteration concepts

The Amazighe writing system is an alphabetical writing system. That means that both spelling-based and phonetic-based transliterations are feasible. The phonetic-based transliteration uses phonemes which are the smallest unit of the speech (Karimi, 2007). This transliteration category uses the common phonetic representation to find the mapping between source and target words. It has some limitations that hinder an accurate phonetic correspondence such as missing sounds in the target language. The phonetic-based transliteration isn't direct because grapheme in the source word should be transformed into a phoneme in the source

language that is mapped to a most similar phoneme in the target language which is finally transformed intographemes in the target language.

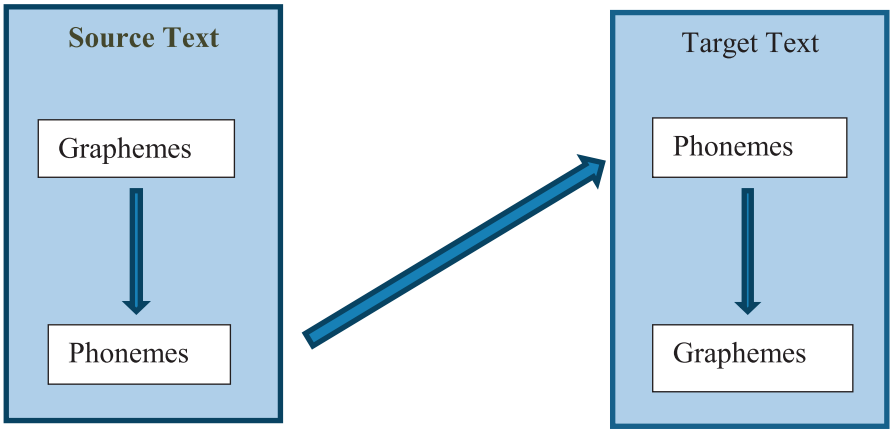


Figure 4: Phonetic-based transliteration

The spelling-based transliteration uses graphemes that are the fundamental units of the writing system. In the Amazighe language the graphemes are alphabet letters. This method is based on a one to one mapping between the source and the target alphabets. It didn't pay attention the pronunciation accuracy.

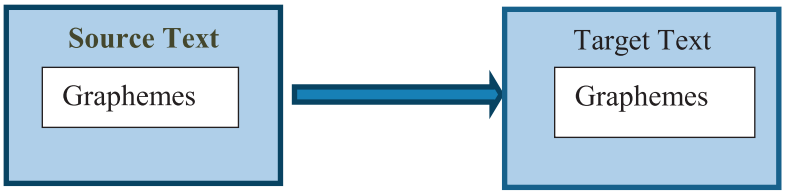


Figure 5: Spelling-based transliteration

The purpose of this work is to enable users to use both the two methods according to their needs. The phonetic-based method should be used by users with accurate pronunciation intention. However, spelling-based method is intended to overcome ICT limitations.

Transliteration criteria

In both transliteration categories we use English alphabet letters because they are all encoded using ASCII encoding. In the phonetic-based method some source letters are transliterated to a set of target letters that have the same aural rendering.

Figure 6: yash phonetic-based transliteration

Amazighe Letters with similar pronunciation are transliterated to the same Latin equivalent letter (Al-Onaizan, 2002). We are not attempting to reproduce the exact copy of the original transliterated text, but the nearest equivalent.

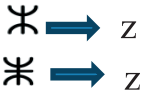


Figure 7: Same phonetic-based transliteration for similar sounds

The aim of the spelling-based method is to have a one-to-one mapping between source and target alphabet (Nasreen, 2003). One of the criteria that are used to choose target corresponding letters when there isn't one target letter with similar pronunciation is letters glyphs similarity. The letter yash is similar to the letter c.

Figure 8: yash spelling-based transliteration

For some missing sounds in the phonetic-based method we use the most common transliteration: The letter yakh is transliterated as kh. But in the spelling based method we look over the two most used languages in North Africa (French and spanich) to find single letters that have a similar aural rendering. The letter yakh is transliterated as x because it is read kh in the name Xavier in Spanish adding to this a graphological criterion which is glyph similaritybetween the target and the source letters.

Figure 9: yakh spelling-based transliteration

Both transliteration categories in this work use ASCII characters to guarantee that transliterated text will be portable i.e. characters keep their encoding values. They aren't neither lost nor confuse to other characters.

Optimal text readability is reached even in the spelling-based transliteration because of the use of graphical similarity and phonetic correspondence in the main spoken languages in the North Africa.
In order to help people who want to learn Amazighe or those who want to communicate using the transliteration, it was designed to be more mnemotechnic.

User interface

The user interface is compound of four components: menu that enables the common used functionalities of a simple text editor with the ability of saving both a source document and its corresponding target transliterated document. When a source or its target document is opened they are both opened in the same transliteration tool window.
The user should input its source document in the source text field and click on the transliterate button to run the transliteration process. The resulting transliterated text is displayed in the target text field. The choice of the transliteration type is enabled by the two radio buttons. The default transliteration method is the phonetic-based one.

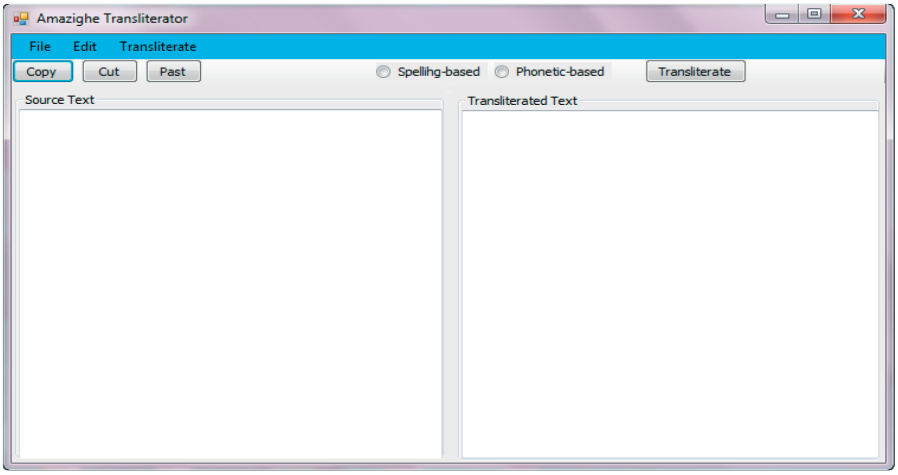


Figure 10: Amazighe transliteration tool screen shot

Conclusion

Amazighe, like several non-Latin-basedscript languages, lack support in ICT environments. This is the most persisting problem that makes the transliteration operation an inevitable practice. The main purpose of this work is to address some issues related to Amazighe text transliteration, and to implement a tool that can be used for both spelling and phonetic based transliterations. The back transliteration will be dealt with in a future work. Indeed, some Amazighe letters are associated to

the same English letter. The machine back transliteration process will be more complicated than the direct transliteration process. Extractive or generative machine transliteration is unavoidable to deal with this problem (Karimi, 2011).

Références

Banouni M., Lazrek A. and Sami K. (2002), Une translittération arabe/roman pour un e-document, Proceedings of *5e Colloque International sur le Document Électronique, Conférence Fédérative sur le Document, Hammamet, Tunisie*, 123-138.

Rachidi A., Mammass D. (2005), Informatisation de La Langue Amazighe : Méthodes et Mises En Œuvre, Proceedings of *3rd International Conference: Sciences of Electronic Technologies of Information and Telecommunications, TUNISIA*

Haralambous Y. (2007), *Fonts & Encodings*, Sebastopol, O'Reilly

Al-Onaizan Y., Knight K. (2002) Translating Named Entities Using Monolingual and Bilingual Resources. Proceedings of *ACL 2002*, 400-408

Karimi S., Scholer F., Turin A. (2007) Collapsed Consonant and Vowel Models: New Approaches for English-Persian Transliteration and Back-Transliteration, Proceedings of The *45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, 648-655

Karimi S., Scholer F., Turpin A. (2011),Machine Transliteration Survey, *ACM Computing Surveys*, Vol 43, Issue 4

Nasreen A., Larkey L. (2003): Statistical transliteration for english-arabic cross language information retrieval. CIKM, 139-146

Whistler K., Karlsson K., Kuhn M. (2009): GSM 03.38 to Unicode, http://www.unicode.org/Public/MAPPINGS/ETSI/GSM0338.TXT

POS tagging in Amazighe using tokenization and n-gram character feature set

Mohamed Outahajala (1, 4), Yassine Benajiba (2), Paolo Rosso (3), Lahbib Zenkouar (4)

(1) Royal Institut for Amazighe Culture, Morocco,

(2) Philips Research North America, Briacliff Manor, USA,

(3) Natural Language Engineering Lab – EliRF, DSIC, Universidad Politécnica de Valencia, Spain,

(4) Ecole Mohammadia d'Ingénieurs, Morocco,

outahajala@ircam.ma, yassine.benajiba@philips.com, prosso@dsic.upv.es, zenkouar@emi.ac.ma

Résumé - Abstract

L'objectif de cet article est de présenter le premier étiqueteur grammatical Amazighe. Très peu de ressources ont été développé pour l'Amazighe et nous croyons que le développement d'un outil d'étiquetage grammatical est la première étape dont on a besoin pour faire le traitement automatique de textes. Afin d'atteindre cet objectif, nous avons formé deux modèles de classification de séquences en utilisant Support Vector Machines (SVM) et Conditional Random Fields (CRFs) en utilisant une phase de segmentation. Nous avons utilisé la technique de 10 fois validation croisée pour évaluer notre approche. Les résultats montrent que les performances des SVMs et des CRFs sont très comparables. Dans l'ensemble, les SVMs ont légèrement dépassé les CRFs au niveau des échantillons (92,58% contre 92,14%) et la moyenne de précision des CRFs dépasse celle des SVMs (89,48% contre 89,29%). Ces résultats sont très prometteurs étant donné que nous avons utilisé un corpus de seulement ~ 20k jetons.

The aim of this paper is to present the first Amazighe POS tagger. Very few linguistic resources have been developed so far for Amazighe and we believe that the development of a POS tagger tool is the first step needed for automatic text processing. In order to achieve this endeavor, we have trained two sequence classification models using Support Vector Machines (SVMs) and Conditional Random Fields (CRFs) after using a tokenization step. We have used the 10-fold technique to evaluate our approach. Results show that the performance of SVMs

and CRFs are very comparable. Across the board, SVMs outperformed CRFs on the fold level (92.58% vs. 92.14%) and CRFs outperformed SVMs on the 10 folds average level (89.48% vs. 89.29%). These results are very promising considering that we have used a corpus of only ~20k tokens.

Keywords:

Étiquetage grammatical automatique, langue Amazighe, apprentissage supérvisé.
Automatic POS tagging, Amazighe language, supervised learning.

Introduction

The part-of-speech (POS) tagging consists of annotating each word in a sentence with its lexical category, i.e. part-of-speech. It is the first layer above the lexical level and the lowest level of syntactic analysis. Hence, all the NLP tasks dealing with higher linguistic levels resort to the POS tags, namely: phrase chunking; word sense disambiguation; grammatical function assignment (Cutting et al., 1992) and named entity recognition (Benajiba et al. 2010a; Benajiba et al. 2010b). In conjunction with partial parsing, POS-tagging is used in more complex tasks (Manning and Schütze, 1999) e.g.: lexical acquisition, information extraction, finding good indexing terms in information retrieval and question answering.
In the literature, proof is abound that the most effective approaches to build an automatic POS-tagger are based on supervised learning machines (See Section 2), i.e. relying on a manually annotated corpus and often other resources, such as dictionaries and word segmentation tools, to pre-process the text and extract features. Similarly, in our approach we use sequence classification techniques based on two state-of-art machine learning approaches, namely: SVMs and CRFs, to build our automatic POS-tagger. We use a ~20k tokens manually annotated corpus (Outahajala et al., 2011) to train our models and a very cheap feature set consisting of lexical context and character n-grams to help boost the performance.
The rest of the paper is organized as follows: in Section 2 we present related work on POS tagging techniques in other languages. Then, in Section 3 we give an overview on the Amazighe language and the employed tag set in the Amazighe corpus. In Section 4 we present the two supervised approaches based on SVMs and CRFs that have been employed for POS tagging. Section 5 describes experiments and discusses results. Finally, in Section 6 we draw some conclusions and describe the work to be done in the future.

Related work on POS tagging

The very first POS taggers were mainly rule-based systems. Building such systems requires a huge manual effort in order to handcraft the rules and to encode the linguistic knowledge which governs the order of their application. For instance, in

1970 Green and Robin (Greene, Rubin, 1971) developed a system named TAGGIT containing about 3,000 rules and achieving an accuracy of 77%. Later on, machine learning based POS-tagging proved to be both less laborious and more effective than the rule based ones. In the literature, many machine learning methods have been successfully applied for POS tagging, namely:
- The Hidden Markov Models (HMMs) (Charniak, 1993) whose states are tags or tuples of tags. For a bigram tagger for instance, the states of the HMM are tags, transition probabilities are probabilities of a tag given the previous tag and emission probabilities are probabilities of a word given a tag;
- The transformation-based error driven system (Brill, 1995) consisting in assigning the most frequent tag to each word by using an annotation reference. It proceeds afterwards by selecting the rule that yields the greatest error. This process is iterated as long as the annotation results are not close enough to the annotation reference;
- The decision trees (Schmid, 1999) based on decision support tool that uses a model and their possible consequences constructed using an annotation reference;
- The maximum entropy model (Ratnaparkhi, 1996) permitting the combination of diverse forms of contextual information without imposing any assumptions on training data, where the goal is to maximize the entropy of a distribution subject to certain constraints contained in the annotation reference.
- Learning algorithms that acquire a language model from a training corpus: based on previously learned examples, taggers based on this approach decide on the tag to attribute to the word (Kudo, Matsumoto, 2000; Lafferty et al. 2001).
Results produced by machine learning taggers obtain about 95%-98% of correctly tagged words. There are also, hybrid methods that use both knowledge based and statistical resources.
Though these methods have good performance, the accuracy for unknown words is much lower than that for known words, and this is a non-negligible problem where training data is limited. The tag set size may vary widely, for instance in POS tagging Arabic (Diab et al., 2004) the authors used a tagset containing 22 tags and 75 tags in (Diab et al., 2007).

Amazighe

In this section we give a gentle introduction to the Amazighe language and describe the adopted tag set in our experiments in Subsections 3.1. and 3.2., respectively.

The Amazighe language

The Amazighe language is spoken in Morocco, Algeria, Tunisia, Libya, and Siwa (an Egyptian Oasis); it is also spoken by many other communities in parts of Niger and Mali. It is used by tens of millions of people in North Africa mainly for oral

communication and has been introduced in mass media and in the educational system in collaboration with several ministries in Morocco. It belongs to the Hamito-Semitic/"Afro-Asiatic" languages[1], with rich templatic morphology (Chafiq, 1991). In linguistic terms, the language is characterized by the proliferation of dialects due to historical, geographical and sociolinguistic factors (the orthographic details are discusses in (Ameur et al., 2006).

In Morocco, for instance, one may distinguish three major dialects: Tarifit in the North, Tamazight in the center and Tashlhit in the southern parts of the country; it is a composite of dialects of which none have been considered the national standard.

Due to its complex morphology as well as the use of the different dialects in its standardization (Tashlhit, Tarifit and Tamazight the three more used ones), the Amazighe language presents interesting challenges for NLP researchers which need to be taken into account. Some of these characteristics are:

1. It does not support capitalization in its script.

2. It is written from left to right with its own alphabet called Tifinaghe (Zenkouar, 2008).

3. It is a complex morphology language.

4. Nouns, quality names (adjectives), verbs, pronouns, adverbs, prepositions, focalizers, interjections, conjunctions, pronouns, particles and determinants consist of a single word occurring between blank spaces or punctuation marks. However, if a preposition or a parental noun is followed by a pronoun, both the preposition/parental noun and the following pronoun make a single whitespace-delimited string. For example: ⵖⵔ (ɣr) "to, at" + ⵉ (i) "me (personal pronoun)" results into ⵖⴰⵔⵉ/ⵖⵓⵔⵉ (ɣari/ɣuri) "to me, at me, with me".

5. It is not an exception when it comes to POS ambiguity, i.e. same surface form might be tagged with a different POS tag depending on how it has been used in the sentence. For instance, ⵗⵯⴰ (ig°ra[2]) may have many meanings; as a verb, it means 'lag behind' while as a noun it refers to the plural noun of ⴰⵔⵓ

[1] http://en.wikipedia.org/wiki/Berber_languages

[2] The amazighe Latin transliteration used in this paper is the one defined in (Outahajala et al., 2010)

(agru) meaning 'a frog'. Some stop words such as "ⴷ" (d) might function as a preposition, a coordination conjunction, a predicate particle or an orientation particle.

6. Like most of the languages which have only recently started being investigated for the NLP tasks, Amazighe lacks annotated corpora and tools and still suffers from the scarcity of language processing tools and resources.

Our tag set

Defining the adequate tag set is a core task in building an automatic POS tagger. It aims at defining a processable tag set which is neither so large that it can hurt the performance of the learning machines, nor so small that there is not enough information to be used by the potential federate systems. In (Outahajala et al., 2010), a tag set containing 13 elements (verb, noun, adverb…etc.) was developed. For each element we define morpho-syntactic features and two common attributes: "wd" for "word" and "lem" for "lemma", whose values depend on the lexical item in question. The defined Amazighe elements and their attributes are set out in (Outahajala et al., 2011). The utilized tag set comprises 13 tags representing the major parts of speech in Amazighe, as it is summarized in Table 1. This tag set is derived from the larger one presented in (Outahajala et al., 2010). Gender, person and number information have not been included in the tag set and were considered as a separate investigation subject to be pursued in the future.

Labeled class	Designation
V	Verb
N	Noun
A	Quality name/Adjective
AD	Adverb
C	Conjunction
D	Determinant
S	Preposition
FOC	Focalizer mechanism
I	Interjection
P	Pronoun
PR	Particle
R	Residual (foreign, number, date, currency, mathematical and other)
F	Punctuation

Table 1. tag set.

Supervised learning for POS tagging

In this section we describe the theoretical foundations of supervised learning in general and of SVMs and CRFs in particular, being proved to give good results for sequence classification (Kudu, Matsomoto, 2000; Lafferty et al., 2001).
In this paper, stems + affixes and punctuation marks are referred to as tokens.

Supervised learning

In supervised learning the goal is to learn a function:
$h : X \rightarrow Y$ (1)

Where $x \in X$ are inputs and $y \in Y$ are outputs. The input objects are called instances, or examples, and they can be any kind of object, depending on the particular learning task: in NLP they could be for example documents to classify, strings of words to tag with POS-sequences which is our case. Depending on the nature of the output space Y, learning tasks can be categorized into several types:

- Binary classification: $Y = \{-1,+1\}$;
- Multiclass classification: $Y = \{1, \ldots, K\}$ (finite set of labels);
- Regression: $Y = R$;
- Structured prediction: here the outputs in Y are complex. For example, in a sequence labeling task such as POS-tagging, $Y = \{1\ldots,K\}$ n, i.e. the output is a sequence of labels of length n equal to the length of the input string.

Support Vector Machines

SVMs were first introduced by Vapnik (Vapnik, 1995); they are known for their good generalization performance and have been used for different recognition problems. For instance, in NLP SVMs are applied to text categorization (Kudu, Matsomoto, 2000), name entity recognition (Benajiba et al., 2010), base phrase chunking (Diab et al., 2007) and others. Many POS taggers based on SVMs have been achieved for many languages, for instance: Arabic (Diab et al., 2004; Diab et al., 2007), Bengali (Ekbal, Bandyopadhyay, 2008) etc .They are reported to have achieved a high accuracy without over fitting even with a large number of features. SVMs are also known for coping well with sparse and noisy data.
With respect to the task of POS tagging in Amazighe, the training process has been carried out by YamCha[3], an SVM based toolkit. For classification, we have used the TinySVM-0.09[4] classifier, a publicly available toolkit for the problem of pattern recognition.

[3] http://chasen.org/~taku/software/yamcha/
[4] http://chasen.org/~taku/software/TinySVM/

Conditional Random Fields

CRFs are undirected graph models. They are a generalization of Maximum Entropy Markov Models (MEMMs) and are oriented toward segmenting and labeling data (Lafferty et al., 2001). Conditional model specifies the probabilities of possible label sequences given an observation sequence. In addition of having the advantages of MEMMs, CRFs also overcome the label bias problem. We can think of CRFs as a finite state model with unnormalized transition probabilities. CRFs are applied to many NLP fields name entity recognition (Benajiba et al., 2010), shallow parsing (Sha, Pereira, 2003), information extraction from tables (Pinto et al., 2003). Dealing with POS tagging CRFs were used for many languages, for instance the Amharic (Adafre, 2005) and Tamil the Indian language (Lakshmana, Geetha, 2009).

We have used CRF++[5], an open source implementation of Conditional Random Fields for segmenting and labeling data, using the same data set as the used with YamCha.

Experiments and Error analysis

Corpus

Our corpus consists of a list of texts extracted from a variety of sources such as: Amazighe version of IRCAM's web site[6], the periodical "Inghmisn n usinag" (IRCAM newsletter) and primary school textbooks, annotated using AnCoraPipe tool (Bertran et al., 2008). Annotation speed of this corpus was between 80 and 120 tokens/hour. Randomly chosen texts were revised by different annotators. On the basis of the revised texts inter-annotator agreement is 94.98%.Common remarks and corrections were generalized to the whole corpora in the second validation by a different annotator.

The input format for YamCha and CRF++ is the same (see Figure 1 and Figure 2), where the first column for is for tokens, the last column is for the labeled class and if any features are used (see Figure 2) they are listed in the columns in between.

Here is an example, where we don't use segmentation of the input format for the sentence "ar as ttHyyaln i tmGra ann sg usgg°as lli izrin." [English translation: "They were preparing for the weddings from the last year"]:

ar	PR
as	S_P
ttHyyaln	V
I	S
tmGra	N
ann	D
sg	S
usgg°as	N

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

[6] http://www.ircam.ma/

lli	P
izrin	V
.	F

Fig.1. An extract from the training corpus

ar	a	-	-	-	r	-	-	-	PR
as	a	-	-	-	s	-	-	-	S_P
ttHyyalnt		tt	ttH	ttHy	n	ln	aln	yaln	V
i	-	-	-	-	-	-	-	-	S
tmGra	t	tm	tmG	tmGr	a	ra	Gra	mGra	N
ann	a	an	-	-	n	nn	-	-	D
sg	s	-	-	-	g	-	-	-	S
usgg°as	u	us	usg	usgg	s	as	°as	g°as	N
lli	l	ll	-	-	i	li	-	-	P
izrin	i	iz	izr	izri	n	in	rin	zrin	V
.	-	-	-	-	-	-	-	-	F

Fig.2. An extract from the training corpus using lexical features

In this paper, we explore two experiments sets. Both of them are based on SVMs and CRFs with and without lexical features (see sub section 5.2). In the first experiment set, we do not segment the compound words. In these experiments we use “S_P” and “N_P” to designate prepositions and kingship nouns respectively when put together with personal pronouns. In the second experiment set, we segment prepositions and kingship nouns when used together with pronouns. However, this is a problematic case since reverse function is not deterministic. We will use most frequent words in the corpus. For instance, the union of either the two morphemes “dg” and “s” can give either “digs” or “dags”, meaning [in it]. Hence, once we split the compound word, e.g. “digs”, into its component morphemes, i.e. “dg” and “s”, and given that it is not possible to compute the original form after POS tagging we will use “digs” since it is the most used in the corpus. In all our experiments, we have used the derived tag set presented above and the same one plus “S_P” and “N_P” corresponding to prepositions and kingship nouns respectively.

Features

In this paper, we explore two features sets. Both of them are based on the actual text and are very cheap to extract. In the first feature set (illustrated in Figure 1), we use:

1- the current token;
2- the surrounding words in a window of -/+2; and

3- the previous label.

In the second feature set (illustrated in Figure 2), we add to the first feature set character n-gram feature which consists of the last and first i character n-gram, with i spanning from 1 to 4.

10-fold experiments

In our first experiments about POS tagging (will appear in final version), we have shown that learning curve is increasing with training corpus size. In this experiment set, we have run 10-fold cross validation over the corpus, i.e., training on 90% of sentences and tagging the remaining 10%, with the experiment repeated 10 times, each time taking a different slice of the corpus.

Fold#	SVMs	SVMs (with lexical features)	CRFs	CRFs (with lexical features)
0	81,01	86,86	83,19	86,95
1	76,02	83,86	80,7	84,98
2	85,64	**91,66**	87	90,86
3	82,56	88,34	86,45	88,58
4	83,55	88,24	85,8	88,87
5	83,28	89,99	86,24	90,48
6	76,59	85,38	79,98	85,38
7	79,07	86,6	81,79	87,96
8	87,35	91,38	88,88	91,14
9	84,64	90,41	86,79	**91,35**
AVG	**81,97**	**88,27**	**84,68**	**88,66**

Table 2. 10-fold cross validation results.

By splitting the compound words, namely kingship nouns and prepositions when used together with pronouns, we obtained better results as shown in Table 4:

Fold#	SVMs	SVMs (with lexical features)	CRFs	CRFs (with lexical features)
0	82,85	87,94	84,46	87,31
1	78,27	85,06	81,55	85,9
2	87,59	**92,58**	87,9	91,42

3	83,95	89,62	87,39	89,22
4	85,06	89,02	86,93	89,26
5	86,08	91,38	87,6	91,62
6	79,27	86,42	82,9	87,18
7	81,34	86,96	83,69	88,96
8	**88,54**	92,47	**89,32**	91,79
9	86,45	91,49	88,65	**92,14**
AVG	**83,93**	**89,29**	**86,01**	**89,48**

Table 3. 10-fold cross validation results using tokenization.

Experiments and Result Discussion

For a better understanding of our system's behavior, we have examined the confusion matrix for the experiment which gave the highest accuracy. The analysis of the confusion matrix presents all the misclassified tags as shown in Tables 4 and 5.

Analyzing the most frequent errors in the two confusion matrices given by SVMs and CRFs, we found that adjectives are frequently tagged as nouns. This is due to the fact that adjectives may act as nouns. In line with this, many Amazighe linguists gave the name of quality nouns to adjectives. However, dropping the distinction between nouns and adjectives we obtained an improvement of 0.73 and a better score of 90.02% using 10 fold cross validation. However, by doing the same experiment with CRFs, we obtained an improvement of 0.77 and a better score of 90.25%.

Error rate of pronouns is also high due to the large overlap between them and the determinants. Another common source of errors is verbs. The POS tagger based on CRFs tagged 4.1% of verbs as nouns and adjectives and 1.6% as prepositions, whereas the POS tagger based on SVMs tagged 5.7% of verbs as nouns and adjectives. Besides SVMs based POS tagger have better results in tagging, pronouns, determinants, adverbs, focalizers and particles.

	N	A	V	P	D	S	C	AD	PR	FOC	F	I	R
N	93,1	0,3	1,8	0,6	3,9	0	0	0	0	0	0,3	0	0
A	18,2	63,6	18,2	0	0	0	0	0	0	0	0	0	0
V	5,4	0,3	93	0	0	0,7	0	0	0,7	0	0	0	0
P	0,7	0	0,7	91	5,5	0,7	0,7	0	0,7	0	0	0	0
D	3,3	0	1,1	9,9	84,6	0	0	1,1	0	0	0	0	0
S	0,5	0	1	0,5	0	94	2,1	0,5	1,6	0	0	0	0
C	0	0	0	2,1	2,1	2,1	83,3	4,2	4,2	2,1	0	0	0
AD	23,2	0	7,1	1,8	1,8	3,6	1,8	60,7	0	0	0	0	0
PR	0	0	0	0	1,9	0,6	0	0,6	96,8	0	0	0	0
FOC	0	0	0	0	40	0	0	0	0	60	0	0	0
F	0	0	0	0,2	0	0	0	0	0	0	99,8	0	0
I	36,4	0	0	0	0	0	0	0	18,2	0	0	45,4	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4. Confusion matrix in percentage when using SVMs with lexical features.

	N	A	V	P	D	S	C	AD	PR	FOC	F	I	R
N	94,6	2,4	2,1	0,2	0,3	0,1	0,2	0	0	0	0	0	0,1
A	12,6	82,3	4,6	0	0	0	0	0	0	0,6	0	0	0

Table 5. Confusion matrix in percentage when using CRFs with lexical features.

Some particles are confused for example conjunctions like “d” which has many eventual tags depending on the context. For instance in the following sentences, the word “d” might be:

- A coordination conjunction: “tamaziGt d tiknulujiyin timaynutin” [Amazighe and information technologies];
- A preposition: “iman d ubrid” [he went with the road];
- A predication particle: “d argaz” [he is a man];
- Or an orientation particle: “asi d tikint tamjahdit” [bring a large bowl].

Analyzing training and test sets, we observed that unknown words in the test set are important due to small size of the data set; also some errors still exist in the hand annotated corpora. Overall, misclassified words are unseen in the training set.

Conclusions and Further Work

In this paper we describe the morpho-syntactic features of the Amazighe language. We have addressed the design of two tag sets and two POS taggers based on SVMs and CRFs. The obtained accuracy achieved 92.58% and we have used the 10-fold technique to further validate our results. In this way, the POS tagger based on CRFs achieves 89.48% whereas the POS tagger based on SVMs and using lexical

features yields the accuracy of 89.29% based on a small corpus of ~20k tokens manually annotated.
We are currently trying to improve the performance of the POS tagger by using additional features and more annotated data based on semi-supervised techniques and active learning. In addition, we are planning to approach base phrase chunking by hand labeling the already annotated corpus with morphology information.

References

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

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

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

Benajiba Y., Zitouni I., Diab M., Rosso P. (2010b), Arabic Named Entity Recognition: Using Features Extracted from Noisy Data. In: Proc. of the 48th Annual Meeting of the Association for Computational Linguistics, ACL-2010, Uppsala, Sweden, July 11-16, pp. 281-285.

Bertran, M., Borrega, O., Recasens, M., Soriano, B. (2008), AnCoraPipe A tool for multilevel annotation. Procesamiento del lenguaje Natural, nº 41. Madrid, Spain.

Brants, T. (2000), TnT - A Statistical Part-of-Speech Tagger.

Brill, E. (1995), Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging

Chafiq, M. (0880), أربعة وأربعون درسا في الأمازيغية (Forty four lessons in Amazighe). éd. Arabo-africaines.

Charniak, E. (1993), Statistical Language Learning MIT Press, Cambridge

Cutting, D., Kupiec, J., Jan Pedersen, J. Sibun, P. (1992), Practical Part-of-Speech Tagger. Xerox Palo Alto Research Center.

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

Diab, M., Hacioglu, K., Jurafsky, D. (2007), Arabic Computational Morphology: Knowledge-based and Empirical Methods, chapter 9. Springer.

Ekbal, A.; Bandyopadhyay, S. (2008), Part of Speech Tagging in Bengali Using Support Vector Machine. In Information Technology, ICIT '08, pp. 106-111.

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

Kudo, T., Matsumoto, Y. (2000), Use of Support Vector Learning for Chunk Identification.

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

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

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

Outahajala M., Zenkouar L., Rosso P., Martí A. (2010), Tagging Amazighe with AncoraPipe. In: Proc. Workshop on LR & HLT for Semitic Languages, 7th International Conference on Language Resources and Evaluation, LREC-2010, Malta, May 17-23, pp. 52-56.

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

Outahajala, M., Zenkouar, L. (2008), La norme du tri, du clavier et Unicode. In Proceedings of la typographie entre les domaines de l'art et l'informatique. Rabat, Morocco. pp. 223—238.

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

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

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

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

Vapnik, Valdimir N. (1995), The Nature of Statistical Learning Theory. Springer Verlag, New York, USA.

Zenkouar L. (2008), Normes des technologies de l'information pour l'ancrage de l'écriture Amazighe, revue Etudes et Documents Berbères n°27, pp. 159-172.

Utilisation des réseaux de neurones et le modèle de Markov pour la reconnaissance des caractères Tifinagh manuscrits

B.EL KESSAB, C.DAOUI, B.BOUIKHALENE, M.FAKIR
Equipe de Traitement de l'Information et Télécommunications,
Faculty of Science and Technology, PB 523, Béni Mellal, Morocco,
E-mail: bade10@hotmail.fr, daouic@yahoo.com, bbouikhalene@yahoo.fr, fakfad@yahoo.fr,

Résumé – Abstract

Dans ce papier, nous proposons un système de reconnaissance des caractères Tifinagh manuscrits, basé sur l'utilisation de réseaux de neurones (le perceptron multicouches PMC) et le modèle de Markov caché (MMC) et sur la morphologie mathématique en phase d'extraction. Notre approche est testée sur une base de données de caractères Tifinagh manuscrits isolés de taille conséquente (900 images en apprentissage et 1275 exemples en test). Le taux de reconnaissance trouvé est de 95,55%. Les classificateurs utilisés (PMC et MMC) montrent des résultats suffisamment bons.

In this paper, we propose a system for recognizing handwritten characters Tifinagh, with the use of neural networks (the multi layer perceptron MLP) and hidden Markov model (HMM). And a feature extraction method based on mathematical morphology, this method is tested on a database of handwritten isolated characters Tifinagh (900 images in learning and 1275 test). The recognition rate obtained is 95.55%. The MLP and HMM classifiers show good enough results.

Keywords – Mots Clés

Reconnaissance des caractères manuscrits, Caractères Tifinagh, Réseau de Neurones, Modèle de Markov Caché, Morphologie mathématique.

Recognition of handwritten characters, characters Tifinagh, Neural Network, Hidden Markov Model, mathematical morphology.

Introduction

La reconnaissance automatique de caractères imprimés est depuis plusieurs années l'objet de plusieurs recherches. Cela a plusieurs applications : Dans le domaine des multimédias, fournir un aide aux non-voyants, la compression d'image etc.…
La reconnaissance automatique d'un caractère tifinagh s'effectue en trois étapes : la première phase est celle de prétraitement pour diminuer le bruit la deuxième pour extraire les caractéristiques, et la troisième pour faire la classification (Les réseaux de neurones, Modèle de Markov caché).
Les réseaux de neurones sont des systèmes de calcul largement utilisés pour la reconnaissance des images (Apurva A, 2010), (Baidyk T., et al. 2002), (Ososkov G., 2003), (Vitabile S., et al 2002). Dans l'apprentissage on utilise l'algorithme de descente du gradient. Dans le Modèle de Markov Caché on utilise le vecteur d'extraction comme une suite d'observation et on cherche à trouver la meilleure probabilité qui maximise le modèle. Pour l'apprentissage on utilise l'algorithme de Baum-Welch.
Nous proposons un système de reconnaissance des caractères Tifinagh (figure 1) implémenté sur une base de caractères Tifinagh manuscrits. Ce papier est organisé comme suit : La section 2 est consacrée à la base de données de test. Dans la section 3, nous décrirons la méthode d'extraction des caractéristiques basée sur la morphologie mathématique. Dans la section 4, nous présentons un aperçu sur les réseaux de neurones PMC. Les résultats expérimentaux des réseaux de neurones sont présentés à la section 5. Le Modèle de Markov Caché MMC est présenté dans la section 6. Les résultats expérimentaux sur Modèle de Markov Caché sont présentés dans la section 7.

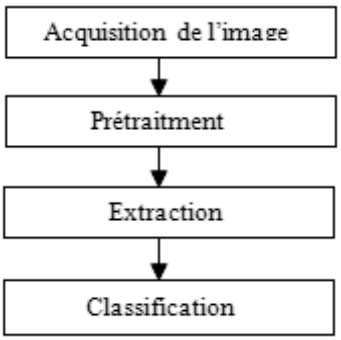


Figure 4 : Le Processus de reconnaissance

Base de données test

La base de données utilisée est Tifinaghe (Y. Ouguengay, M. Taalabi 2009), elle est composée de 2175 caractères Tifinagh écrites par différentes mains (900 en apprentissage) et (1275 en test). Le nombre des caractères Tifinagh selon IRCAM est 33 caractères (figure 2).

Figure 2 : Les caractères Tifinagh

Prétraitement

Dans la phase du prétraitement, les images des caractères sont écrites avec la main puis s rendus numériques avec un scanner, en suite on rendre l'image binaire avec le seuillage. Après on fait la normalisation pour les images extraites dans un carré normalisé de la taille 150x150.

Extraction

La méthode utilisée pour l'extraction est basée sur la morphologie mathématique (Serra J., 1982), (Soille P., 2003). On cherche à détecter pour chaque image quatre zones caractéristiques : la zone Ouest, la zone Est, la zone Nord et la zone Sud. Ces zones caractéristiques sont détectées grâce à la dilatation de l'image traitée dans les quartes directions.

La dilatation de l'image :

La dilatation est une transformation basée sur l'intersection entre l'objet de l'image A (les pixels blancs) avec un élément structurant B (une demi-droite). Elle est définie par la formule suivante

$$\text{Dilatation}(A, B)=\{x \in A / B_X \cap A \neq \emptyset\}$$

où A est l'objet de l'image (les pixels blancs), B l'élément structurant qui est un ensemble particulier de centre x, de géométrie et de taille connue (dans ce travail c'est une demi droite).

La figure 3 illustre un caractère dilaté vers l'Est.

Figure 3 : La dilatation du caractère vers L'Est

La figure 4 montre la dilatation pour les directions Ouest, Nord et Sud.

Figure 4 : Les dilatations du caractère vers l'Ouest, Nord et Sud

La détection des zones caractéristiques de l'image :

On détermine pour chaque caractère les paramètres discriminants (zones). Les zones caractéristiques peuvent être détectées en faisant des intersections de dilatations trouvées vers l'Est, l'Ouest, le Nord et le Sud. On définit cinq types de zones caractéristiques : Est, Ouest, Nord, Sud, et la zone Centrale.

1. ***A*** *Extraction de La zone caractéristique Est :*

Un point de l'image (Figure 5) appartient à la zone caractéristique Est (Figure 6) si et seulement si :

- Ce point n'appartient pas à l'objet du caractère (les pixels blancs).
- A partir de ce point, en se déplaçant en ligne droite vers l'Est, on ne croise pas l'objet (les pixels blancs).
- A partir de ce point, en se déplaçant en ligne droite vers le Sud, le Nord, l'Ouest on croise l'objet (Figure 5). Le résultat de l'extraction est illustré à la Figure 6.

Figure 5 : Image du caractère

Figure 6: La zone caractéristique Est (ZE)

*1. **B** Extraction de La zone caractéristique Ouest :*

Un point de l'image (Figure 7) appartient à la zone caractéristique Ouest si et seulement si :

- Ce point n'appartient pas à l'objet du caractère (les pixels blancs).
- A partir de ce point, en se déplaçant en ligne droite vers l'Ouest, on ne croise pas l'objet.
- A partir de ce point, en se déplaçant en ligne droite vers le Sud, le Nord, l'Est on croise l'objet (Figure 7). Le résultat de l'extraction est illustré à la Figure 8.

Figure 7 : Image du caractère

Figure 8: La zone caractéristique Ouest (ZO)

1. C Extraction de La zone caractéristique Sud :

Un point de l'image (Figure 9) appartient à la zone caractéristique Nord si et seulement si :

- Ce point n'appartient pas à l'objet du caractère.
- A partir de ce point, en se déplaçant en ligne droite vers le Sud, on ne croise pas l'objet.
- A partir de ce point, en se déplaçant en ligne droite vers le Nord, l'Est et l'Ouest on croise l'objet (Figure 9). Le résultat de l'extraction est illustré à la Figure 10.

Figure 9 : Image du caractère

Figure 10 : La zone caractéristique Sud (ZS)

1. D Extraction de La zone caractéristique Nord :

Un point de l'image appartient à la zone caractéristique Nord si et seulement si :

- Ce point n'appartient pas à l'objet du caractère.
- A partir de ce point, en se déplaçant en ligne droite vers le Nord, on ne croise pas l'objet.
- A partir de ce point, en se déplaçant en ligne droite vers le Sud, l'Est et l'Ouest on croise l'objet (Figure 11). Le résultat de l'extraction est illustré à la Figure 12.

Figure 11 : Image du caractère

Figure 12: La zone caractéristique Sud (ZS)

***1. E** Extraction de La zone caractéristique centrale :*

Un point de l'image appartient à une zone caractéristique centrale si:

- Ce point n'appartient pas aux autres zones caractéristiques Ouest, Est, Nord et Sud (les pixels blancs).

Figure 13 : La zone caractéristique centrale (ZC)

Chaque caractère est caractérisé par cinq composantes : NZW, NZE, NZN, NZS, NZC.
Avec :
NZW, NZE, NZN, NZS, NZC : Le nombre de pixels de valeur 1 respectivement dans les zones caractéristiques Ouest, Est, Nord, Sud, Centrale. Donc le vecteur d'extraction sera défini comme suit :

Vext=[ZW, ZE, ZN, ZS, ZC]

Avec :
Npixels : Nombre de pixels dans l'image traitée de taille 150x150

ZW = NZW / (Npixels).
ZE = NZE / (Npixels).
ZN = NZN / (Npixels).
ZS = NZS / (Npixels).
ZC = NZC / (Npixels).

Réseaux de neurones

Les réseaux de neurones (Dreyfus et. al, 2001), (Bishop C., 1995), (Haykin, 1998), (Hertz, et al, 1991), (Thiria, et al 1997) : s'inspirent de propriétés du cerveau pour construire des systèmes de calcul mieux capables de résoudre le type de problèmes que les êtres vivants savent résoudre. Ils possèdent plusieurs modèles, l'un de ces modèles est le perceptron.

5.1 Le perceptron multicouches (PMC).

Dans la phase de classification on procède comme suit :
Le nombre de neurones utilisé dans le réseau est :
- Cinq neurones dans la couche d'entrée (le nombre Cinq correspond aux valeurs trouvées dans le vecteur d'extraction).
- Dix-huit neurones dans la couche de sortie (le nombre Dix-huit correspond aux nombres des caractères utilisés).
Le nombre de neurones utilisé dans la couche cachée est choisi selon ces trois conditions :
- égal au nombre de neurones dans la couche d'entrée.
- égal à 75 % de nombre de neurones de la couche d'entrée.
- égal à la racine carrée du produit de deux couches de sortie et d'entrée.

Selon ces trois conditions on varié le nombre de neurones de la couche cachée entre cinq et dix neurones. La méthode utilisée pour l'apprentissage est celle de la descente du gradient (algorithme de rétro-propagation de gradient) (Flitcher, R.(1987), (Ciarlet, P.G. 1989), (Le Cun Y., et al 1990), (LeCun Y, et al 1998), (Jan A., 2005). Les matrices de connexion sont notées respectivement par Z et W

- La correction des erreurs pour la couche de sortie : On cherche à modifier les valeurs de Z. Le gradient de E par rapport à Z se calcule en utilisant la règle de dérivation des fonctions composées (ré-écrite en notation matricielle) (Marsden J.E. et al 1981).

$$\begin{aligned} Z_{t+1} &= Z_t + \Delta_Z E \\ &= Z_t - \alpha \times \nabla_Z E \\ &= Z_t - \alpha \times \frac{dE}{dZ} \end{aligned}$$

- La correction des erreurs pour la couche cachée : On cherche à modifier les valeurs de W comme suit

$$\begin{aligned} W_{t+1} &= W_t + \Delta_W E \\ &= W_t - \alpha \times \nabla_W E \\ &= W_t - \alpha \times \frac{dE}{dW} \end{aligned}$$

Les couches	Neurones		La constante d'apprentissage
Entrée	5		α = 0.9
Cachée	9		
Sortie	18		
L'erreur quadratique			La fonction d'activation
$E = \frac{1}{2}(t-o)^2$ t : La sortie théorique. o : La sortie voulue.			La fonction de sigmoïde $F(x) = \frac{1}{1+e^{-x}}$

Figure 15 : Les détails du réseau de neurones

Classification

Après l'extraction des données caractéristiques de l'image traitée, on fait la classification par les réseaux de neurones. On calcule alors les coefficients des poids et les sorties voulues.

Résultats expérimentaux

Les valeurs des vecteurs caractéristiques obtenues à la phase de l'extraction des caractéristiques sont introduites à l'entrée du réseau de neurones et on force le réseau de converger vers un état final précis (L'apprentissage supervisé). Chaque caractère est caractérisé par un vecteur d'extraction de cinq composantes. Pour la formation du réseau (Le perceptron multicouche PMC), on commencé par un ensemble de dix-huit images, et en fin avec 50 ensembles de 900 images, pour trouver les meilleurs paramètres qui maximise le réseau (Figure.16).

Ensembles des caractères manuscrits	Nombres des caractères	Base de Test
1 ensemble	18	70.86
10 ensembles	180	86.67
20 ensembles	360	92.02
30 ensembles	540	93.66
40 ensembles	720	95.35
50 ensembles	900	95.55
Nombres des images pour le test		1275 Images

Figure 16 : Résultats expérimentaux pour le PMC

Figure 17 : Les caractères utilisés pour la reconnaissance avec PMC

Modèles de Markov caché :

Les modèles de Markov cachés (Hidden Markov Models ou HMMs), (Choisy C., et al 2002), (Choisy C., et al 2002), (Choisy C., 2002) sont des modèles statistiques paramétriques de production du signal, largement utilisés en reconnaissance de la parole et plus tardivement en reconnaissance de l'écrit.

Chaine de Markov

Une chaine de Markov discrète d'ordre n est un processus stochastique discret $X = \{X_t \mid t = 1, \ldots, T\}$ avec des variables aléatoires discrètes, vérifiant la propriété de Markov :

$$P(X_t=q_{it}|X_{t-1}= qi_{t-1},\ldots,X_1= qi_1)= P(X_t=qi_t|X_{t-1}= q_{it-1},\ldots,X_{t-n}= qi_{t-n})$$

Ou Q={q_1,….,q_n} représente l'ensemble des états.

Chaine stationnaire

Une chaine de Markov d'ordre un est stationnaire si pour tout t et k on a :

$$P(X_t=q_i|X_{t-1}= q_j)= P(X_{t+k}=q_i|X_{t+k-1}= q_j)$$

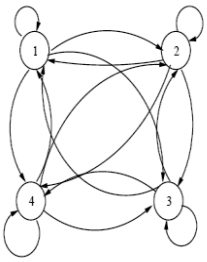
Dans ce cas, on définit une matrice de probabilité de transition A = (a_{ij}) telle que :

$$a_{ij} =P(X_t=q_j|X_{t-1}= q_i)$$

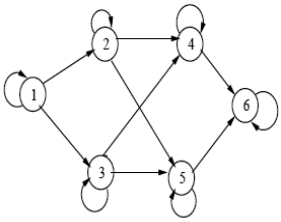
À un instant donné d'un processus quelconque.

Types de HMMs :

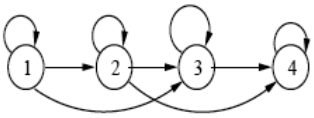
Les principaux types de modèles de Markov cachés sont le modèle ergodique et le modèle gauche _ droite.



Ergodique



gauche-droite : parallèle



gauche-droite : séquentiel

Evaluation de la probabilité d'observation :

Soient O=o1o2…oT une suite d'observations et Q = q1q2…qT la suite d'états associée.

La probabilité d'observation de O, étant donné le modèle β (ou classe), est égale à la somme sur tous les chemins d'états possibles Q des probabilités conjointes de O et Q.

$$P(O/\lambda) = \sum P(O,\mathrm{Q}/\gamma)$$

Calcul de $P(O/\gamma)$ à l'aide de la fonction Forward-Backword :

L'observation peut se faire en deux temps :

- Emission de début de l'observation O (1 : t).
- Emission de la fin de l'observation O (t+1 : t).

L'évaluation de l'observation est donnée par :

$$P(O/\gamma) = \sum \alpha(t,qi) \times \beta(t,qi)$$

Reconnaissance :

Connaissant la classe à laquelle appartient le caractère celui-ci est comparé aux modèles λ_k,
k= 1,……, L de sa classe. Le modèle retenu sera celui qui fournira la meilleure probabilité correspondant à l'évaluation de sa suite de primitives c.-à-d. : **max** $P(O/\lambda_k)$
Avec :
O : La suite d'observation dans ce travail c'est la vecteur d'extraction.
λ_k: c'est le modèle Markov constitué de la matrice de transition A, la matrice d'observation B et la matrice d'initialisation Πi.

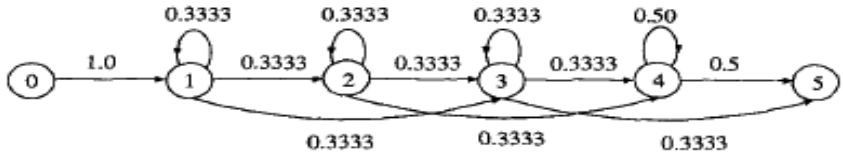


Figure 18 : Transitions initiales

Chaine de traitement

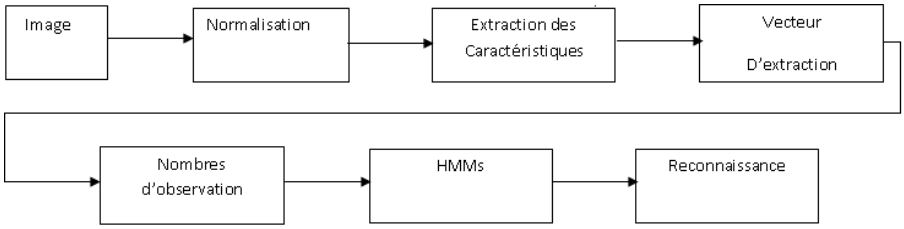


Figure 19: Le processus de traitement pour le MMC

Résultats expérimentaux

Pour la classification avec le modèle de Markov caché, on considère les valeurs des vecteurs caractéristiques obtenues dans la reconnaissance des caractères comme une suite d'observations, on initialise le modèle et on cherche avec l'algorithme de

Baum-Welch (L'apprentissage) de trouver la meilleure probabilité qui maximise les paramètres du modèle (A : la matrice de transition, B : La matrice d'observation, Πi : La matrice d'initialisation). Chaque caractère est caractérisé par un vecteur d'extraction de cinq composantes. Les résultats expérimentaux sont illustrés dans la figure suivante (Figure.20).

Nombre des caractères manuscrits dans la base	Type des caractères utilisés	Base de Test
1275	18	93.80

Figure 20 : Résultats expérimentaux pour le MMC

Figure 21 : Les caractères utilisés pour la reconnaissance avec MMC

Conclusion

Dans ce travail, nous avons utilisé une méthode basée sur les réseaux de neurones (le perceptron multicouches et la rétro propagation) et le modèle de Markov Caché pour le classification des caractères Tifinagh manuscrits. Une technique d'extraction basée sur la morphologie mathématique est utilisée dans la phase d'extraction des caractéristiques avant la mise en œuvre de la classification des caractères. Le taux de reconnaissance est 95.55% pour le PMC, et pour le MMC est 93.80% avec une base de Test qui contient 1275 images. En comparant les deux méthodes, on déduit que le principal inconvénient des PMC est celui de la reconnaissance relativement faible à la vitesse. Les classificateurs PMC et MMC pourraient être utilisés pour la reconnaissance d'écriture.

Références

Youssef Ait ouguengay, Mohamed Taalabi, "Elaboration d'un réseau de neurones artificiels pour la reconnaissance optique de la graphie amazighe : Phase d'apprentissage", 5e Conférence internationale sur les "Systèmes Intelligents : Théories et Applications", Paris : Europia, cop. 2009 (impr. au Maroc), ISBN 978-2-909285-55-3.

Apurva A. Desai (2010) Gujarati handwritten numeral optical character reorganization through neural network. Pattern Recognition 43 (2010) 2582–2589

Baidyk T., Kussul E. (2002), Application of neural classifier for flat image recognition in the process of microdevice assembly, Proceedings of the International Joint Conference on Neural Networks, Hawaii, USA 1 (2002) 160–164.

Ososkov G., (2003) Effective neural network approach to image recognition and control, Proceedings of International Conference on Physics and Control 1 (2003) 242–246.
Vitabile S., Gentile A., Sorbello F., (2002) A neural network based automatic road signs recognizer, Proceedings of the 2002 International Joint Conference on Neural Networks 3 2315–2320.
Dreyfus et. al (2001) : Réseaux de neurones. Méthodologie et applications. Eyrolles.
Bishop C. (1995) : Neural networks for pattern recognition. Clarendon Press - Oxford.
Haykin (1998) : Neural Networks. Prentice Hall.
Hertz, Krogh & Palmer (1991): Introduction to the theory of neural computation. Addison Wesley.
Thiria, Gascuel, Lechevallier & Canu (1997) : Statistiques et méthodes neuronales. Dunod.
Rosenblatt, F. (1962). Principles of Neurodynamics. Spartan.
Rumelhart, D. E., Hinton, McClelland, and Williams, R. J. (1986), Learning Internal Representations by Error Propagation.
Kussul E.M., Baidyk T.N., et al., (2001) Rosenblatt perceptrons for handwritten digit recognition, Proceedings of International Joint Conference on Neural Networks IJCNN 2 1516–1520.
Flitcher, R.(1987). Practical methods of optimization. New York: Wiley.
Ciarlet, P.G. (1989) Introduction to numerical linear algebra and optimisation. Cambridge: C.U.P
Le Cun, Y., Boser B., Denker J., Henderson D., Howard R., Hubbard W., and Jackel L, (1990). Handwritten digit recognition with a back-propagation network. Advances in neural information processing systems. San Mateo, Morgan Kaufmann. 396-404.
LeCun Y., Bottou L, Bengio Y., Haffner P., (1998) Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) 2278–2344.
Jan A. Snyman (2005). Practical Mathematical Optimization : An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms. Springer Publishing
Marsden, J.E., Tromba A.J., (1981) Vector Calculus, 2e édition, W.H. Freeman and Company, New York.
Serra J., 1982, Image Analysis and Mathematical Morphology. Academic Press, London.
Serra J., 1988, editor. Image Analysis and Mathematical Morphology.II : Theoretical Advances. Academic Press, London.
Soille P., 2003, Morphological Image Analysis : Principles and Applications. 2nd edition.

Choisy C., Belaid A., (2002) Couplage d'une vision locale par HMM et globale par RN pour la reconnaissance de mots manuscrits. In Conférence Internationale Francophone sur l'écrit et le document.
Choisy C., (2002) Modélisation analytique de l'écriture manuscrite par une segmentation basée sur des champs de Markov. Ph.D thesis, Université de Nancy2.
Chevalier S., Geoffrois E., et Preteux F., (2003) A 2D dynamic programming approach for Markov random field-based handwritten character recognition. In Proceedings of International Conference on Image and Signal Processing (ICIP), volume 2, page 616-629

Transformation de Fourier et Moments Invariants Appliqués à la Reconnaissance des Caractères Tifinaghe

R. EL AYACHI (1), M. FAKIR (1) & B. BOUIKHALENE (2)

(1) Equipe de traitement de l'information et de télécommunications (TIT), Facultés des Sciences et Techniques, Université Sultan Moulay Slimane, Béni Mellal, Maroc
rachidieea@yahoo.fr, fakfad@yahoo.fr
(2) Equipe de traitement de l'information et de télécommunications (TIT), Facultés Poly disciplinaire, Université Sultan Moulay Slimane, Béni Mellal, Maroc
bbouikhalene@yahoo.fr

Résumé – Abstract

La Reconnaissance des Caractères Optiques OCR est un outil qui vise à donner la possibilité aux ordinateurs de lire les caractères sans intervention humaine. Le problème de l'OCR est celui de faire reconnaitre les caractères indépendamment de leurs positions, rotation et taille. Pour remédier à ce problème on utilise les descripteurs invariants en l'occurrence la transformée de Fourier et les moments invariants.
Le système développé dans ce travail utilise la transformation de Fourier et les moments invariants pour extraire les attributs et la programmation dynamique et les réseaux de neurone pour la classification.

Optical Character Recognition OCR is a tool that aims to provide opportunities for computers to read characters without human intervention. The objective of OCR is characterization of a character by invariant descriptors in translation, rotation and scaling.

In this paper, the OCR developed use invariant moments and Fourier transform in extraction phase. In the recognition phase, dynamic programming and neural network are adopted.

Keywords – Mots Clés

Reconnaissance des Caractères Optiques, transformation de Fourier, moment invariant, programmation dynamique, réseaux de neurones.

Optical Character Recognition OCR, Fourier transform, invariant moment, Dynamic programming, Neuronal Network.

1 Introduction

La Reconnaissance de Caractères Optique ROC (Optical Character Recognition OCR) (Fakir, 2001 ; Fakir et al, 2000 ; El Ayachi et al, 2010 ; El Ayachi, Fakir, 2009 ; Es Saady et al, 2010) est un sujet de recherche populaire dans le domaine de la reconnaissance de forme.
Les applications de l'OCR sont nombreuses et s'appliquent à des domaines aussi variés que le guidage automatique de véhicule, la reconnaissance d'objet, la numérisation d'ouvrage, les transactions banquières,
Dans cet article, l'objectif de l'OCR (Fig.1) est de reconnaitre les caractères Tifinaghe (Fig.2). Le système adopté comporte plusieurs phases : Prétraitement, extraction d'attributs et reconnaissance.
La base de données utilisée est Tifinaghe (Y. Ouguengay, M. Taalabi 2009), elle est composée de 2175 caractères Tifinagh écrites par différentes mains (900 en apprentissage) et (1275 en test).

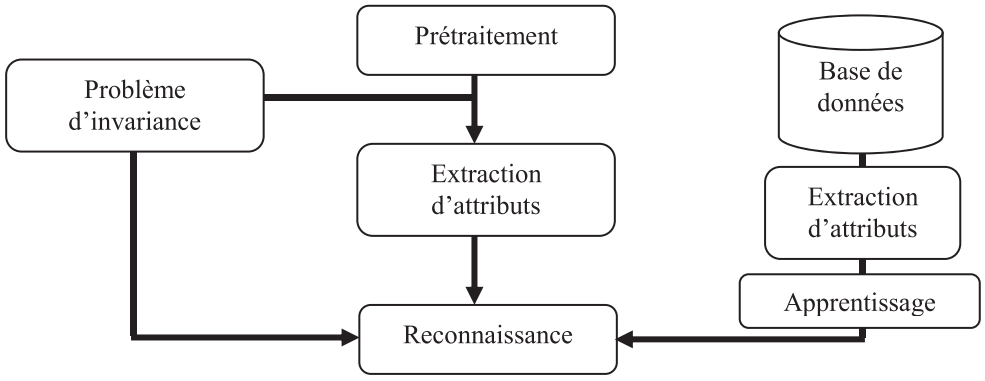


Figure 1 : L'OCR

La structure de l'article est organisée de la façon suivante : la section (2) s'intéresse à la phase de prétraitement qui contient plusieurs opérations, section(3) représente la phase d'extraction d'attributs en utilisant la transformation de Fourier et les moments invariants, section (4) traite la phase de la reconnaissance à l'aide des réseaux de neurones et la programmation dynamique, section (5) résoudre le problème d'invariance, section (6) illustre les résultats obtenues et section (7) donne les différentes conclusions tirées à partir de ce travail.

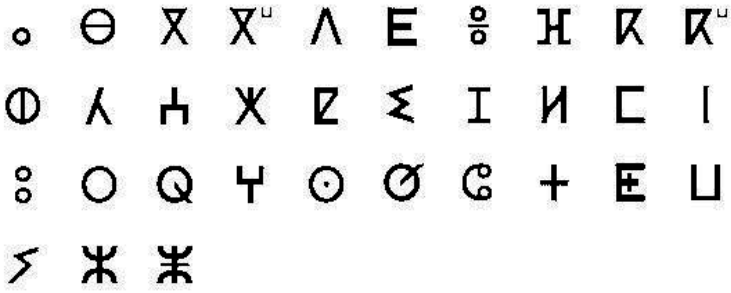


Figure 2 : Les caractères de Tifinaghe – IRCAM

2 Prétraitement

Après l'acquisition de l'image, le système de reconnaissance commence par la phase de prétraitement à fin d'améliorer la qualité d'image et de réduire le temps d'exécution. Le prétraitement comporte la normalisation, la correction de l'inclinaison et la segmentation.

2.1 Normalisation

Pour éliminer les zones inutiles dans une image, la méthode d'histogramme (El Ayachi et al, 2010) est utilisée. Le principe repose sur le calcul des histogrammes horizontale et verticale, ensuite, le parcourt des deux histogrammes pour déterminer les marges à supprimer.

2.2 Correction d'inclinaison

L'inclinaison est un phénomène qui influence négativement sur la phase de segmentation et par conséquent sur le système de reconnaissance, donc il est indispensable d'utiliser une méthode efficace pour la correction de l'inclinaison.

La correction d'inclinaison passe par deux étapes: la détection de l'angle d'inclinaison et la rotation dans le sens approprié.
Pour détecter l'angle d'inclinaison, nous avons adopté la transformation de Hough (Fakir et al, 2000 ; El Ayachi et al, 2010) qui est connu par robustesse dans ce type de problèmes.

2.3 Segmentation

La fonction de segmentation consiste premièrement à détecter les lignes dans une image contenant un texte en Tifinaghe, secondement, segmenter ces lignes en caractères.
Aussi, dans cette fonction de segmentation, nous avons utilisé la méthode d'histogramme (El ayachi et al, 2010). Le balayage de l'histogramme horizontal de haut en bas pour déterminer les lignes. Le parcourt de l'histogramme vertical de gauche à droite afin de segmenter les lignes en caractères.

3 Extraction d'attributs

L'extraction d'attributs est la deuxième phase à appliquer dans un OCR, elle joue un rôle primordial dans la reconnaissance, car elle doit prendre en considération la représentation du caractère en quelques situations telles que: la translation, la rotation et l'échelle. C'est la raison derrière l'utilisation des moments invariants (Fakir, 2001 ; El Ayachi et al, 2010) et la transformation de Fourier (Ghorbel, 1993) dans le système traité dans ce travail.

3.1 Moments invariants

Soit f une fonction définie par : $f(x,y)=1$ sur une région R fermée et délimitée et $f(x,y)=0$ ailleurs.
On définit le moment d'ordre (p,q) comme suit :

$$m_{pq}=\iint_R x^p y^q f(x,y)dxdy \quad , \quad \text{pour} \quad p,q=0,1,2,... \tag{1}$$

Les moments centraux peuvent être exprimés par :

$$\mu_{pq} = \iint_R (x-\overline{x})^p (y-\overline{y})^q f(x,y)dxdy \quad , \text{ avec} \quad \overline{x} = \frac{m_{10}}{m_{00}} \quad , \quad \overline{y} = \frac{m_{01}}{m_{00}}$$

(2)

Pour une image numérique, l'équation (2) devient :

$$\mu_{pq} = \sum_{(x,y)\in R}\sum (x-\overline{x})^p (y-\overline{y})^q f(x,y)$$

(3)

On peut facilement vérifier que les moments centraux d'ordre $p+q \leq 3$ peuvent être calculés par les formules suivantes :

$$\mu_{00} = m_{00}$$

$$\mu_{10} = 0$$

$$\mu_{01} = 0$$

$$\mu_{11} = m_{11} - \overline{y}m_{10}$$

$$\mu_{20} = m_{20} - \overline{x}m_{10}$$

$$\mu_{02} = m_{02} - \overline{y}m_{01}$$

(4)

$$\mu_{12} = m_{12} - 2\overline{y}m_{11} - \overline{x}m_{02} + 2\overline{y}^2 m_{10}$$

$$\mu_{21} = m_{21} - 2\overline{x}m_{11} - \overline{y}m_{20} + 2\overline{x}^2 m_{01}$$

$$\mu_{30} = m_{30} - 3\overline{x}m_{20} + 2\overline{x}^2 m_{10}$$

$$\mu_{03} = m_{03} - 3\overline{y}m_{02} + 2\overline{y}^2 m_{01}$$

Les moments centraux sont invariants par translation. Ils peuvent être normalisés, pour conserver l'invariance, par changement d'échelle et on obtient les moments centraux normalisés :

$$\alpha_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \quad , \quad \text{avec} \quad \gamma = \frac{p+q}{2} + 1 \quad , \quad \text{pour} \quad p+q = 2,3,...$$

(5)

Les invariants du moment suivants ont été obtenus par Hu (1962) et fréquemment utilisés comme caractéristiques pour la reconnaissance des formes.

$$\varphi_1 = \alpha_{20} - \alpha_{02}$$

$$\varphi_2 = (\alpha_{20} - \alpha_{02})^2 + 4\alpha_{11}^2$$

(6)

$$\varphi_3 = (\alpha_{30} - \alpha_{12})^2 + (3\alpha_{12} - \alpha_{03})^2$$

$$\varphi_4 = (\alpha_{30} + \alpha_{12})^2 + (\alpha_{21} + \alpha_{03})^2$$

$$\varphi_5 = (\alpha_{30} - 3\alpha_{12})(\alpha_{30} + \alpha_{12})[(\alpha_{30} + \alpha_{12})^2 - 3(\alpha_{21} + \alpha_{03})^2]$$
$$+ (3\alpha_{21} - \alpha_{03})(\alpha_{21} + \alpha_{03})[3(\alpha_{30} + \alpha_{12})^2 - (\alpha_{21} + \alpha_{03})^2]$$

$$\varphi_6 = (\alpha_{20} - \alpha_{02})[(\alpha_{30} + \alpha_{12})^2 - (\alpha_{21} + \alpha_{03})^2]$$
$$+ 4\alpha_{11}(\alpha_{30} + \alpha_{12})(\alpha_{21} + \alpha_{03})$$

$$\varphi_7 = (3\alpha_{21} - \alpha_{30})(\alpha_{30} + \alpha_{12})[(\alpha_{30} + \alpha_{12})^2 - 3(\alpha_{21} + \alpha_{03})^2]$$
$$+ (3\alpha_{12} - \alpha_{03})(\alpha_{21} + \alpha_{03})[3(\alpha_{30} + \alpha_{12})^2 - (\alpha_{21} + \alpha_{03})^2]$$

Hu a montré que ces quantités $\varphi_i, (1 \le i \le 7)$ sont invariantes par changement d'échelle, translation et rotation.

3.2 Transformation de Fourier et moments invariants (étude théorique)

Pour $f : IR^2 \to IR$, sa transformée de Fourier est donnée par :

$$\hat{f}(u,v) = \int_{-\infty-\infty}^{+\infty+\infty} f(x,y) e^{-i2\pi(xu+yv)} dxdy$$

(7)

Cette transformation admet l'approximation discrète :

$$\hat{f}(u,v) = \sum_x \sum_y f(x,y) e^{-i2\pi(xu+yv)}$$

(8)

On définit aussi le moment d'ordre (p,q) par l'équation (1) dont l'approximation discrète est donnée par l'équation :

$$m_{pq} = \sum_x \sum_y x^p y^q f(x,y)$$

(9)

<u>La relation entre les moments et la transformation de Fourier :</u>

Si on calcule

$$\frac{\partial^{p+q}\widehat{f}(u,v)}{\partial u^p \partial v^q} = \int_{-\infty-\infty}^{+\infty+\infty}\int (-i2\pi x)^p (-i2\pi y)^q f(x,y) e^{-i2\pi(xu+yv)} dxdy$$

$$= (-i2\pi)^{p+q} \int_{-\infty-\infty}^{+\infty+\infty}\int x^p y^q f(x,y) e^{-i2\pi(xu+yv)} dxdy$$

(10)

Pour $(u,v) = (0,0)$

On obtient

$$\frac{\partial^{p+q}\widehat{f}(0,0)}{\partial u^p \partial v^q} = (-i2\pi)^{p+q} \int_{-\infty-\infty}^{+\infty+\infty}\int x^p y^q f(x,y) dxdy$$

$$= (-i2\pi)^{p+q} m_{pq}$$

(11)

Alors

$$m_{pq} = \frac{\dfrac{\partial^{p+q}\widehat{f}(0,0)}{\partial u^p \partial v^q}}{(-i2\pi)^{p+q}} = Fe\left[x^p y^q f(x,y)\right]_{(0,0)}$$

(12)

D'où

$$\frac{\partial^{p+q}\widehat{f}(u,v)}{\partial u^p \partial v^q} = Fe\left[(-i2\pi)^{p+q} x^p y^q f(x,y)\right]$$

$$= \sum_x \sum_y (-i2\pi)^{p+q} x^p y^q f(x,y) e^{-i2\pi(xu+yv)}$$

$$= (-i2\pi)^{p+q} \sum_x \sum_y x^p y^q f(x,y) e^{-i2\pi(xu+yv)}$$

(13)

On a

$$m_{00} = \widehat{f}(0,0)$$

$$m_{10} = \frac{\dfrac{\partial\widehat{f}(0,0)}{\partial u}}{(-i2\pi)} = Fe\left[xf(x,y)\right]_{(0,0)}$$

$$m_{01} = \frac{\dfrac{\partial\widehat{f}(0,0)}{\partial v}}{(-i2\pi)} = Fe\left[yf(x,y)\right]_{(0,0)}$$

(14)

$$m_{11} = \frac{\dfrac{\partial^2 \hat{f}(0,0)}{\partial u \partial v}}{(-i2\pi)^2} = Fe\left[xyf(x,y)\right]_{(0,0)}$$

$$m_{02} = \frac{\dfrac{\partial^2 \hat{f}(0,0)}{\partial v^2}}{(-i2\pi)^2} = Fe\left[y^2 f(x,y)\right]_{(0,0)}$$

$$m_{12} = \frac{\dfrac{\partial^3 \hat{f}(0,0)}{\partial u \partial v^2}}{(-i2\pi)^3} = Fe\left[xy^2 f(x,y)\right]_{(0,0)}$$

$$m_{20} = \frac{\dfrac{\partial^2 \hat{f}(0,0)}{\partial u^2}}{(-i2\pi)^2} = Fe\left[x^2 f(x,y)\right]_{(0,0)}$$

$$m_{21} = \frac{\dfrac{\partial^3 \hat{f}(0,0)}{\partial u^2 \partial v}}{(-i2\pi)^3} = Fe\left[x^2 yf(x,y)\right]_{(0,0)}$$

$$m_{03} = \frac{\dfrac{\partial^3 \hat{f}(0,0)}{\partial v^3}}{(-i2\pi)^3} = Fe\left[y^3 f(x,y)\right]_{(0,0)}$$

$$m_{30} = \frac{\dfrac{\partial^3 \hat{f}(0,0)}{\partial u^3}}{(-i2\pi)^3} = Fe\left[x^3 f(x,y)\right]_{(0,0)}$$

Alors la conclusion que nous tirons à partir de cette étude théorique c'est que nous pouvons utiliser la transformation de Fourier et ces dérivées pour calculer les valeurs moyennes pondérées d'une image f .

4 Reconnaissance

La reconnaissance est la tâche la plus délicate dans un OCR, car la réussite de ce système de reconnaissance repose sur la décision résultante. Donc, il est nécessaire d'utiliser un processus de classification efficace qui a un taux de reconnaissance élevé et par conséquence un taux d'erreur diminué.

Dans cette étape, nous avons utilisé deux méthodes (la notion de réseau de neurones (El Ayachi et al, 2010; El Ayachi, Fakir, 2009) et la programmation dynamique (El Ayachi et al, 2010; Chevalier et al, 2003) pour

la reconnaissance des caractères Tifinaghe, parce qu'ils sont connus de leurs succès dans le domaine de la reconnaissance des caractères.

4.1 Réseau de neurones

La figure (Fig.3) représente l'exemple de réseau de neurones utilisé, c'est un réseau multicouche qui contient une couche caché.
Cet exemple de réseaux comporte :

- Une couche d'entrée à 7 cellules d'entrées (vecteur du moment invariant) $E_i = X_i$
- Une couche cachée de 3 neurones d'activations Y_j
- Une couche de sortie de 6 neurones d'activations Z_k
- 7 × 3 connexions entre la couche d'entrée et la couche cachée, chacune pondérée par V_{ji}
- 3 × 6 connexions entre la couche cachée et la couche de sortie, chacune pondérée par W_{kj}
- X_0 et Y_0 sont des scalaires

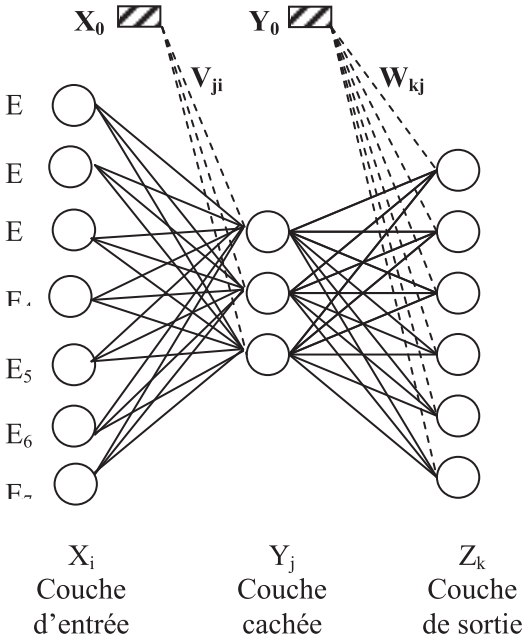


Figure 3 : Réseau de neurones

Le principe de fonctionnement du réseau de neurones (Fig.3) repose sur un ensemble détapes :

Etape1: (Initialisation des poids des connexions)

Les poids sont pris aléatoirement.

Etape2: (Propagation des entrées)

Les entrées E_i sont présentées à la couche d'entrée : $X_i = E_i$

La propagation vers la couche cachée se fait à l'aide de la formule suivante :

$$Y_j = f\left(\sum_{i=1}^{7} X_i V_{ji} + X_0\right) \tag{15}$$

Ensuite de la couche cachée vers la couche de sortie, on adopte :

$$Z_k = f\left(\sum_{j=1}^{3} Y_j W_{kj} + Y_0\right) \tag{16}$$

X_0 et Y_0 sont des scalaires

f est la fonction d'activation (fonction sigmoïde) :

$$f(a) = 1 \Big/ 1 + \exp(-a) \tag{17}$$

Etape3: (Rétropropagation de l'erreur)

Au niveau de la couche de sortie, l'erreur entre la sortie désirée S_k et la sortie réelle Z_k est calculée par :

$$E_k = Z_k(1 - Z_k)(S_k - Z_k) \tag{18}$$

L'erreur calculée est propagée sur la couche cachée en utilisant la formule suivante:

$$F_j = Y_j(1 - Y_j)\sum_{k=1}^{6} W_{kj}.E_k \tag{19}$$

Etape4: (Correction des poids des connexions)

On corrige les poids de connexions entre la couche d'entrée et la couche cachée par :

$$\Delta V_{ji} = \eta.X_i.F_j \quad \text{et} \quad \Delta Y_0 = \eta.F_j \tag{20}$$

Puis, on modifie les connexions entre la couche cachée et la couche de sortie par :

$$\Delta W_{kj} = \eta.Y_j.E_k \quad \text{et} \quad \Delta X_0 = \eta.E_k \tag{21}$$

η Un paramètre à déterminer empiriquement.

Etape5: (Boucle)
Boucler à l'étape2 jusqu'à l'obtention d'un critère d'arrêt à définir (seuil d'erreur, nombre d'itérations).

Après l'apprentissage et l'exécution d'OCR, on utilise la distance euclidienne pour identifier les caractères de Tifinaghe :

$$d(t_k, o) = \left(\sum_{i=1}^{6} (t_{ki} - o_i)^2 \right)^{1/2} \quad (22)$$

Avec, t_k la sortie désirée et o la sortie du réseau.

4.2 Programmation dynamique

La programmation dynamique est la deuxième méthode utilisée dans la reconnaissance. Elle est basée sur deux étapes suivantes :

Etape1 : Calculer la matrice d entre le vecteur du caractère segmenté V_{car} et chaque vecteur des caractères Tifinaghe de référence V_{ref}

La matrice est donné par:

$$d(x, y) = \left| V_{car}(x) - V_{ref}(y) \right| \quad (23)$$

Avec $x, y = 1,2,...,7$

Etape2 : Calculer le chemin optimal à partir du point $(1,1)$ jusqu'à point (x, y) en utilisant la formule récursive suivante :

$$S(x, y) = d(x, y) + \min \begin{Bmatrix} S(x-1, y) \\ S(x-1, y-1) \\ S(x, y-1) \end{Bmatrix} \quad (24)$$

Avec $S(x, y)$ est la distance cumulée le long du chemin optimal à partir du point $(1,1)$ jusqu'à point (x, y).

$S(x, y)$ est évalué dans l'espace $[1,7]*[7,1]$ ce qui est parcouru colonne par colonne et ligne par ligne à partir du point $(1,1)$.

Etape3 : Calculer l'indice de dissemblance :

$$D(V_{car}, V_{ref}) = \frac{S(7,7)}{7} \quad (25)$$

5 Problème d'invariance

Un OCR efficace c'est un système de reconnaissance qui prend en considération l'invariance du caractère en quelques situations (translation, rotation et taille). Cependant, pour les caractères Tifinaghe, cette caractéristique d'invariance pose un problème dans la phase de reconnaissance, car il y en a des caractères qui se ressemblent (et , et , et). Donc, il faut ajouter un autre traitement pour régler ce problème. Ce qui est l'objectif de cette section, ou bien, dans la phase d'extraction d'attributs, il faut utiliser une autre méthode variante.

Pour cela, nous allons enregistrer le caractère segmenté puis exécuter processus pour calculer d'autres variables (SUP, INF et HORIZ) sui vont être intervenir lors de la reconnaissance. Ce processus comporte deux traitements : le premier pour les caractères (et) et (et) et le deuxième pour les caractères (et).

Le premier traitement (Fig.4) contient les étapes suivantes:

- Diviser l'image en deux parties horizontalement (SUP et INF).
- Calculer le nombre de pixels noirs dans chaque partie.
- Attribuer 1 au variable SUP et 0 au variable INF si le nombre de pixels noirs dans la partie SUP est supérieur au nombre de pixels noirs dans la partie INF.
- Attribuer 0 au variable SUP et 1 au variable INF si le nombre de pixels noirs dans la partie SUP est inférieur au nombre de pixels noirs dans la partie INF.

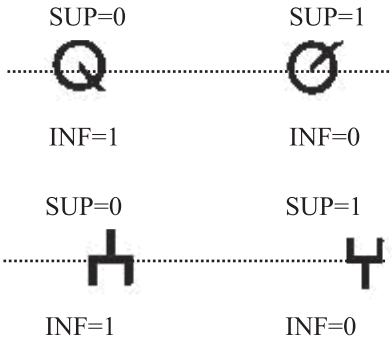


Figure 4 : Premier traitement

Le tableau suivant indique les valeurs des variables SUP et INF :

Variables				
SUP	0	1	0	1
INF	1	0	1	0

Figure 5 : Les variables SUP et INF

Le deuxième traitement (Fig.6) se déroule de la façon suivante :

- Tracer le contour du caractère.
- Enlever le contour externe.
- Calculer l'histogramme horizontal.
- Parcourir l'histogramme, s'il contient deux zones représentant les pixels noirs, alors attribuer 2 au variable HORIZ, et s'il contient une seule zone, alors attribuer 1 au variable HORIZ.

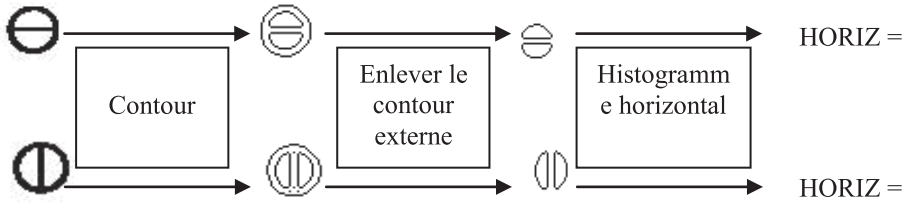


Figure 6 : Deuxième traitement

Le tableau suivant indique les valeurs du variable HORIZ :

Variable		
HORIZ	2	1

Figure 7 : La variable HORIZ

6 Résultats

L'image de la figure (Fig.8) représente un exemple de texte en Tifinaghe.

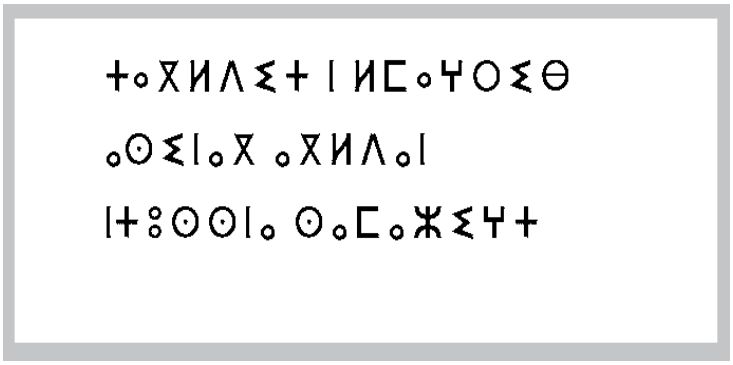
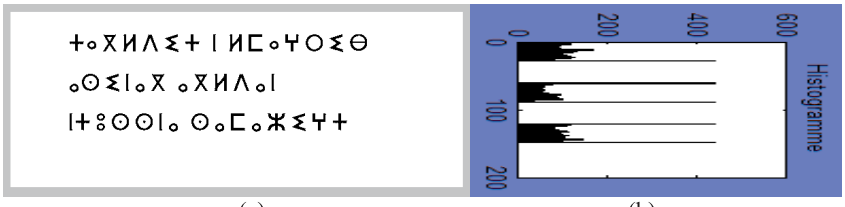
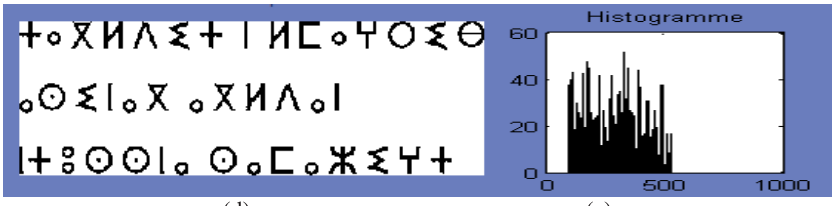


Figure 8 : Exemple de texte en Tifinaghe

Après l'application de la phase de prétraitement sur la figure (Fig.8), nous obtiendrons les figures (Fig.9 et Fig.10) suivantes :

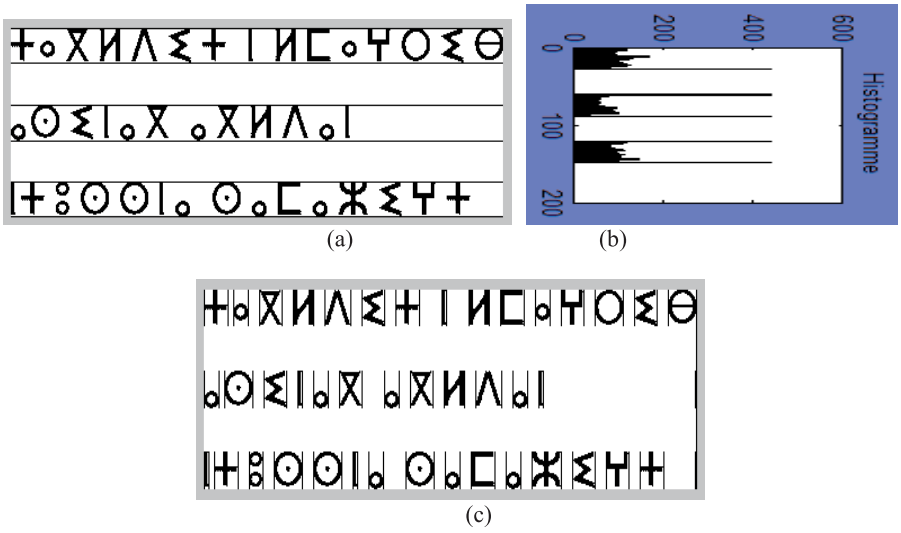


(a) (b)



(d) (c)

Figure 9 : (a) Avant normalisation, (b) Histogramme horizontal
(c) Histogramme vertical et (d) Après normalisation

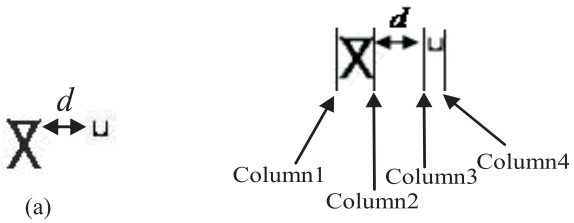


(a) (b)

(c)

Figure 10 : (a) Segmentation en lignes, (b) Histogramme horizontal et (c) Segmentation en caractères

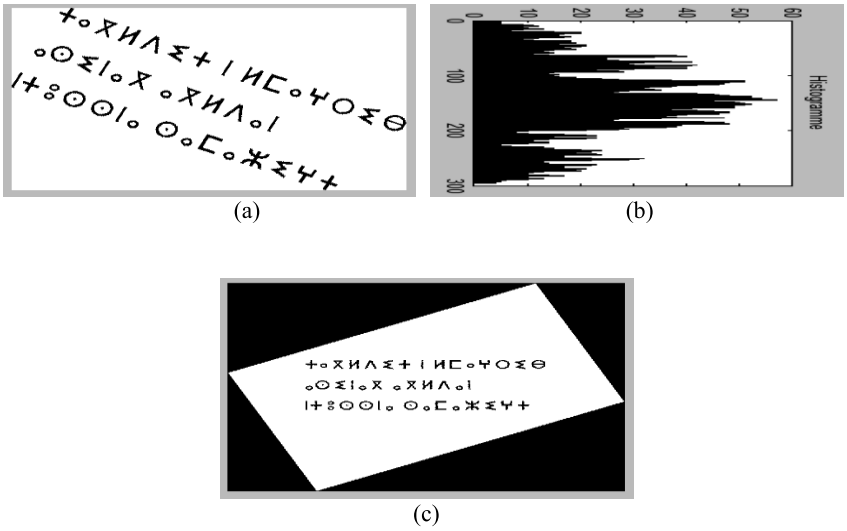
La méthode d'histogramme rencontre un problème dans le cas de la segmentation des caractères (et), elle divise chacun de ces caractères en deux, ce qui est faux, et par conséquence, elle pose une difficulté lors de la reconnaissance. Pour remédier à ce problème, nous faisons recours à la distance d représentée dans la figure (Fig.11). Tout d'abord, nous calculons la distance d, en suite, si $d \leq 5 pixels$ alors nous annulons les colonnes 2 et 3 afin d'obtenir un seul caractère.



(a) (b)

Fig.11 : (a) Distance d, (b) Colonnes de segmentation

Dans le cas d'un texte incliné, la fonction de la correction d'inclinaison permet de résoudre le problème, ce qui est illustré dans la figure (Fig.12).



(a) (b)

(c)

Figure 12 : (a) Incli naison, (b) Histogramme horizontal
et (c) Correction d'inclinaison

La comparaison effectuée, entre les deux approches appliquées dans la phase de reconnaissance (la programmation dynamique et le réseau de neurones) en se basant des attributs extraits à l'aide de la transformation de Fourier et les moments invariants, est illustrée dans le tableau suivant :

	Transformation de Fourier et moments invariants	
	Taux de reconnaissance	*Taux d'erreur*
Programmation dynamique	*92.76%*	*7.24%*
Réseau de neurones	*93.27%*	*6.73%*

Figure 13 : Taux de reconnaissance et taux d'erreur

Les résultats du tableau3 montrent que :

- Le taux de reconnaissance calculé en utilisant le réseau de neurones est supérieur que le taux de reconnaissance trouvé par la programmation dynamique.
- Le taux d'erreur du réseau de neurones est inférieur que le taux d'erreur de la programmation dynamique.

Dans ce système, la source des erreurs repose sur deux points :

- Problème d'invariance ( et  ,  et  , et ) : L'existence de quelques caractères qui sont identiques en quelques situations.
- Le vecteur caractéristique extrait contient uniquement sept éléments (ce qui est insuffisant pour une reconnaissance efficace).

Pour remédier à ce problème de source d'erreurs, les solutions qu'on peut proposer sont :

- Pour le cas d'invariance, la correction peut se faire à l'aide de l'utilisation de la méthode détaillée dans la partie (5) de cet article, ou bien, l'adoption d'une méthode (phase d'extraction) qui n'a pas le problème d'invariance.
- Pour le cas du nombre d'éléments du vecteur caractéristique, on peut utiliser des méthodes qui produisent des vecteurs dont le nombre d'éléments est supérieur à sept, ou bien, regrouper plusieurs méthodes pour obtenir un seul vecteur dont le nombre d'éléments est supérieur à sept.

La base de données utilisée est constituée de 627 images. La configuration de l'ordinateur portable utilisé est :

Processor: Intel (R) Core (TM) 2 Duo, CPU T5870 @ 2.00 Ghz.

RAM: 2.00 Go.

7 Conclusion

Dans ce travail on a traité un exemple de système de reconnaissance des caractères Tifinaghe. Ce système est composé d'un ensemble de phases : prétraitement, extraction d'attributs et reconnaissance.

La phase de prétraitement a regroupé plusieurs fonctions à appliquer sur l'image d'entrée ; la fonction de normalisation permettant l'élimination des zones indésirables afin de réduire le temps d'exécution et minimiser l'espace mémoire utilisé, la fonction de correction d'inclinaison qui facilite l'étape suivante concernant la segmentation en lignes et en caractères.

Dans l'extraction d'attributs, nous avons utilisé une méthode, qui garde l'invariance du caractère en quelques situations (translation, rotation et

taille), appelée moments invariants, ces moments invariants sont calculés à l'aide de la transformation de Fourier. Cette invariance a posé un problème lors de la phase de reconnaissance, ce problème revient à l'existence des caractères de Tifinaghe identiques (s'ils ont subi une rotation par exemple). Il est résolu par l'exécution d'un traitement en parallèle avec le système de reconnaissance.
La reconnaissance c'est la phase la plus délicate dans un OCR, car l'efficacité de ce système se base sur la classification trouvée, c'est la raison derrière le choix du réseau de neurones et programmation dynamique comme deux méthodes de reconnaissance grâce à leurs succès dans ce type de domaine de recherche.
Les résultats obtenus ont montré la réussite de ce système traité dans la reconnaissance des caractères Tifinaghe, c'est-à-dire le succès de l'utilisation de la transformation de Fourier dans le calcul des moments invariants.

Références

F. Ghorbel (1993), Application de la Transformée de Fourier Généralisée au Problème de l'invariance en Reconnaissance de Formes à Niveau de Gris, Quatorzième *Colloque GRESTI*, *JUAN-LES-PINS*- du 13 au 16 septembre 1993.

M. Fakir. (2001), Reconnaissance des Caractères Arabes Imprimés, *Thèse*, pp : 28-36

M. Fakir, M.M. Hassani and C.Sodeyama. (2000), On the recognition of Arabic characters using Hough transform technique, Malysian *Journal of Computer Science*, Vol. 13, No.2, Dec.2000, pp: 39-47.

R.El Ayachi, K. Moro, M. Fakir and B. Bouikhalene. (2010), On The Recognition Of Tifinaghe Scripts, *JATIT*, vol. 20, No. 2, pp: 61-66, 2010.

R. El Ayachi and M. Fakir, Recognition of Tifinaghe Characters Using Neural Network, 978-1-4244-3757-3/09/$25.00 ©2009 IEEE.

S. Chevalier, E. Geoffrois, and F. Prêteux. (2003), A 2D Dynamic Programming Approach for Markov Random Field-based Handwritten Character Recognition, *Proceedings IAPR International Conference on Image and Signal Processing (ICISP' 2003)*, Agadir, Morocco, 2003, p. 617-630.

Y. Es Saady, A. Rachidi, M. El Yassa, D. Mammass. (2010), Printed Amazigh Character Recognition by a Syntactic Approach using Finite Automata, *ICGST-GVIP Journal*, Volume 10, Issue 2, June 2010.
Youssef Ait ouguengay, Mohamed Taalabi, "Elaboration d'un réseau de neurones artificiels pour la reconnaissance optique de la graphie amazighe : Phase d'apprentissage", 5e Conférence internationale sur les "Systèmes Intelligents : Théories et Applications", Paris : Europia, cop. 2009 (impr. au Maroc), ISBN 978-2-909285-55-3.

Intégration de l'amazighe dans un OCR OpenSource :

Ocropus comme modèle

Ait ouguengay Youssef
aitouguengay@ircam.ma
Institut Royal de la Culture Amazighe
Rabat-Maroc

Résumé - Abstract:

Le travail présenté dans cet article consiste en l'intégration du système graphique amazighe, le Tifinaghe, dans uns système OCR opensource. C'est une exploitation d'un travail de standardisation du Tifinaghe mené á l'IRCAM (Institut Royal de la Culture Amazighe) depuis 2002, tant au niveau linguistique que technique. Le système OCR utilisé est celui utilisé par le moteur de recherche Google. Une base de données, contenant quelques milliers de patterns présentant différents polices de caractères amazighes, a été générée pour l'entrainement du moteur de classification du système. Les premiers résultats donnent un bon taux de reconnaissance des patterns utilisés dans l'entraînement.

Mots Clés - OCR, Tifinaghe, entrainement, OCRopus, classification, Tesseract.

The work presented in this paper aims to integrate the Amazigh writing system,Tifinaghe, in an open source OCR system. It is a continuation of the work led to standardization of Tifinaghe by IRCAM (Royal Institute of Amazigh Culture) since 2002, in both computing and linguistic fields. The used OCR system is the one used by the Google search engine and sponsored by the Google funding. A database containing several thousands of patterns with different fonts Amazigh, was generated to drive the classification engine system. Early results give a good rate of recognition of patterns used in training.

Keywords – OCR, Tifinaghe, training, Ocropus, classification, Tesseract.

1. Introduction :

Le système graphique amazighe: le Tifinaghe, comme tout autre script alphabétique, a besoin de suivre l'évolution des nouvelles technologies. Longtemps vécu comme script historique plus qu'un vrai système alphabétique aménagé, le

« Tifinaghe » est, depuis l'an 2002, le caractère officiel de l'écriture de l'amazighe au Maroc. En fait, « *Tifinaghe »* est une graphie d'un age de 25 siècle qui a son histoire, et ses formats d'apparition n'ont cessé de se développé depuis l'apparition de l'écriture rupestre amazighe jusqu'au néo-Tifinaghe. Depuis son adoption comme graphie officielle pour l'écriture de l'amazighe au Maroc, le Tifinaghe a subit plusieurs opérations d'aménagement, aux niveaux linguistique et technologique. Ces travaux réalisés et appuyés par les différents centres de recherche de l'Institut Royal de la Culture Amazighe (IRCAM) ont été couronnés par plusieurs résultats notamment par la normalisation internationale du Tifinaghe et son intégration dans le standard Unicode et dans la norme ISO/ CEI 10646 depuis l'an 2004. Le Tifinaghe jouit, depuis, d'un statut de système alphabétique complet pour la langue amazighe.

Le projet proposé dans ce papier vient cooronner ses efforts et exposer l'alphabet amazighe à la communité internationale du développement. Il consiste dans son objectif globale à intégrer l'amazighe dans un systéme opensource de reconnaissance optique de caractères tirant profis des expériences et travaux précédents dans ce domaine.

Dans la première section nous situerons notre travail par rapport à l'éxistent. La deuxiéme section présente un état d'art des systèmes OCR en prenant Ocropus comme système open source modèle. Nous détaillons dans la troisiéme section le travail effectué pour l'intégration du Tifinaghe dans Ocropus et particulièrement sur l'entrainement du moteur tesseract et le moteur interne d'Ocropus pour la reconnaissance du Tifinaghe. Dans la conclusion nous décrirons les perspectives de notre travail.

2. Situation du travail :

Un système OCR complet comporte plusieurs niveaux de traitements incluant : l'analyse de la mise en page des documents, segmentation et reconnaissance, post traitement, support multilinguisme, etc.

En mettant à coté l'énorme effort consenti dans les problématiques de la reconnaissance optique des alphabets latins, plusieurs travaux ont été effectués dans ce domaine sur les alphabets non latins. Nottamment, les travaux sur le thaï [1], le Bangali [2], le Korian [3], le chinois [4] ou l'arabe [5]. Pour l'amazighe aussi, même s'il n'en est qu'à ces débuts informatiques, la reconnaissance optique des caractères constitue un centre d'intêrét des différents laboratoires de recherches au Maroc. Ainsi quelques travaux de recherche, de fond, sur l'OCR amazighe ont étudié plusieurs méthodes de classification : les réseaux de neurones ([6], [7]), les automates à états finis [8], etc. Néanmoins, un système complet de reconnaissance optique pour le Tifinaghe, couvrant tout le processus de la reconnaissance, est toujours un besoin d'actualité. Le travail décrit dans ce papier essaye de s'identifier à ce besoin.

Nous proposons dans une premiére étape d'adapter le noyau d'un des systèmes OCR présent dans la scène international : Ocropus, qui s'occupe de la classification en se basant sur son moteur ocr : Tesseract. Ce travail nous permettera de se concentrer sur l'entrainement et le test du noyau de système pour la classification des caractères amazighes avant d'attaquer selon les besoins les autres aspects du système.

3. Le système Ocropus:

Parmi les systèmes OCR du marché libre, Ocropus est réputé à être un système modèle complet et open source pour la reconnaissance optique des documents. Le projet étant supporté par google et d'autres organisations, l'un des objectifs majeurs visés est de produire un système OCR multilingues. Les contributeurs au projet, développeurs et chercheurs, peuvent intégrer leurs propres langages et scripts dans le projet.

3.1. Architecture générale modèle :

Dans une application réelle, la reconnaissance des caractères doit accepter en entrée un document image et produire en sortie un fichier texte. Nous avons, donc, besoin non seulement de développer le processus de reconnaissance proprement dit, comportant les phases du pré-traitement, entrainement, apprentissage et post-traitement, mais aussi de s'occuper de la partie de modélisation des patterns (caractères) à reconnaître.

L'étape de prétraitement est conçue pour adapter l'image d'entrée á un format normalisé. Elle comporte le processus d'élimination du bruit, l'alignement de la page, l'affinage et la normalisation. La segmentation des lignes et/ou des caractères transforme le document d'entrée en un ensemble d'unités isolés et normalisés qui vont être analysé par le classificateur et comparés avec les donnés d'apprentissage.

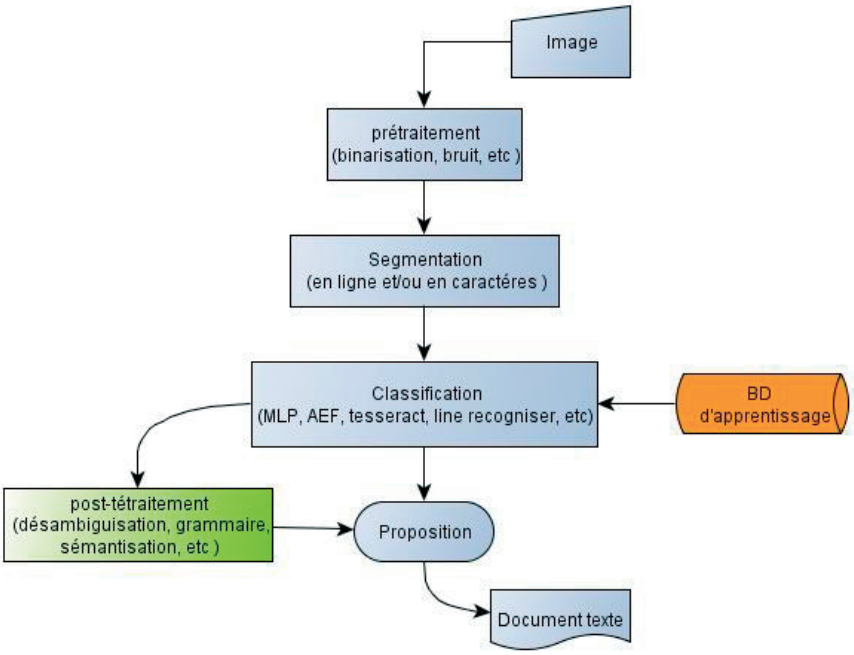


Fig.1. Structure modèle de la reconnaissance optique

*Apprentissage.*L'apprentissage est une procédure qui consiste á configurer le système de reconnaissance afin d'arriver á accomplir les tâches qui lui sont affféctés. Habituellement, l'étape de l'apprentissage permet d'induire au systéme des modéles á partir des exemples.

Trois types d'apprentissage éxistent :

Apprentissage supervisé (dit aussi manuel) : dont lequel on connaît les valeurs que doit avoir la sortie en fonction des entrées correspendants. Dans ce type d'apprentissage on utilise un nombre de données siginificatif : Dans le cas des caractéres imprimés on utilise des polices de caractéres qui représentent bien les caractères.

Apprentissage non supervisé (ou automatique) qui, sans indiquer les classes de sortie, fait apprendre au systéme de reconnaissance un grand nombre de données toute en éliminant les ambiguïtés entre les données significatives d'une classe donnée. En fait en disposant d'un grand ensemble de données, on cherche á les regrouper dans des classe, selon des critéres de ressemblance qui sont inconnus á priori (agrégation)

Apprentissage continu : où les caractéres reconnu en cours de production peuvent être ajoutées á la base d'apprentissage. La base de données du systéme est actualisée á chaque fois. Les caractéres non reconnus peuvent étre reconnus par vérification léxicale ou sont retournées á l'utilisateur pour les identifier.

post-traitement. Le post traitement des résultats d'un système de reconnaissance optique est utilisé pour améliorer les résultats du ROC. Deux méthodes distinctes peuvent être envisagées : la combinaison de plusieurs classificateurs ou l'utilisation d'un modèle linguistique de langage. Ce qui permet de résoudre le problème du pourcentage d'erreur résultant du bruit et de ressemblance des caractères, par une analyse des caractères en groupe.
Dans le cas de la combinaison des classificateurs, deux principes sont considérés dans cette méthode : L'intégrité : l'activation et configuration de chaque classificateur sont dérigés par un algorithme de vote. Pour avoir un degré d'intégrité haut : l'algorithme de vote n'active que les meilleurs classificateurs selon la situation rencontrée. La représentation des résultats du classification : qui peut se faire d'une manière abstraite oú chaque classificateur donne seulement un résultat, ou d'une manière ordonnée, dans ce cas le système donne plusieurs résultats ordonnées du meilleur au plus mauvaise. Un degré de certitude peut être définis, en donnant une probabilité définissant la pertinence des résultats de la classification.

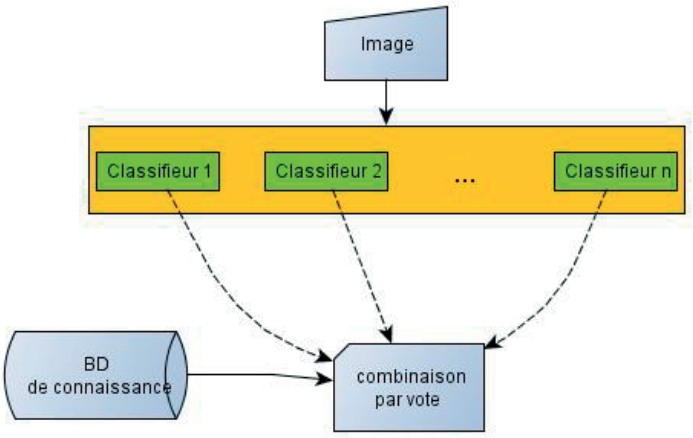


Fig.2.19 Chaîne de la combinaison des classificateurs

Utilisation des informations contextuelles. L'analyse contextuelle essaye de corriger les erreurs dérivées des décisions prises á partir de l'analyse des caractéristiques locales. En fait, dans le cas d'incertitude, les hypothèses générées par le classificateur local de caractère courant sont complétées avec les hypothèses des caractères voisins.
Pour utiliser les informations des caractères d'entourage, deux contextes peuvent être exploité : le contexte géométrique et le contexte linguistique.
Les techniques basées sur le contexte géométrique utilisent la probabilité qu'un mot de n lettres á reconnaître soit un mot valide du langage en question. Pour les

caractères qui présentent une incertitude á leur reconnaissance, on prenne la décision selon une probabilité, prédéfinit, la plus haute.

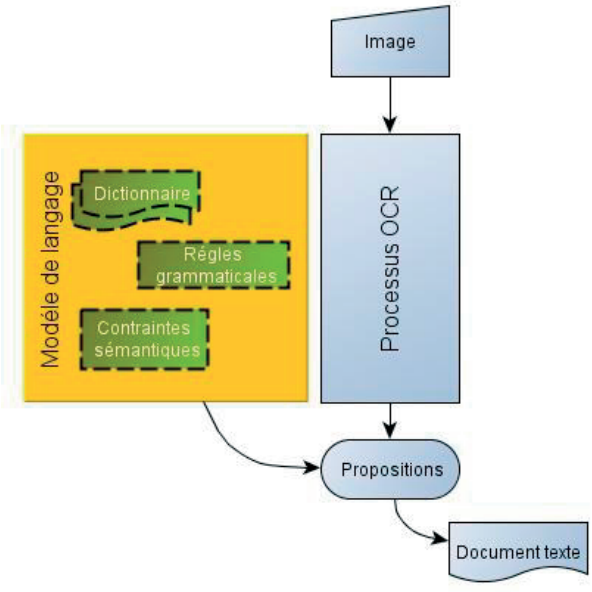


Fig. Amélioration de la sortie OCR par TAL

Une autre technique est basée sur des grammaires et dictionnaires, en utilisant un modèle linguistique du langage en traitement pour valider le résultat de la ROC. Ils sont similaires aux premiers saufs qu'elles permettent de traiter des chaînes de caractères á longueur variable selon les règles grammaticales. Avec cette technique, on construit un dictionnaire avec l'ensemble des mots corrects permettant ainsi la correction orthographique du texte. Pour remédier aux problèmes des mots qui n'apparaissent pas au dictionnaire, des structures de représentation de dictionnaire d'accès rapides, sont construites, pour une possibilité de mise á jours par l'utilisateur.

3.2. La Classification dans Ocropus :

Ocropus dispose de plusieurs classifieurs abstraits concrétisés par des composants (routines) internes qui sont appelés durant l'entrainement du systéme. Le classifieur reçoit en entrés un vecteur seriélisé de caractéristiques extraites. En mode reconnaissance le classifieur assigne les classes aux vecteurs caractéristiques du pattern à reconnaitre selon les probablités calculés.
Deux types de modèle de classifications sont utilisés par Ocropus :

3.2.1.les K voisins les plus proches (k-Nearest Neighbors)

C'est une méthode de classification qui utilise plusieurs modèles pour chaque classe. Ayant une image, on prend les k modèles les plus proches de l'image. L'image est classifiée dans la classe qui apporte le plus de modèles à l'ensemble des k-modèles.

On définit la distance :

$$D_i(x) = \left| V_x^{(i)} - V_x \right| \text{ [1]}$$

Où : V_x : l'ensemble des k modéles les plus proch et $V_x^{(i)}$: les modéles appartenant à la classe i.

Ou encore :

$$D_i(x) = \sum_{j=1}^{n} \left| x^j - x_k^j \right|$$

et chaque classe i est définis par :

$$i = élements(\min\{d(x, x_k) / k = 1,..., n_k\}$$

3.2.2.Les réseaux de neurons (RNA)

Plusieurs types de réseaux de neurones artificiels (RNA) existent. les plus utilisés, pour la classification, sont les réseaux multi-couches à rétropropagation dits aussi perceptrons multicouches. Le même modéle est utilisé par Ocropus.

Dans le cas de la classification par RNA on peut prendre l'image de caractére isolé comme entrée qui génére un vecteur de caractéristiques de son image. le RNA est organisé en couches; dans chaque couche un nombre détérminés de neurones. La premiére couche correspend aux élements du vecteur des caractéristiques et la dérniére couche correspend aux différentes classes. Les couches intérmidiaires représentent les sous-vecteurs caractéristiques de la couche suivante. La valeur de chaque neurone est calculée à partir de l'application d'une fonction des valeurs des neuods de la couche en aval, pondérées par un vecteur de poids :

$$x_j^k = b_j^k + \sum_{i=1}^{n(k-1)} w_{ji}^k x_i^{k-1}$$

Oú :

x_i^k : la valeur du neuod i de la couche k

$n(k)$: nombre du neuods de la couche k

b_i^k : biais du neuod i de la couche k

w_{ji}^k : le poid de la connexion entre les neuods j de la couche k et le neuod i de la couche k-1.

---

[1] Les formules mathématiques sont données ici uniquement à titre indicatif.

Au moment du conception du RNA, on décide du nombre des couches et du nombre des neuods de chaque couche. La phase d'apprentissage sert à détérminer les valeurs des poids des connexions inter-neuronales qui minimisent l'erreur de classification.

4. Intégration du Tifinaghe dans Ocropus :

Le sous système graphique Tifinaghe ciblé par ce travail a un faible niveau de difficultés : caractérisé par un ensemble réduit de 33 caractères utilisés dans le système éducatif marocain. Les caractères peuvent être imprimés de multiples fontes ou manuscrits séparés. Le répertoire qui nous intéresse comporte aussi les signes de ponctuations habituelles et les chiffres arabes.

Ils existent plusieurs façons d'intégrer un nouveau script dans Ocropus : En interfaçant un classificateur externe à base de ligne, déjà entrainé, à Ocropus comme le cas du Tesseract que nous traitons dans cette section, ou en entrainant un des classificateurs internes d'Ocropus.

4.1. Apprentissage du Tifinaghe par le classificateur Tesseract

Tesseract est un moteur de reconnaissance optique, à base de ligne de commande, assez complet et supportant Unicode. Il a été intégré comme classificateur dans la version 0.2 d'Ocropus. La procédure d'intégration du Tifinaghe dans Tesseract passe par la création de plusieurs fichiers de données qui seront exploités dans l'entrainement.

Dans un premier lieu, nous nous sommes limités aux 33 caractères Tifinaghe largement utilisés aux Maroc avec les ponctuations usuelles et les chiffres romans. La même procédure est applicable aux restes des caractères. Les données crées peuvent être par la suite incrémentés par les nouveaux entrées.

–	:	’	2	ⵔ	ⵅ	ⵔ	ⵣ
!	?	+	²	ⵖ	ⵀ	ⵇ	ⵣ
#	@	<	3	ⵠ	ⵠ	ⵄ	
&	[	=	4	ⵠⵓ	ⵆ	ⵙ	
(	]	>	5	ⵏ	ⵋ	ⵁ	
)	_	«	6	ⵄ	ⵉ	ⵛ	
*	{	»	7	ⵒ	ⵑ	+	
,	}	§	8	ⵅ	ⵇ	ⵄ	
.	~	0	9	ⵅⵓ	ⵏ	ⵍ	
/	‘	1		ⵙ	ⵗ	ⵋ	

Fig. : ensemble des unités graphiques traités

La base de données d'apprentissage brut a été construite à partir d'un large corpus amazighe en codage Unicode de manière aléatoire. Des images textes de 1746 caractères sous 16 fontes différents ont été générées. Le graphe suivant montre la fréquence d'apparition des lettres Tifinaghe dans la base de données d'apprentissage.

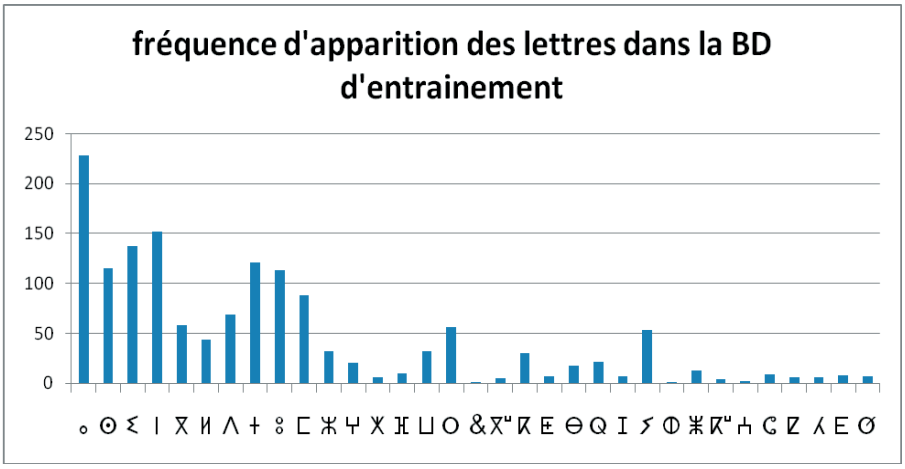


Fig. fréquence des lettres dans la BD d'apprentissage

La partie la plus cruciale est la création des fichiers Box, à partir des images de la base de données déjà établies pour ce besoin. Ces fichiers donnent des informations géométriques sur chaque caractère et à partir desquels les caractéristiques seront extraites durant l'entrainement. Chaque fichier box correspond à une image (.tif) contenant une seule fonte. Un ensemble de 16 fichiers ont été crées durant ce travail.

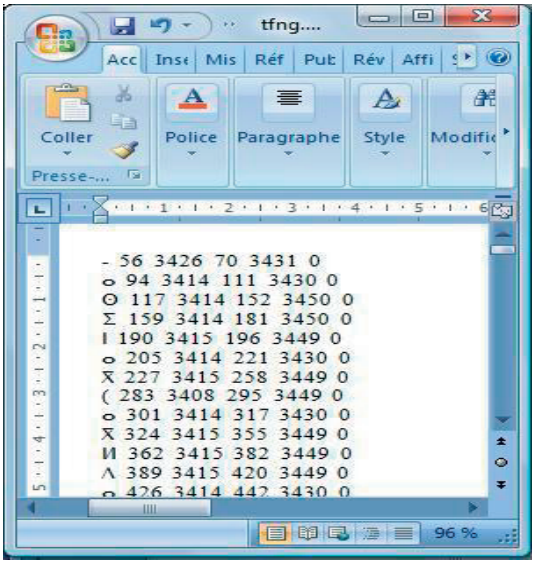


Fig. : Aperçu du fichier BOX pour le tifinaghe

Le fichier Box doit par la suite être édité manuellement pour rectifier les erreurs éventuelles qui peuvent être automatiquement générés. Ces erreurs sont dû généralement aux différents diacrités du Tifinaghe, aux ambigüités causés par le système lui-même (qui n'est pas encore entrainé) ou aux unités graphiques composés (e.i ⵅ, « , etc)

L'utilisation du Tesseract en mode entrainement génère les fichiers .tr nécessaires pour l'apprentissage, contenant les caractéristiques des unités graphiques. Tous les fichiers .tr sont ensuite agrégés dans un seul fichier dit *normproto*. Deux autres fichiers de données ont été aussi générés : le fichier *unicharset* listant l'ensemble du répertoire des caractères Unicode à apprendre avec leurs types (ponctuation, numérique, alphabet, etc) et un fichier de lexique amazighe optionnel qui est utilisé dans la modélisation de langage pour améliorer la sortie du système. Un dernier fichier optionnel qui n'a pas été intégré à ce stade est le fichier de désambigüisation « *unicharambigs*»

4.2.Résultats et discussions :

Dans un premier temps les tests ont été effectués sur des caractères isolés et des textes composés avec des fontes d'apprentissage différents. 98 % des caractères Tifinaghe isolés de différentes fontes sont reconnus sans ambiguïté. D'autres tests sont en cours sur des images textes avec un mélange de fontes appris et d'autres qui ne le sont pas.

Le rendement observé dans le premier test est causé par le faite d'absence du fichier de désambigüisation dans ce travail. Son intégration à la base de données d'apprentissage est en perspective.

5. Conclusion :

Dans ce papier, nous avons intégré le script Tifinaghe dans le moteur de classification Tesseract du système OCR Ocropus. A ce point, tous les résultats ne sont pas aussi satisfaisants qu'on peut le souhaiter mais n'est il pas une première pour le Tifinaghe : d'avoir un système complet de reconnaissance.
Le taux d'erreur de la reconnaissance de l'ensemble des patterns de validation (non appris par le système) reste encore à tester d'où une suite de ce travail s'impose. Nous projetons, ainsi, d'améliorer l'apprentissage du Tifinaghe par le moteur Tesseract et de procéder à l'entrainement du système interne de classification d'Ocropus.

Bibliographie :

[1] S. Tangwongsan and B. Suvacharakulton, "A Highly Effective System for Printed Thai Character Recognition by Word Prediction Method", ICITA 2009.

[2] Arif Billah Al-Mahmud Abdullah and Mumit Khan, "A Survey on Script Segmentation for Bangla OCR", Dept. of CSE, BRAC University, Dhaka, Bangladesh,

[3] K. Hyuk-Chul , H. Ho-Jeong , K. Min-Jung , L. Seong-Whan , « Contextual postprocessing of a Korean OCR system by linguistic constraints", 3th ICDAR '95, IEEE, ISBN: 0-8186-7128-9

[4] *R. Romero, R. Berger, R. Thibadeau, D. Touretsky, "* Neural Network Classifiers for Optical Chinese Character Recognition",

[5] T. Kanungo, G. E. Marton, and O. Bulbul, "Performance Evaluation of Two Arabic OCR Products", Proc. of AIPR Workshop on Advances in Computer Assisted Recognition, SPIE Vol. 3584, Washington DC, 1998.

[6] Y. Ait ouguengay, M. Taalabi « Elaboration d'un réseau de neurones artificiels pour la reconnaissance optique de la graphie amazighe : Phase d'apprentissage », « Systèmes intelligents-Théories et applications, 2009, ISBN-102909285553

[8] Y. Es Saady, A. Rachidi, M. El Yassa, D. Mammass, « Reconnaissance de Caractères Amazighes Imprimés par le Formalisme des Automates à états finis » , SITACAM'09, IRCAM, 2009

[7] B.Bouikhalene, M.Fakir, S.Safi Et B.El Kessab, « Reconnaissance Des Caracteres Tifinaghe Par L'utilisation Des Reseaux De Neurones Multicouches », SITACAM'09, IRCAM, 2009.

Reconnaissance automatique de la parole Amazigh à base de la transcription en alphabet Tifinagh

A. EL GHAZI (1), C. DAOUI(3), M. FAKIR(1), B. BOUIKHALENE(2), N. IDRISSI(1)

(1) Equipe de Traitement de l'Information et Télécommunications, Département d'Informatique, Faculté des Sciences et Techniques, Université Sultan Moulay Slimane, PB 523, Béni Mellal, Maroc, hmadgm@yahoo.fr, fakfad@yahoo.fr, najlae_idrissi@yahoo.fr

(2) Equipe de Traitement de l'Information et Télécommunications, Faculté Poly disciplinaire, Université Sultan Moulay Slimane bbouikhlene@yahoo.fr

(3) Laboratoire de Modélisation et Calcul, Département de Mathématique Faculté des Sciences et Techniques, Université Sultan Moulay Slimane, PB 523, Béni Mellal, Maroc, daouic@yahoo.com

Résumé – Abstract

Ce papier présente un système de reconnaissance automatique de la parole amazigh basé sur la transcription en alphabet Tifinaghe délivré par l'Institut Royal de la Culture Amazigh(IRCAM) et le modèle de Markov Caché.

In this paper we present a systéme of automatic speech recognition based on Amazigh transcription alphabet Tifinagh issued by the Royal Institute of Amazigh Culture (IRCAM). The method is based on the Hidden Markov Model.

Mots Clés – Keywords

Reconnaissance de la parole, parole amazigh, HMM, modèle acoustique.
Speech recognition, speech Amazigh, HMM, acoustic model.

1. INTRODUCTION

Récemment la reconnaissance automatique de la parole RAP intéressent différents chercheurs : phonéticiens, informaticiens, mathématiciens et d'autres. Leur objectif commun est la réalisation d'un système qui produit les capacités proches de celles de l'être humaine dans ce domaine. La reconnaissance

automatique de la parole est l'un des domaines de recherche prometteur du traitement automatique de la parole, son principe est d'extraire le message oral contenu dans un signal de parole. Cela a plusieurs applications dans la vie courante et dans le domaine technologique (contrôle à l'aide de la parole, composition orale des numéros, les logiciels de traitement de textes vocaux, indexation automatique de la parole etc.…

Le principe de la RAP s'effectue sur un flux acoustique continu. La conception et l'implémentation d'un système de reconnaissance automatique de la parole SRAP nécessite une modélisation de l'ensemble des phénomènes observés dans le signal (les respirations, les hésitations, les fragments de mots, les brouillons de paroles etc.…). Dans ce travail, Le système SRAP permet de transcrire un message oral, extraire une information linguistique à partir d'un signal de la parole. Le modèle de Markov caché (Hidden Markov Model : HMM) permet de modéliser les unités constituants les mots et les phrases d'une langue. Notre approche consiste aussi à faire modéliser la parole amazigh afin de réaliser un système de reconnaissance qui permet de transformer un signal en une suite significative de lettre Tifinagh. Il faut aussi mentionner que plusieurs outils d'implémentation existent, les plus connus sont : HTK (Barbara Resch(2003)) source, CMU sphinx (H. Satori et M. Harti 2004). Dans ce travail nous avons utilisé CMU sphinx à base le modèle de Markov caché. Le système CMU sphinx est librement disponible (open source), populaire, robuste et puissant. Notre système permet la mise en place des bases de construction d'un système de reconnaissance automatique de la parole amazigh basé sur sphinx 4.

La réalisation d'un système de reconnaissance automatique de la langue amazigh a marqué l'événement ces dernières années vue le nombre d'applications qui peut être mise en place grâce à ce système, parmi ces applications (Ali sadiqui , Noureddine chenfour 2010) on cite :

- Système de dialogue Homme-machine : pour les gens qui ne maitrisent les autres langues.
- Systèmes de traduction : pour traduire des énoncés amazighs.
- Apprentissage de la langue amazigh.

Ce travail est organisé comme suit : la deuxième section est consacrée à la présentation des différentes composantes du système CMU Sphinix. Dans la troisième section on décrit les étapes de la reconnaissance à savoir la conception et l'implémentation du système SRAP. A la quatrième partie on présente les résultats expérimentaux.

2. PRESENTATION DU SYSTEME CMU SPHINX

Le système sphinx est un projet open source réalisé à l'université Carnegie Mellon (CMU) pour la réalisation d' un environnement pour guider les recherches dans le domaines de la reconnaissance de la parole. Le système sphinx est un ensemble de classes et de librairies crées avec le langage Java sous forme de fichier compressés de type .jar.

Le sphinxTrain est un ensemble de scripts perl délivré avec sphinx pour faciliter la création du modèle acoustique.

2.2. INSTALLATION

2.2.1 Sphinx 4

L'installation du sphinx 4 nécessite les sources suivantes :

- Les librairies sphinx 4 téléchargés sous forme compressé.
- Java 2 SDK
- Ant : serveur qui permet d'exécuter, en utilisant un fichier de configuration XML, les différentes programmes pour construire les exécutables .exe.
- IDE : pour construire l'application java.

2.2.2 SphinxTrain

L'utilisation du sphinxTrain nécessite l'installation des logiciels suivants :

- ActivePerl : Il permet de prendre en compte les fichiers .pl par le système et de les exécuter.
- Visual C++ 6.0 : Pour compiler les différents fichiers.

3. RECONNAISSANCE DE LA LANGUE AMAZIGH

La langue amazigh et l'une des plus anciennes langues du monde, son histoire a commencé dans l'Afrique du nord par la création du noyau de l'alphabet tifinagh qui a commencé à se propager dans la population amazigh au nord de l'Afrique. Actuellement le systéme alphabétique amazigh, appelé Tifinagh, connait actuellement une implantation dans les programmes scolaire marocain et il est utilisé dans la recherche historique amazigh.

Le systéme alphabétique amazigh comme il est donné par l'IRCAM (Meftaha Ameur , Aïcha Bouhjar , Fatima Boukhris 2004)), (Ali sadiqui , Noureddine chenfour 2010):

- 27 consonnes : les labiales (ⴱ,ⴼ,ⵎ),les dentales (ⵜ,ⴷ,ⵟ,ⵜⵜ,ⵏ,ⵔ,ⵙ ,ⵣ), les alvéolaires (ⵚ, ⵛ, ⵥ, ⵞ), les palatales (ⵊ,ⵉ), les vélaires (ⴽ,ⵅ), les

labiovélaires (ⵇᵘ, ⵅᵘ), les uvulaires (ⵖ, ⵅ, ⵇ), les pharyngales (ⵄ, ⵃ) et la laryngale (ⴰ).

- 2 semi-consonnes : ⵢ et ⵡ.
- 4 voyelles : trois voyelles pleines (ⴰⵉⵓ) et la voyelle neutre ⴻ.

3.1 BASE D'APPRENTISSAGE

La base d'apprentissage pour notre système est constituée de la prononciation de dix chiffres amazigh par 16 personnes de différents âge et dans des conditions différentes. Chaque locuteur est invité à prononcé 10 fois le même mot.

Les signaux subissent à un traitement avant l'enregistrement :

- Suppression du bruit de début et de fin ainsi que le bruit entre deux prononciation successives.
- Enregistrement à 16khz, 16 bit format '.sph'.

La figure 3 représente les mots utilisés et leur notation en tifinagh ainsi que leurs transcription phonétique.

3.2 MODEL DE MARKOV CACHE

Le modèle de Markov caché MMC est un outil stochastique permettant de modéliser un signal acoustique par le principe des états, chaque signal représentant un mot est modélisé par un ensemble de HMM, chacun de ces derniers corresponds à une suite des états qui représente un phonème.

Les signaux vocaux subissent à un ensemble de transformation pour extraire les coefficients de Mel appelés MFCC (Mel Fréquence Coefficient Cepstrum). Dans ce stade le signal vocal sera représenté par une série de vecteurs acoustique, chaque vecteur représente 10ms du signal acoustique avec encombrement à ½ de deux fenêtre successives (Fig. 1).

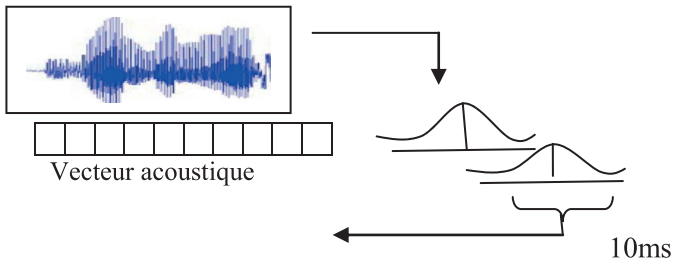


Fig 1 : fenêtrage de Hamming applique sur chaque 10ms du signal

Le principe du MMC consiste à associer à un état une suite de vecteurs acoustique, ces dernières sont représentés par les paramètres suivants :

- Vecteurs des moyens
- Matrice de covariances
- Matrice de transition
- Poids des gaussiens

La figure 2 représente le principe de la modélisation des signaux vocaux avec le modèle de Markov caché

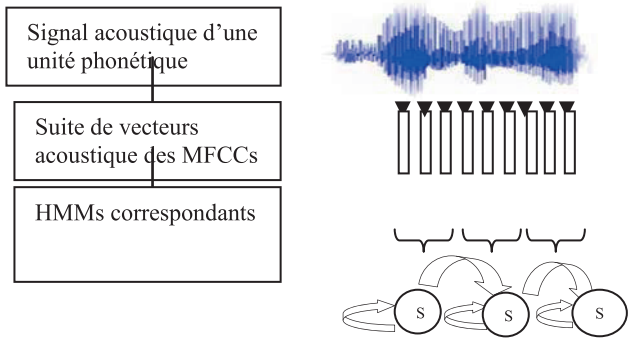


Fig. 2 : Présentation du principe de modélisation des vecteurs acoustique par les MMC

3.3 PRINCIPE DE LA RECONNAISSANCE

Le principe de la reconnaissance peut être expliqué comme le calcule de la probabilité P(W/S) : la probabilité qu'une suite de mots W correspond au signal S et de déterminer la suite de mots qui maximise cette probabilité (Thomas Pellegrini et Raphael Duée juin 2003).

Selon la formule de Bayes la probabilité P (W/S) peut s'écrire :

$$P(W/S)=P(w).P(S/W)/P(S)$$

Avec :

P(W) : Probabilité a priori de la suite de mots W (Modèle de langage).

P(S/W) : Probabilité du signal S, étant donnée la suite de mots W (Modèle acoustique).

P(S) : probabilité du signal acoustique S (indépendant de W).

3.4 CORPUS D'APPRENTISSAGE

Le corpus d'apprentissage est constitué de dix mots amazighs. Le tableau de la figure 3 représente les mots utilisés ainsi que leur transcription. La base d'apprentissage comprend cinq prononciations pour chaque mot par 16 personnes de différentes âge et dans des différentes situations.

L'ensemble des symboles ou phones utilisés pour l'apprentissage des états du modèle de Markov caché est décrit dans le tableau 4 (Meftaha Ameur, Aïcha Bouhjar, Fatima Boukhris 2004)

Symbole	Représentation
	I
	L
	E
	M
	Y
	N
	SS
	C (prononciation moyen atlas)
	K
	U
	Z
	S
	TT
	T

Fig 3 : Symboles phonétiques utilisés pour la reconnaissance des chiffres amazighs

Mot	Transcription phonétique
0	**I L E M**
1	**Y E N or Y A N**
2	**SS I N**
3	**C R A D**
4	**K O Z**
5	**S E M**
6	**S E D**
7	**S A**
8	**TT A M**
9	**T Z A**

Fig4 : Corpus d'apprentissage et ses transcriptions phonétiques

La transcription phonétique permet de présenter un mot ou une prononciation sous forme d'une suite de phonèmes ce qui permet au systéme de modéliser chaque unité par un modèle de Markov Caché (HMM) . La figure 5 représente une illustration du modèle de Markov caché pour le mot (Y A N) (Vincent Luba et AinaneYounes 2004-2005).

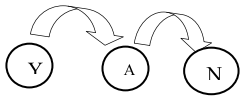


Fig. 5 : représentation du HMM pour le mot 'YAN'

La matrice de transition gère les transitions entres les phonèmes et le modèle de langage gère la transition entre les mots pour la reconnaissance de la parole continue.

3.5 FICHIER D'APPRENTISSAGE

La phase d'apprentissage nécessite un certains nombre de fichiers qui vont être utilisés par des scripts perl pour entrainer les fichiers acoustiques. Les fichiers utilisés sont :

- Project_name. phone : contient une liste des phonèmes représentants les mots.
- Project_name.fileids : contient la liste des fichiers acoustiques.
- Project_name.DIC : contient la liste des mots et leur représentation phonétique
- Project_name.TRANSCRIPTION : contient la localisation des fichiers audio ainsi que leurs contenus.

En plus de ces fichiers on a besoin du fichier de grammaire qui définit l'usage des mots dans une application appelé aussi modèle de langage. La figure 6 définit le fichier de grammaire utilisé :

```
#JSGF V1.0;

/**
 * JSGF Grammar tifinagh
 */

grammar tifinagh;

public <tifinagh> = (0 | 1 | 3 | 4 | 5 | 6 | 7| 8| 9);
```

Fig. 6 : fichier de grammaire utilisé pour la reconnaissance des chiffres amazigh de 0 à 9.

La figure 7 représente les caractéristiques de la base d'apprentissage utilisée.

Durée de la Base	Nombre de personnes	Nombre de prononciation
1h45min de prononciation.	- 3 personnes adultes - Un homme qui ne parle pas l'amazigh. - 2 filles - 14 garçons	- 20 répétitions multipliées par 5 pour les adultes implique au totale 100 prononciations pour chaque personne. - 5 répétitions pour les garçons multiplié par 10 ce qui donne au totale 50 enregistrements pour chaque personne. Le nombre d'enregistrement total dans la base est égal à 2100 mot prononcé par 20 personnes.

Fig 7 : caractéristiques de la base d'apprentissage

4 RESULTATS EXPERIMANTAUX

La base de données de teste est constitué de 300 prononciations réalisées par 6 personnes, chacune de ces dernières est invités à prononcé 5 fois le même chiffre. Les résultats obtenus, en calculant le taux de reconnaissance donnée par la relation (1), sont donnés dans le tableau 7.

$$t = \frac{\text{nombre de mot reconu}}{\text{taille de la base de teste}} \qquad (1)$$

Base de test	Base d'apprentissage	Résultats
300 prononciations	1h45min	T=90%

différentes en introduisant les fichiers audio plus bruité		

Fig 7 : Résultats obtenus pour un système de reconnaissance de la parole amazigh

La figure 8 présente une interface java qui permet de faciliter la tache de reconnaissance.

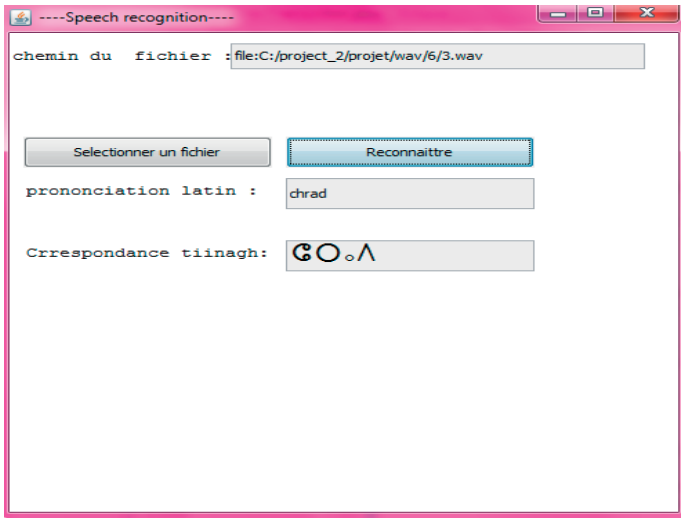


Fig. 8 : interface java pour le système de reconnaissance automatique de la langue amazigh

5. CONCLUSION

Ce travail permet la mise en place d'une application d'un système de reconnaissance vocale pour la langue amazighe diffusée dans le nord de l'Afrique. L'amazigh est l'une des langue la plus complexe au niveau phonétique et au niveau de la différenciation régionale. Le travail réalisé permet de donner une idée sur la phonétique qui peut être utilisé pour la reconnaissance de cette langue. Les résultats obtenus sont très satisfaisantes, vue le nombre limité des locuteurs et la taille de la base de données audio.

REFERENCES

Ali sadiqui & Noureddine chenfour (2010) "Reconnaissance de la parole arabe basé sur CMU sphinx", Anale. Seria Informatica. Vol VIII fasc. 1.

H. Satori & M. Harti " Système de la reconnaissance automatique de la parole", Faculté des Sciences, B.P. 1796, Dhar Mehraz Fès, Morocco.

Divejver and J. Killer (1982),"Pattern recognition" in Pattern Recognition: a statistical approach"; Prentice Hall.

Thomas Pellegrini et Raphael Duée (juin 2003), "Suivi de la voix parlée grâce aux modèles de Markov Caché", lieu : IRCAM 1 place Igor Stravinsky 75004 PARIS.

Meftaha Ameur - Aïcha Bouhjar - Fatima Boukhris (2004), IRCA: publication : "initiation a la langue amazigh", Institut Royale de la Culture Amazigh, Maroc.

Benjamin LECOUTEUX (2008) "Reconnaissance automatique de la parole", UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE-France.

Barbara Resch(2003) "Automatic Speech Recognition with HTK", A Tutorial for the

Course Computational Intelligence http://www.igi.tugraz.at/lehre/CI.

Vincent Luba et AinaneYounes (2004-2005)," Modèles de Markov cachés Reconnaissance de la parole", FACULTE POLYTECHNIQUE DE MONS

Some Aspects of Amazigh Clause Structure

Naima Omari
Faculty of Chariaa
B.P. 52, Lamzar, Ait Melloul
Al Quaraouiyine University
Agadir, Morocco
Naimaomari2004@yahoo.com

Résumé

L'objet de cet article est de présenter quelques propriétés générales de la structure de la phrase de l'Amazighe dans le cadre du programme minimaliste (MP). En particulier, nous discutons la morphologie des verbes et la structure de la phrase simple en Tachelhit. Le travail vise à adapter l'Amazighe aux analyses informatiques modernes, et préparer le terrain pour les chercheurs d'explorer la maniére d'utiliser des modèles formels comme le MP pour le traitement de la phrase.

Abstract

The object of this paper is to present some general properties of Amazigh clause structure within the framework of the minimalist program (MP). In particular, we discuss the morphology of verbs and the functional structure of Tashelhit simple verbal sentences. The work is mainly construed to pave the way for researchers on computational linguistics to explore ways of using formal models like the MP in modeling the nature of sentence processing.

Mots Clés : La morphologie du verbe, temps et aspect, vérification des traits, principes d'économie, fusion et déplacement, le programme minimaliste, traitement de la phrase.

Key Words : Verb morphology, tense and aspect, feature checking, economy principles, merge and move, the minimalist program, sentence processing.

1. Introduction

The object of this paper is to discuss the morphology of verbs and the functional structure of Amazigh clauses within the MP (Chomsky 1992, 1995). The discussions are based on data from the Tashelhit variety, spoken in the southwest of Morocco. This work is mainly construed to pave the way for researchers on

computational linguistics to explore ways of using the MP to understand sentence processing.

The paper is organized as follows. Section 1 considers the verb morphology in matrix clauses in Amazigh. To this end, we first identify the head V and the morphemes it contains. Then, after classifying the verbal stems into three forms, we proceed to identify the different aspectual and temporal information they express. Section 2 presents the different feature specifications which characterize the elements that head the functional projections in order to make predictions about head movement. In the course of the discussion, we explore perspectives on applying MP theory to the construction of sentence parsing algorithm.

2. Verb morphology

2.1. The head V

A typical aspect of Amazigh is the nonconcatenative nature of morphological structure ((Iazzi, 1991), (El Moujahid ,1993)). Typically, the consonantal root carries the meaning and is recurrent in all the forms of the same derivational family, while patterns of vowels and consonants spell the grammatical categories of the forms in question.

The representation of the verb ‘skar’ (is doing / making) and ‘skir’ (not do/make), for example, consists of the three grammatical elements, the consonantal root /skr/ which ranges over the semantic field of the predicate, the vocalic melody which expresses aspect, and the CV tier which represents the morphological template on which the other two tiers are mapped. The association of the consonantal root and the vocalic melody defines the lexical category and its content:

(1) a. Semantic field: skr

Aspect:

-i-

CV tier :

CCVC

b. Semantic field: skr

-a-

CCVC

Aspect:

CV tier :

Examples such as (1) show the difficulty with Pollock’s (1989) style affixation-based analysis. This analysis depends on the assumption that the range of possible morphological operations is restricted to affixation- that is, to the addition to a root of elements with segmental material that can be directly represented in tree structures and manipulated by rules of syntax. Under such an analysis, it is not clear how to account for Amazigh which provides evidence of morphological processes other than affixation, such as vowel alternation (1), vowel epenthesis,

and gemination (see section 2.2). This problem is resolved by Chomsky's (1992/1995) checking theory where it is argued that inflectional morphemes are lexically generated on the verb, thus removing word formation, in the sense of merging elements that make up words, from the domain of syntax.

Having identified the head V, we have to identify which morphemes it contains. As we have just seen, there is no guarantee that this can be done by morphological parsing (segmentation). In what follows, we will see that a morpheme (feature value) can be identified in a verb if this verb minimally contrasts with another for this morpheme (feature value).

2.2. The Verbal stems

The verb in Amazigh consists of the stem (the root and its vowel and/or consonantal melody) and an agreement affix[1]. In this section, we describe the morphology of the three verbal stems in Amazigh, which can be labelled the aorist, the imperfective, and the perfective.

Generally speaking, the aorist stem is defined as being a verbal form that expresses the verbal action (or state) without reference to its aspectual or temporal values. Since the aorist in Amazigh does not carry any temporal or aspectual information, it cannot occur as a root indicative sentence:

(2) * i – ftu s lxdmt.
he- go+A to the-work
"He goes to work."

(2) has no temporal reference and cannot receive an aspectual characterization. Being in the aorist, the verb ftu is uninterpretable. The fact that the aorist is tenseless and aspectless leads to the assumption that it functions as the base form in Amazigh ((Aspinion, 1953), (Chaker, 1973), (Cadi, 1981), (Ouhalla, 1988)). However, there is strong evidence that the verbal base cannot be handled in this way; the derivation of the perfective and the imperfective from the aorist cannot uniformly be accounted for, as shown in (3), from (Makhad, 1996:28):

(3)	Aorist	Perfective	Imperfective
Gloss			
a.	amz	umz	tt- amz
'catch'			

[1] See Omari (2001) for two related issues in the grammar of Amazigh clause: agreement and word order.

b.	asi	usi	tt- asi	
‘carry’				
c.	af	ufi /a	tt- afa	
‘find’				
d.	af	uf	tt- af	‘be
better than’				
e.	ddu	ddi /a	tt- dda	
‘go’				

At first glance, the forms in (3a–d) seem to satisfy the notion of the aorist as the base: the initial vowel a in the aorist changes systematically into u in the perfective. Nevertheless, this position fails to adequately account for other forms in this system. (3 (c) and (d)), for example, have identical aorist forms, yet they behave differently with regard to their perfective and imperfective forms. On the other hand, (3 (c) and (e)) have identical behavior in perfective and imperfective, though they are forms with different bases.

Admittedly, then, the above empirical limitation points to the idea that the aorist cannot function as the base form in Amazigh. To solve the base problem, we propose that the three verbal stems have the same lexical entry, which is a complex of three types of features: phonological, semantic, and syntactic features. Other features are chosen as the verbs enter Numeration. For example, if a verb stem is specified for the formal feature [+Aorist], this will be an instruction that it bears no thematic affix.

To express future tense reference, the aorist is combined with a tense antecedent, either the tense marker rad or the matrix tense. In conjunction with the future particle rad, the aorist expresses future time. The same interpretation is true when it appears in control and purposive constructions. This is shown in the following examples:

(4) a. rad γr -ħ lktab.
Fut read+ A -I the-book
“I will read the book.”

b. ri -ħ ad sγ -ħ lktab.
want+perf-I that buy+A -I the book
“I want to buy the book.”

Turning to the imperfective stem, it involves the imperfective aspect which is concerned with the non-finished aspect of the verbal event. Unlike the aorist, the imperfective stem has an aspectual value, namely progressive or habitual interpretation. Moreover, unlike the aorist stem, the imperfective stem is morphologically marked. It involves the following morpho-phonological processes: the prefixation of tt (and the emergence of a vowel), the gemination of a consonant,

and a pre-final-position-vowel epenthesis. In the processes of the imperfective, some forms, sometimes, do not restrict these derivations to one single operation, but combine more than one process to delineate their aspectual values. (5) shows the morphologically-marked form of the imperfective stem when compared to the aorist stem:

(5)	Aorist	Imperfective	Gloss
a.	amz	tt-amz	'catch'
b.	fsi	fssi	'unfasten'
c.	siggl	siggil	'look for'
d.	ggal	tt-galla	'swear'

There is an extra morpheme that occurs along with the imperfective morphology. This form can be used as habitual or progressive:

(6) ar i -ttara tabrat.
part he-write+Imperf the-letter
"He writes / is writing the letter."

In Amazigh, the imperfective is a 'dependent' form, so that (6) never occurs as a root indicative sentence, as depicted by the ungrammaticality of (7):

(7) *i - ttara.
he - write+ Imperf

Concerning the grammatical value of the particle ar, there are two hypotheses at least. On the one hand, we might argue that the particle ar is a realization of present tense ((Guerssel, 1983), (Makhad, 1996)). On the other hand, we might argue that there is no morphological realization of the present tense and that the particle is a purely aspectual morpheme ((Dell, El Medlaoui, 1989), (Boukhris, 1998)).
The first hypothesis predicts that whenever the particle is present, a present tense interpretation should obtain. This prediction is not borne out by (8):

(8) ar t -alla idgam.
Part she-cry+ Imperf yesterday
"She was crying yesterday."

Given the assumption that tense and temporal adverbs must have the same temporal reference, the sentence in (8) is expected to be ungrammatical due to the incompatibility between the adverb 'yesterday' and the tense information represented by ar, a wrong prediction. Let us then consider the second hypothesis. If ar is an aspectual particle, the question arises about the status of the vocalic and/or consonantal melody associated with the imperfective stem. We argue that it carries aspectual information. This is clear from the fact that the form aqra in (9) expresses the imperfective aspect though ar is absent:

(9) ur ufi -ħ ad aqra -ħ γiyyd.
Neg able+perf-I that studying-I the-night
"I cannot study during the night".

Thus, reference to Aspect in (6) and (8) is indicated both by ar and the thematic affix on the verb. Note that, in many cases, the unmarked interpretation of the imperfective stem in Amazigh is that of present tense, as in (6). To render (6) in the future, the future particle rad is added, as in (10):

(10) rad i - ttara tibratin.
Fut he-write+Imperf the-letters
"He will be writing letters."

Concerning the perfective stem, it is morphologically marked. The following processes are used in forming perfective stems: vowel alternation, null affix, and consonant gemination, as illustrated in (11):

(11)	Imperfective	Perfective	Gloss
a.	arm	urm	'try'
b.	skar	skr	'do/ make'
c.	tt- ini	nni /a	'say'

The perfective stem involves the perfective aspect which expresses the accomplishment of the verbal action (Comrie,1985). One could assume that perfective morphology is a realization of the past tense in Amazigh. However, this analysis would wrongly predict that whenever the affix is present, a past tense interpretation should obtain (Boukhris, 1998: 124):

(12) ass-a lli- ħ, aska mmut- ħ .
day-this be+perf-I tomorrow die+perf-I
"I am alive today; I will die tomorrow." = "I am not eternal."

The two clauses in (12) express present tense and future tense respectively irrespective of the perfective morphology on the verbs. Additional data against this analysis comes from stative verbs. These verbs carry the same inflection as the verbs in the past tense but are restricted to sentences that express present tense:

(13) a. ffeγ - ħ. b. ssen - ħ ħmad.
go+perf - I know+perf - I Hmad
"I went out." "I know Hmad."

Notice the identical vowel e on the active verb ffeɣ-ħ and the stative verb ssen. With the former, the perfective expresses a completed action in the past, whereas with the latter, the perfective stem expresses present time. This shows clearly that the vowel on the perfective stem carries aspect only. This leads to the following question: How is tense expressed? We argue that tense is an abstract morpheme that does not have any specific phonological realization. However, the abstract tense is syntactically active in that it licenses adverbs:

(14) i - mmudda igdam.
he- travel+perf yesterday
"He travelled yesterday."

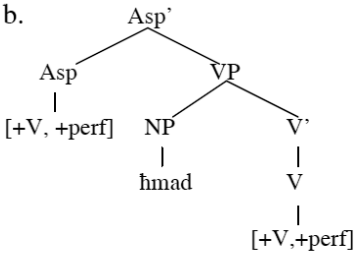
3. Functional categories

3.1. Aspect

In so far as Amazigh encodes aspect, it by and large expresses a binary distinction, namely imperfectivity and perfectivity. We take this to mean that there is a category aspect, immediately above V, with a categorial feature [+V] and a syntactic feature [+/-perf(ective)], as part of the syntactic representation of these sentences:

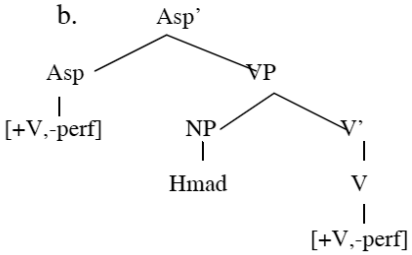
(15) a. i - fta ħmad.
he-leave+perf Hmad
"Hmad left."

b.



(16) a. ar i- ttmuddu ħmad.
Asp he-travel+Imperf Hmad
"Hmad is travelling."

b.



When the verb is specified for [+perf], it carries the morphological feature of the perfective stem, (15). When it is specified for [-perf] it carries the morphological feature of the imperfective stem and is preceded by the aspectual morpheme ar,

(16). Notice that ar is not inflectional in the sense that it does not attach to the verb: the verb and ar can be seperated by adverbs and clitics:

(17) a. ar yadlli y -aqra .
Asp formerly he-study+imperf
"He used to study."

b. ar -tn y -aqra.
Asp -them he-study+imperf
"He is reading them."

This suggests that ar has its own lexical entry and is specified for the categorial feature [+V] and the aspectual feature [-perf]. Following Boukhris (1998), we argue that the aspectual projection is headed by a null morpheme in the context of sentences with perfective interpretation, and by ar in the context of sentences with imperfective interpretation. Independent evidence that ar heads AspP in the syntax comes from the fact that it is attracted by clitics. In the example (17b), the clitic appears attached to the aspectual morpheme ar. On the assumption that clitics are head categories, then ar movement is essentially a head movement which results in the adjunction of the moved category to another head category[2].

Concerning V-movement, there is clear evidence that it takes place overtly. For example, in Amazigh, lexical verbs may precede VP adverbs, which are adjoined to VP (Pollock, 1989):

(18) a. i -γra mliħ.
he -study+perf well
"He studied well."

b. *mliħ i - γra.
Well he-study+perf

(19) a. ar y -aqra mliħ.
Asp he-study+imperf well
"He is studying well."

b. *ar mliħ y -aqra.
Asp well he-study+imperf

The fact that the verb overtly precedes the VP adverb is an indication that movement to Asp has taken place in the overt syntax.
Similary, if the postverbal subject in Amazigh, as in (20) and (21) below, is the specifier of VP, as predicted by the VP-internal subject hypothesis ((Speas, 1986), (Koopman, Sportiche, 1991), among others), then this lends further support to our claim that V has overtly moved over the subject to Asp:

(20) a. i - mmudda ħmad.

(21) a. ar i - tt - mmuddu ħmad.

[2] See Omari (2001) for an analysis of clitics in Amazigh.

he-travel+perf Hmad	Asp he- imperf+travel Hmad
"Hmad travelled"	"Hmad travelled."
b. *ħmad i - mmudda.[3]	b. ar ħmad i - tt - mmuddu.
Hmad he-travel+perf	Asp Hmad he-imperf+ travel

3.2. Tense

Amazigh distinguishes two general classes of tense: future and non-future. This latter class includes present and past. This opposition results from the observed fact that future is morphologically realized by the verbal particle rad and past and present are not:

(22) rad i - mmuddu ħmad.
Fut he- travel+A Hmad
"Hmad will travel."

(23) a. i -mmudda ħmad.
he-travel+perf Hmad
"Hmad travelled."

b. ar i - tt – mmuddu ħmad.
Asp he- travel+imperf Hmad
"Hmad travels/ is travelling."

Consideration of the sentences in (22) and (23) reveals that only in (22) is tense morphologically realized. Both (23a) and (23b) are not overtly inflected for temporal values, which legitimizes the [+/-Fut(ure)] opposition in Amazigh, where [+Fut] corresponds to a morphological realization of tense and [-Fut] to its absence. Hence, tense in Amazigh exhibits overt morphology. By overt morphology we mean morphology that can be phonologically detected by opposition. [-Fut] tense, for instance, is overtly marked as such, not by an affix, but precisely, by the absence of an affix and by its standing in opposition to future tense. This opposition predicts word order.

[3] The intended reading for (20b) to be ungrammatical is not the one where ħmad is in a topic position with a pause setting it off from the common part.

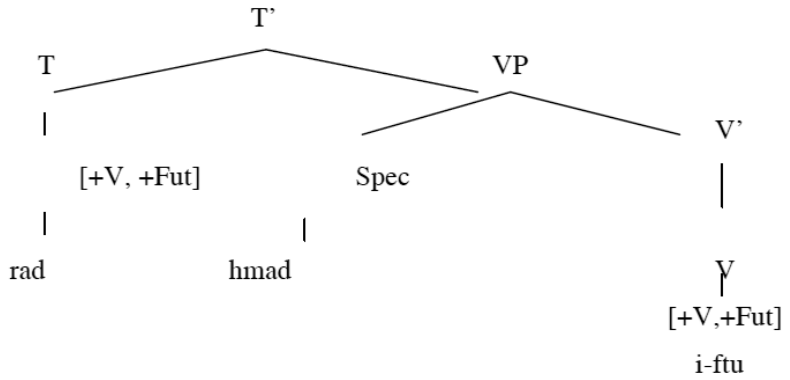
When T is specified for the feature [+Fut], the tense morpheme is a free particle: There is no need for overt V-movement to T; the free particle rad is inserted under T and will check the relevant feature. That V does not move overtly to T is confirmed by the fact that rad and V can be seperated by accusative clitics and adverbs:

(24) a. rad - tnγr -ħ.
Fut - them read+A -I
"I will read them."

b. rad ddaħ i- mmuddu.
Fut again he- travel +A
"He will travel again."

Note, however, that in cases like (22), the verb precedes the lexical NP ħmad. This entails that V has moved to T. Consider the derivation of (22), represented as in (25):

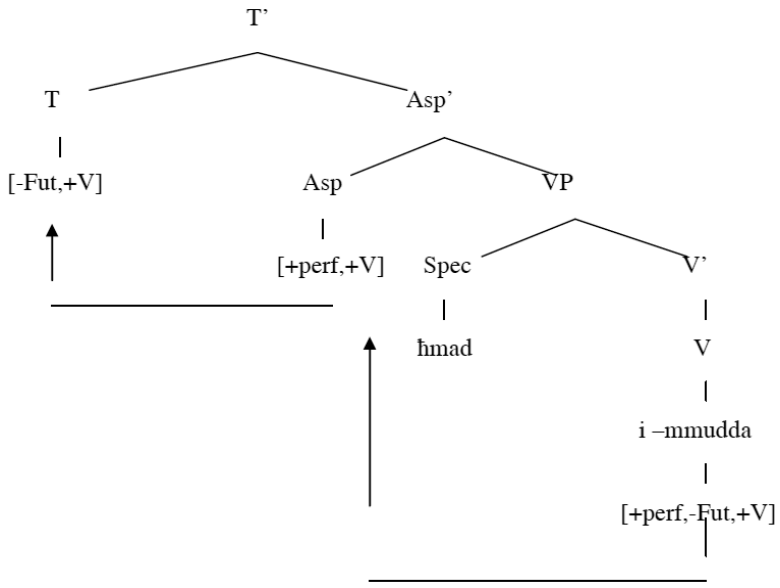
(25)



As shown in (25), the verb iftu moves to T in overt syntax. This may be accounted for if we assume that in languages like Amazigh, there is a morphological constraint on verbal chains formalism which is needed for e(vent)-binding (Jamari p.c). According to this constraint, no lexical XP can intervene between elements of a verbal chain. Notice that in (25) where Asp has no realized morphological content the clause is just a projection of T'. This is in line with Chomsky's (1995) proposal that functional categories will appear in syntactic structure only so far as they are needed for the derivation to converge; e.g. for the checking of a syntactic feature that will be left unchecked.

It is obvious that in (23a), where no tense particle projects, the verb is allowed to occupy tense via a stepwise raising operation. We take this to mean that the feature [-Fut] is strong, and so necessitates explicit checking. Thus, overt V–raising takes place up to T to check the strong V-features of this functional category. The two stages of the derivations are given in (26):

(26)



The evidence for overt movement to T involves constructions with adverbs and accusative clitics. In such constructions, the verb is ordered to show up in a position higher than adverbs and accusative clitics:

(27) a. i -mmudda yadlli.
h -travel+perf formerly
"He used to travel."

b. * yadlli i - mmudda.
formerly he-travel+perf

(28) a. i- sγa -tn.
he-buy+perf-them
"He bought them."

b. *tn i - sγa.
them he - buy+perf

As for (23b), since we are claiming that the feature [-Fut] is strong, this requires that checking take place in the overt syntax. The verb i-ttmuddu, though a potential checker, is not attracted to T by the main features it can attract, namely [+V] and [-Fut]. Ar, being of a verbal nature, blocks the potential landing site of the main verb

due to the minimal link condition (MLC)[4]. Ar is closer to T, and it can enter into a checking relation with T. Thus, the MLC prohibits T from attracting V. We could suppose that V moves and merges with ar and then the complex [V+Asp] subsequently moves to T. However, this analysis would predict that in all instances, this merger should be obligatory, but in fact, in Amazigh, the merger of ar and V does not take place, as shown in (17), repeated here as (29):

(29) a .ar yadlli y -aqra. b. ar - tn y -aqra.
Asp formerly he -study+perf Asp-them he -read+perf
"He used to study." "He is reading them."

For this reason, we argue that overt V- movement is restricted to Asp, while ar moves from Asp to T to check the categorial feature [+V]. Overt V-movement to T would result in the ordering of the verb prior to the aspectual form ar, which is not the situation, as indicated by the ungrammaticality of (30):

(30) * i - ttmuddu ar.
he- travel+imperf Asp

4. Conclusion

In this paper, we have explored the morphology of the three verbal stems in Amazigh: the aorist, the perfective, and the imperfective. We have shown that the verbal stems in combination with the verbal particles express different temporal and aspectual information. We have observed that the perfective and the imperfective carry an aspectual value which is expressed by the vocalic and/or consonantal melody. The aorist, by contrast, is not specified for any aspectual feature. We have noted that the verbal stems are not overtly inflected for temporal values, yet they encode temporal distinctions. Future is realized by the future particle rad, which precedes the aorist and the imperfective. As for the present and past, they are realized by an abstract morpheme on the perfective and the imperfective. We have also argued that the morphological differences between the verbal stems reflect the different feature specifications that characterize the elements that head the functional projections.

---

[4] The MLC imposes locality restrictions on the operation Move, by requiring that each movement be into the closest relevant position. This condition reads as follows

Minimal link Condition (Chomsky, 1995:311):

K attracts αonly if there is no β, βcloser to K than α, such that K attracts β

As shown in our analysis of some aspects of Amazigh clause structure, the MP approach provides a rich array of representational assumptions and technical devices to adequately account for the linguistic data. Such a formal model ought to play a role in understanding sentence processing. There is a strong belief in this area that future advances in understanding can be accelerated by the use of formal models that provide explicit characterizations of the cognitive processes that underlie the ability to use language (cf. (Weinberg, 2001), (Gerth, bein Graben, 2009)). Taking into account our proposed syntactic analysis of Amazigh clause structure, there are many interesting issues to be considered by computational linguists interested in Amazigh sentence processing. Some of these issues are stated below:

- the lexicon-syntax relationship as regards processing.
- the content of lexical entries.
- the Minimalist parsing using a bottom-up strategy within two nested loops: one for the domain of Merge and the other for the domain of Move.
- the incorporation of economy conditions to parsing.
- syntax, semantics, and processing: on the nature of constituent order.

Acknowledgements

We would like to thank Abderrahim Jamari and Rachid Laabdellaoui for helful and inspiring discussions concerning this work.

Bibliography

Aspinion R. (1953), *Apprenons le berbère: Initiation aux dialèctes chleuhs*, Rabat: Edition Felix Moncho.

Boukhris F. (1998), *Les clitiques en berbère tamazight: Approche minimaliste*, Ph.D. Dissertation, Rabat.

Boukhris F., Boumalk A., El Moujahid E., Souifi H. (2008), *La nouvelle grammaire de l'amazighe,* IRCAM.

Cadi K. (1981), *Le verbe en tarifit (Maroc-Nord): Formes, structures, et valences*, D.E.S. Thesis, Paris III.

Chomsky N. (1992), A Minimalist program for linguistic theory, *MIT Occasional Papers in Linguistics* n°1, MIT.

Chomsky N. (1995), *The Minimalist program*, Cambridge, Mass.: MIT Press.

Comrie B. (1985), *Tense*, Cambridge: Cambridge University Press.

Dell F., Elmedlaoui M. (1989), Clitic ordering, morphology, and phonology in the verbal complex of Imdlawn Tashelhiyt Berber, *LOAPL 2*, pp.165-194.

El Moujahid E. (1993), *Syntaxe du groupe nominal en berbère tachelhiyt (parler d'Igherm, Souss, Maroc)*, Ph. D. Thesis, Rabat.

Gerth S., Bein Graben P. (2009), Unifying syntactic theory and sentence processing difficulty through a connectionist minimalist parser, *Cognitive neurodynamics,* Vol.3(4), pp.295-316.

Guerssel M. (1983), *An Outline of the structure of Berber*, Unpublished MIT Paper.

Iazzi E. (1991), *Morphologie du verbe en tamazight- parler des Aït Attab, Haut-Atlas central. Approche prosodique*, D.E.S. Thesis, Rabat.

Jamari A. (1992), *Clitic phenomena in Arabic,* Ph.D. Thesis, SOAS, University of London.

Labdellaoui R. (1997), *binyatu l - ɛanaasiri ssarfiyyati fi lloɣati l?amaaziɣiyyati: Halatu ttataabuqi bayna lfiɛli wa lfaaɛili*, D.E.S. Thesis, Oujda.

Makhad H. (1996), *Tense and aspect in Berber*, D.E.S. Thesis, Rabat.

Omari N. (2001), *The syntax of negation in Tashelhit Berber,* Doctorat Dissertation, Rabat.

Ouhalla J. (1988), *The Syntax of head movement: A study of Berber*, Ph. D. Thesis, University College London.

Pollock J.-Y. (1989), Verb movement, UG, and the structure of IP, *Linguistic Inquiry* 20, pp.365-424.

Sadiqi F. (2004), *Grammaire du Berbère,* Afrique orient.

Soudi A., Bosch A , Neumann G. (eds)(2007), *Arabic computational morphology : knowledge-based and empirical methods,* Springer.

Weinberg A. (2001), A minimalist theory of human sentence processing, in Epstein S.D., Horstein N. (eds.), *Working minimalism,*Cambridge : MIT.

Lexical development of bilingual Moroccan children in the Netherlands: analysis of mother-child interactions

Mohammadi Laghzaoui (1) & Esmah Lahlah (2)

(1) Faculty of Humanities – Tilburg University
PO Box 90153- 5000 LE Tilburg - Netherlands
E-mail: mo.laghzaoui@gmail.com

(2) Faculty of Law – Tilburg University
PO Box 90153- 5000 LE Tilburg - Netherlands
E-mail: a.lahlah@uvt.nl

Abstract

The present paper examines the input to which Moroccan Berber children are exposed in terms of rich and diverse lexicon at home in the Netherlands. Previous research reveals that the degree of exposure to a rich and diverse vocabulary determines the vocabulary size of children, which is a crucial predictor of a successful school career (Huttenlocher *et al*, 2002; Weizman & Snow 2001; Brent and Siskind, 2000 and Hoff, 2006). The current study covers both the individual characteristics of the children as well as those of their caretakers in the home environment. Firstly, the maternal language input in relation to children's language development will be discussed. Secondly, children' output will be discussed. The results indicate a large variation both among mothers and children. This can be partially attributed to the families' socio-economic levels (SES) and the literacy practices at home.

Keywords

Lexical development, lexical richness, lexical density, lexical diversity, Berber, Academic language, Moroccans in the Netherlands

Introduction

A large amount of research on language acquisition and vocabulary development within monolingual and bilingual families has shown that vocabulary size and early language skills are strongly related to the language input children are exposed to in early childhood in the home environment (Huttenlocher, Haight, Bryk, Seltzer & Lyons, 1991; Weizman & Snow, 2001). Communication between young children and their caregivers at preschool age is crucial with respect to children's language development next to the social, emotional and cognitive support provided. Earlier studies suggest that the amount of parental speech clearly influences children's vocabulary growth. Smolak and Weinraub (1983) examined mothers' speech of two groups of children with large and small vocabularies. The results showed that mothers of children with a large vocabulary produced by far significantly more speech than mothers with less vocabulary at their disposal. Tomasello *et al* (1986) studied mother-child interactions in experimental laboratory play sessions and found that the number of mothers' utterances significantly correlate with the number of word types produced by their toddlers. Similarly, the studies of Ninio (1993) and Brent and Siskind (2000) show that about 60% of children's early vocabulary are the same words which their parents often use, especially single word utterances such as proper names, action verbs, objects and interjections. All these studies provide evidence that the exposure to a relatively large amount of language facilitates children's vocabulary development.

To be able to develop vocabulary and semantic knowledge, preschool children need to be exposed to rich and varied speech that helps them to understand what objects are called and how words are used in different contexts (Hart & Risley, 1995). Vocabulary growth and development is influenced by context. Children whose social experiences provide more communicative opportunities and richer input build their vocabulary at a faster rate than children with less communicative experiences and less rich input (Hoff, 2006).

Lexical density and lexical diversity

Lexical density

Lexical density refers to the ratio of lexical words used in a text or conversation to convey a message. According to Ure (1971), lexical density is the proportion of tokens that are content words as opposed to function words. This notion of lexical density was used to measure the 'literacy' in written and spoken language. Academic texts are characterized by more lexical density than texts occurring in informal and non-academic settings (Schleppegrell, 2001). Academic language is more dense than informal language and contains a larger amount of nouns. Due to lexical density, academic texts pack more information efficiently into each sentence, while interactional texts are less characterized by lexical *thickness*.

School-oriented texts are typically more dense than texts for informal communication. The latter shows more short utterances with a low proportion of nouns and less subordinate clauses and nominalizations.

To calculate lexical density, different methods have been applied. According to Halliday (1994), lexical density can be described as the number of content words per non-embedded clause in a text. An alternative is the number of content words as proportion of the total number of words in a text (Egging, 1994; Fang, 1997). Another way to calculate the lexical density is to divide the total number of content words by the number of function words (Laufer & Nation, 1995). Since we are dealing in this study with spoken language which is segmented in single utterances, lexical density is measured by dividing the total number of content words by the number of utterances (cf. Halliday, 1994; Schleppegrell, 2001; Schleppegrell, 2004).

Lexical diversity

The other feature considered is lexical diversity. Mother-child interactions are characterised by the use of frequent vocabulary in daily contacts, while at school children are expected to use low-frequency vocabulary (Schleppegrell, 2004). In the school environment children are expected to use/and are exposed to a more sophisticated, technical and less frequent vocabulary. For school tasks, children are encouraged to use different words to talk about the same objects and events. This helps to make their vocabulary more diverse. Lexical diversity indicates the variation of the vocabulary in a written or spoken text. It has been traditionally measured by adopting the type-token ration (TTR), that is the number of different words in a sample of speech or written text divided by the total number of tokens. However, due to TTR's sensitivity to sample size, new methods have been suggested such as Guiraud's index, also known as RTTR. This index is calculated by the number of types divided by the square root of the number of tokens (Guiraud, 1960) and is independent of the text length. Recently, an even more advanced method has been introduced that takes into account the limitations of the older measures. This new algebraic method, which is used in the current study, is called the index D. It has been developed by Malvern, Richards and McKee (Malvern & Richards, 2002; McKee *et al*., 2000) and is calculated by using the computer program *vocd*, which is part of the CLAN system (MacWhinney, 2000).

Objectives and research questions

This study covers both the individual characteristics of the children as well as those of their mothers in the home environment. Our assumption is that the use of a high quantity of speech at home towards children about different subjects provides and

stimulates the kind of language expected at school. Against this background, the first aim of this paper is to investigate the quantity and quality of the lexical input of the caretakers (mothers). This will be done by trying to answer the following research questions:
- What are the characteristics of mothers' input in terms of lexical density and diversity and other vocabulary features?
- What are the characteristics of the children's output in terms of lexical density and diversity and other vocabulary features?

Methodology

Participants

The present paper focuses on a group of 12 Moroccan-Berber children and their mothers. These families participated in the DASH-project[1], a longitudinal study that comprised three measurement times. All children were born in the Netherlands and all the mothers were native speakers of Tarifit-Berber, a regional and only oral variety spoken in the North of Morocco. Most mothers and children are bilingual and speak both Berber and Dutch at home. Mothers reported to be the main caretakers of their children. The families were visited by a Moroccan Berber female research assistant who speaks both Dutch and Tarifit Berber. Data were collected during three measurement times: Time 1 when the children were 3;02 years old, Time 2 when children were 4;02 years old and Time 3 when children were 5;10 years old.

Data collection and analysis

Two tasks of different genres were designed to obtain the data from the mother-child conversations. In every measurement round, two interaction tasks were used: book reading task and picture description task. During the book reading sessions, mothers were asked to read to their children a book that the research assistant brought with her. The mothers were instructed to read the books as spontaneous as they normally do. In the first measurement time, the book was "*Tijn op de fiets*" (Tijn on the bicycle) (Oud & Sluyzer, 1996). In the second measurement time, the book "*Tim op de tegels*" (Tim on the tiles) (de Boer & Veldkamp, 2005) was used.

---

[1] This study is part of the DASH project (Development of Academic language in School and at Home), an in-depth study that deals with the development of academic language of Moroccan-Berber children in communicative contexts at home and in school (The project is financed by Netherlands Organization for Scientific Research; reference number 411-03-063).

The last book "*Emma in het spookhuis*" (Emma in the haunted house) (Waechter, 2005) was used in the third measurement time. For the picture description tasks, three different pictures, taken from Leiber (1999), were used in the three home visits. In the first measurement time, the picture "*op straat*" (on the street) was used. It is a general view of a busy street where a range of activities takes place. The picture used for the second measurement time is *'Circus Pepo'*. It is an arena with various scenes, performers and visitors. Finally, "*het klaslokaal*" (the classroom) was used in the third measurement round. This picture depicts children in different settings in the classroom.

A total of 36-videotaped sessions of mother-child conversations have been transcribed using *Child Language Data Exchange System* programme (CHILDES), following the CHAT conventions (MacWhinney, 2000). While transcribing, and due to Berber's rich morphology, consistent word segmentation has been developed and adopted by distinguishing the stem from prefixes and suffixes. Subsequently, a coding system was developed by the research group to code data according to the target language features of the in-depth studies (DASH, 2006).

Results

Lexical richness in mothers' input

Maternal lexical input in the current paper is approached by considering two dimensions: the quantity (number of utterances, tokens, content words and function words) and the quality (density and diversity). Table 1 presents the means and standard deviations of different input variables of the mothers. These descriptive statistics show maternal lexical input during two tasks at home.

N = 12		**Time 1 (Age 3;2 years)[13]**		**Time 2 (Age 4;2 years)**		**Time 3 (Age 5;10 years)**	
		M	SD	M	SD	M	SD
Nr. of utterances	Picture task	88.1	25.5	107.9	33.5	75.7	22.6
	Book task	105.4	39.8	79.3	19.5	62.4	23.2
Nr. of tokens	Picture task	669.3	292.0	838.2	329.4	458.8	272.5
	Book task	1180.0	511.9	945.3	321.9	716.0	342.0
Nr. of word types	Picture task	335.6	153.4	361.8	143.3	277.7	109.4
	Book task	521.2	218.0	443.2	128.6	341.2	165.9
Nr. of content words	Picture task	116.9	51.2	162.7	65.7	96.5	51.3
	Book task	275.4	130.0	232.1	84.6	175.8	88.3
Nr. of function words	Picture task	136.4	66.7	227.0	104.5	185.7	61.2
	Book task	288.7	129.2	298.7	124.4	184.6	89.0

Table 1: mothers' lexical input in two interactions tasks during three measurement times

At the level of utterances, there is a slight variation over the two tasks. At measurement time 1, mothers produced more utterances during the book task than during the picture task. At time 2 and 3, however, the total number of mothers' utterances during the picture task was higher than in performing the book task. However, the ANOVAs showed no significant task effect ($F(1,11) =1.77$, $p=.21$, $\eta_p^2=.14$). With regard to the book reading task, ANOVA with repeated measures indicates that there is a significant change of the number of utterances produced by mothers across the three measurement times ($F(2,22)=8.39$, $p=.00$, $\eta_p^2=.43$). The main effect of time for the picture task was ($F(2,22)=4.45$, $p=.02$, $\eta_p^2=.29$).

The total number of word tokens was higher in the book task than in the picture task during all three measurement times. It is noteworthy that in the last measurement time (children's age is 5,10 years) mothers produced the lowest number of utterances and word tokens compared to time 1 and 2. The same pattern can be seen at the level of word types. This might be due to the fact that children start talking more at this age and consequently take more time and turns than their mothers. Mothers used more word types during the book task conversations than the picture task. Again at time 3, and in both tasks, mothers produced less word types than in time 1 and 2. Regarding content words, over the three measurement times, mothers used more lexical words during the book reading task than during the picture description task. Again, in the last measurement round, mothers used the lowest number of content words compared with round 1 and 2 when children were younger than 5 years old.

After considering the quantity measures, we now go into the quality variables, density and diversity. These two measures are adopted as academic language components. Table 2 gives the mothers' mean values and standard deviations for both density and diversity during two different tasks at three measurement times.

		Time 1 (Age 3;2 years)		**Time 2 (Age 4;2 years)**		**Time 3 (Age 5;10 years)**	
		M	SD	M	SD	M	SD
Density	Book	2.54	0.48	2.97	0.99	2.74	0.74
	Picture	1.29	0.35	1.49	0.40	1.24	0.45
Diversity	Book	85.92	23.09	78.50	27.46	79.50	18.22
	Picture	59.45	23.61	61.28	30.48	57.49	29.53

Table 2: means and standard deviations of lexical richness measures of mothers at three measurement times

With regard to density and in the analysis for both book reading task and picture task, a significant main effect was found for the factor task ($F(1,11)=101,57$, $p=.00$, $\eta_p^2=.90$). In all measurement times, mothers' lexical density during book reading task was, on average, approximately twice as higher as during the picture task. Longitudinally, in both tasks mothers do not show much variation across the three measurement times. At the level of book reading session, ANOVA with repeated

measures confirms that there is no significant difference between the three measurement times ($F(2,22)=1.02$, $p=.38$, $\eta_p^2=.085$). The results of ANOVA also revealed that the main effect of time during the picture task was not significant ($F(2,22)=1.09$, $p=.35$, $\eta_p^2=.09$). Note that in both tasks, mothers lexical density increased from Time 1 to Time 2 and decreased again from Time 2 to Time 3. Both home tasks in measurement time 2 yielded the highest values of lexical density.

As to the lexical diversity, it is not divergent from the general pattern up till now. Mothers' language was more diverse when involved in book reading activity than in picture description activity. The comparison of the two tasks, using ANOVA with repeated measures, revealed a clear significant main effect of task ($F(1,11)=13.62$, $p=.004$, $\eta_p^2=.55$). Mothers' lexical diversity was also analyzed in a longitudinal perspective. Concerning the book reading task, results of a repeated measures ANOVA across the three measurement times show no significant change ($F(2,22)=.49$, $p=.62$, $\eta_p^2=.04$). The main effect of time was also found not significant during the picture description task ($F(2,22)=.10$, $p=.90$, $\eta_p^2=.009$).

Lexical richness in children' output

Having dealt with the mothers' input, we turn now to deal with the children's language use and their lexical output in particular. The means and standard deviations are set out in Table 3.

		Time 1 (Age 3;2 years)		**Time 2 (Age 4;2 years)**		**Time 3 (Age 5;10 years)**	
		M	SD	M	SD	M	SD
Nr. of utterances	Book task	39.1	20.4	32.6	14.2	25.1	16.7
	Picture task	52	22.8	57.4	21	58.5	14.7
Nr. of tokens	Book task	171.8	103.7	224.3	122.7	189.5	146.7
	Picture task	228.9	130.2	317.8	174.7	331	216
Word types	Book task	90.3	51.7	110.4	59.2	86.4	65.9
	Picture task	103.6	59.3	137.7	75.8	214.2	100.1
Content words	Book task	46.8	25.7	56.3	28.1	42.3	38.2
	Picture task	47.4	26.4	66.3	36.6	86.8	40.7
Function words	Book task	33.1	27.9	74.1	55.2	38.5	33.3
	Picture task	37.4	26.8	67.8	40.5	116.8	61.3

Table 3: Descriptive statistics for children's lexical measures in two interaction tasks during three measurement times at home

On utterance level, an important finding is that children do not show a stable increase across the three measurement times during the book reading task. On average, children produced the highest number of utterances in Time 1, followed by Time 2 and Time 3. At the level of word tokens, however, children used most tokens in Time 2, followed by Time 3 and 1. Concerning word types and content words, measurement time 2 witnessed the highest scores, followed by Time 1 and 3. Regarding picture description task, all variables showed an increase

corresponding with increasing age. The total number of utterances did not increase over the three measurement times. However, the scores of the other variables (word tokens, word types, content words and function words) were considerably higher as the age of children increased. Finally, it should be noted that, in contrast with the mothers, children's scores on different variables were higher during the picture description task than in the book reading task. This can be attributed to the nature of both tasks. In the book reading, it was usually the mothers who took the initiative to discuss the story of the book, while the picture description task was more approachable for children to produce more talk.
In order to assess children' lexical richness in the home setting, quality variables will be treated. In Table 4, the mean scores and standard deviations for lexical density and lexical diversity are presented.

N = 12		**Time 1 (Age 3;2 years)**		**Time 2 (Age 4;2 years)**		**Time 3 (Age 5;10 years)**	
		M	SD	M	SD	M	SD
Density	Book	1.2	.4	1.7	.4	1.5	.6
	Picture	.9	.2	1.1	.4	1.5	.6
Diversity	Book	51.4	21.5	60.3	26.1	77.3	35.5
	Picture	59.2	18.1	48.0	16.5	56.2	38.9

Table 4: Lexical richness of children at home during three measurement times

With regard to lexical density, children's language in Time 1 and Time 2 during book reading task was more dense than in picture task. In Time 3, the lexical density values in both settings were similar. Generally, children's talk during book reading activities generated more dense vocabulary than during picture description tasks. ANOVA analysis indicated a significant main effect of task ($F(1,11)=14.01$, $p=.003$, $\eta_p^2=.56$). Children's lexical density was also analyzed in longitudinal perspective. During the book reading task, ANOVA with repeated measures shows that the main effect of time was found to be significant ($F(2,22)=3.91$, $p=.035$, $\eta_p^2=.26$). In particular, lexical density increased significantly from Time 1 to Time 2. In the picture task, also a clear change can be seen at the level of lexical density which increased over time. ANOVA with repeated measures revealed a significant main effect of time ($F(2,22)=6.93$, $p=.01$, $\eta_p^2=.38$).

As far as lexical diversity was concerned, the ANOVAs showed that there was no significant task effect ($F(1,11)=.035$, $p=.88$, $\eta_p^2=.03$). That is, no significant difference can been seen in children's language diversity across the two home tasks, book reading and picture task. During book reading, when we compare the scores of children's lexical diversity, we can see a considerable progress across the three measurement times. The score values of lexical diversity were 51.4, 60.3 and 77.3 respectively in Time 1, Time 2 and Time 3. However, the results of ANOVA with repeated measures revealed that this difference was not significant

(F(2,10)=.70, p=.51, η_p^2=.12). A consideration of lexical diversity during the picture task shows a different trend. In Time 1, children used the most diverse language (mean=59.2), which was followed by a decrease in Time 2 (mean=48.0). In Time 3, a slight increase was established (mean=56.2). Results of ANOVA with repeated measures indicate that no significant difference could be found across time. The main effect of time was (F(2,22)=.18, p=.73, η_p^2=.03).

Conclusion

The main objective of this paper was to investigate the quantitative and the qualitative aspects of mothers' lexical input in addition to children's output. With regard to the mothers, large individual differences have bee found in both interaction tasks. Mothers show differences regarding the two settings (book reading vs. picture task) in all measurement times. It should be noted that the mothers did not show significant differences across the measurement waves. It turned out that mothers use more lexical input variables in the book reading task than in the picture task. Also differences have been found with respect to the qualitative measures (lexical density and lexical diversity). Mothers tend to use more dense and more diverse vocabulary during the book reading task than during the picture description task.

As far as the children are concerned, no stable increase could be established at the level of lexical variables during the book reading task. However, children showed an increase during the picture task. Concerning children's lexical density and diversity, the book task generated richer vocabulary than the picture task.

Acknowledgments

I gratefully acknowledge the Netherlands Organization for Scientific Research (NWO) for its support (grant number 411-03-063). I am also grateful to the children and their families who participated in this study as well as the assistants who collected the data of this project.

References

Bialystok, E., (2001), *Bilingual Development: Language, Literacy and Cognition*, Cambridge: Cambridge University Press.

Boer, K. de, & Veldkamp, T (2005),*"Tim op de tegels" [Tim on the tiles]*. Houten: Van Goor.

Brent, M., & Siskind, J. M. (2000), The role of exposure to isolated words in early vocabulary development. *Cognition 81*, 33-44.
Cummins, J. (1991), *Conversational and academic language proficiency in bilingual contexts*, *AILA Review,8,75,* 89.
DASH. (2006), *Coding protocol for functional linguistic analysis*, Amsterdam/Tilburg/Utrecht: DASH Research group.
Egging, S. (1994), An Introduction to Systemic Functional Linguistics. London: Pinter Publishers.
Extra, G. & T. Vallen (1997), The sociolinguistic status of immigrant minority groups in the Netherlands, *Sociolinguistica*, 11, 204-214.
Fang, Z. (1997), A study of changes and development in children's written discourse potential. *Linguistics and Education. 9*(1), 341-367.
Field, A. (2009), *Discovering statistics using SPSS for Windows*. London: Sage.
Guiraud, P. (1960), Problèmes et méthodes de la statistique linguistique. Dordrecht: D. Reidel.
Halliday, M. A. K. (1994), *An introduction to functional grammar.* London: Edward Arnold.
Hart, B., & Risley, T.R. (1995), *Meaningful differences in the everyday experiences of young*
American children. Baltimore: Paul H. Brookes Publishing.
Hoff, E. & L. Naigles (2002), How children use input to acquire a lexicon, *Child Development*, 73, 2, 418-433
Hoff, E. (2006), How social contexts support and shape language development. *Developmental Review, 26*, 55-88.
Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991), Early vocabulary growth: relation to language input and gender. *Developmental Psychology, 27*(2), 236-248.
Huttenlocher, J., *et al*. (2002), Language input and child syntax, *Cognitive Psychology, 45*, 337-374.
Laghzaoui, M., & E-Rramdani, Y. (2006), Mother-Child conversations in the first and the second language: A functional approach to school language acquisition at home. In L.R. Miyares, A.M. Alvarado, & C.Ã. Moreno (Eds.), *Proceedings of the 10th International Symposium on Social Communication* (pp. 867-872). Santiago de Cuba: Cuba: Centro de Linguistica Applicada.
Laghzaoui, M. (2008), On discourse conventions in academic and interactional Tarifit-Berber texts. In: A. El Aissati (ed.), *The Amazigh Language at Home and at School: Perspectives on Oral Discourse Structure and Academic Language Skills.* *Berber Studies* Volume 21, Koln: Rudiger Koppe Verlag.
Laghzaoui, M. (2008), Développement de la langue académique chez des enfants marocains amazighes aux Pays-Bas. In J.J. de Ruiter (ed.). *Langues et cultures en*

contact: Le cas des langues et cultures arabe et turque en France et aux Pays-Bas. (pp.123-147) Paris: l'Harmattan, Espaces discursifs.

Laghzaoui, M. (2009), Features of academic language in mother-child interactions: lexical development of Moroccan Berber children in the Netherlands. *Proceedings of the 8th International Language, Literature and Stylistics Symposium*. Turkey-Izmir: Izmir University, Faculty of Arts and Sciences.

Leiber, L. L. (1999), *Wat gebeurt er in de stad? [What happens in the city?]*. Oosterhout: Aartselaar.

MacWhinney, B. (2000), The CHILDES Project: Tools for analyzing talk. Mahwah, NJ: Lawrence Erlbaum Associates.

Malvern, D. D., & Richards, B. J. (2002), Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language testing, 19*(1), 85-104.

McKee, G., Malvern, D., & Richards, B. J., (2000), Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, *15*, 323-338.

Ninio A. (1993), On the fringes of the system: Children's acquisition of the syntactically isolated forms at the onset of speech. *First Language 13*, 291-314.

Oud, P., & Sluyzer, B. (1996), *Tijn op de fiets. [Tijn on the bicycle]*. Amserdam: Kimio.

Schleppegrell, M. (2001), Linguistic features of the language of schooling. In *Linguistics and education*, 12(4), 431-459.

Schleppegrell, M. (2004), The language of schooling: a functional linguistic perspective. London: Lawrence Erlbaum Associates.

Smolak, L. & Weinraub, M. (1983), Maternal speech: Strategy or response? *Journal of Child*

Language, 10, 369-380.

Tomasello, M., Mannle, S. & Kruger, A. (1986), Linguistic environment of 1- to 2-year-old

twins. *Developmental Psychology*, 22, 169-176.

Tomasello, M. (2000), First steps in a usage based theory of first language acquisition, *Cognitive Linguistics*, 11*,* 61-82.

Tomasello, M. (2003), *Constructing a Language: A Usage-Based Theory of Language Acquisition*, Harvard University Press.

Ure. J. (1971), Lexical density and register differentiation. In: Perren, G.E. & Trim, I. L.M. (eds.). *Applications of linguistics: selected papers of the Second International Congress of Applied Linguistics, Cambridge 1969* (pp. 443–452), Cambridge: Cambridge University

Press.
Waechter, F. (2007), *Emma in het spookhuis. [Emma in a ghost train]*. Amsterdam: Em. Querido's Uitgeverij BV.
Weizman, Z.O., & Snow, C.E. (2001), Lexical input as related to children's vocabulary acquisition: effects of sophisticated exposure and support for meaning. *Developmental Psychology*, *37*(2), 265-279.

Technologies de Recherche Sémantiques Appliquées au Tourisme: Cas de la Culture Amazighe

S.Mouhim[1], A.El aoufi[2], M.Eddahibi[3], C.Cherkaoui[4], El.Megder[5], D.Mammass[6]

1,2,3,4,5 Laboratoire IRF-SIC, Faculté des sciences & ENCG
B.P.28/S – Agadir – Maroc.

mouhimsanaa@yahoo.fr, adil@univ-ibnzohr.ac.ma, eddahibi@yahoo.fr, ccherkaoui@yahoo.fr, megderel@yahoo.fr, driss_mammass@yahoo.fr

Résumé – Abstract

L'objectif de ce travail est d'étudier les capacités de quelques moteurs de recherche sémantiques à répondre à certains critères sémantiques. Les critères choisis (*variations morphologiques, synonymie, généralisations et concepts, langages naturel, etc*.), constituent le noyau principal de nos interrogations. Nous n'essayerons en aucun cas d'effectuer une étude comparative de ces moteurs. Nous tenterons cependant de montrer les possibilités existantes en matière de recherche sémantique, de clarifier les approches utilisées et d'examiner quelques potentialités Nous déterminerons également quelques limitations que nous capterons à travers l'usage des ontologies qui s'avère important et qui est censé améliorer les résultats de la recherche. Les requêtes posées aux moteurs rentrent dans le cadre du secteur touristique, en particulier pour ce qui concerne la *découverte de la culture amazigh*.

The goal of this work is to study and discuss the capabilities of some semantic search engines to respond to some semantic criteria like: the morphological variations, synonyms, generalizations and concepts, natural language, etc. In this paper we do not try to make a comparative study of these engines, however we will show what is possible in terms of semantic search, clarify the approaches used and discuss some potential. We will also present some limitations that we will capture through the use of ontologies which is supposed to improve search results. The used queries as examples to test the semantic engines, come from the tourism sector, particularly as regards to the discovery of the Amazigh culture.

Mots Clés - Keywords

Critères sémantiques, Web sémantique, recherche sémantique, moteurs de recherche sémantiques, ontologies.

Semantic criteria, semantic web, semantic search, semantic search engines, ontologies.

1. Introduction

Les moteurs de recherche comme outils d'accès à l'information, sont tout à la fois un sujet *d'émerveillement* et de *frustration* (Mondeca, 2007). Les raisons de la frustration sont généralement multiples et peuvent être liés par exemple, au fait que l'information recherchée n'est pas en ligne et nous ne sommes pas certain de la trouver. D'autre part, nous ne savons pas comment décrire notre problématique de recherche, tout simplement par ce que le moteur utilise un jargon inconnu tels que les *opérateurs logiques*, ou alors le moteur nous inonde par la quantité et l'imprécision des documents présentés (Média, 2009). L'émerveillement peut parfois être due au fait que certains moteurs de recherche trouvent rapidement des réponses à nos requêtes alors que ces informations ne sont pas forcément stockées dans des bases de données.

De nos jours, l'usager est d'autant plus émerveillé de pouvoir directement récupérer une réponse précise à une question posée (Cousin, 2008). D'autres facteurs d'émerveillement proviennent des nouvelles potentialités fournies par les *moteurs de recherche sémantiques*, à savoir : les possibilités de poser des requêtes complètement en langage naturel, de réaliser automatiquement une correspondance entre les expressions de l'utilisateur et des expressions équivalentes, de réaliser des extensions sur les synonymes, acronymes, codes, références, etc.

Le travail que nous présentons dans cet article s'intéresse plus particulièrement à la recherche sémantique dans le cadre du *secteur touristique*. Il s'agit d'un Web Sémantique particulier dans lequel un touriste ou un acteur touristique (agence de voyage, tour opérateur, etc.) sont généralement à la recherche de ressources pour planifier un voyage, découvrir de nouveaux sites, apprendre des choses sur la culture d'un pays ou d'une région, etc. Sans une bonne connaissance des outils de recherche, l'usager et plus particulièrement le touriste émerge dans un vaste espace de ressources qui est loin d'être pertinent pour répondre à des besoins spécifiques. De plus, les clients sont plus exigeant que dans le passé (Ielpa, 2009). La requête classique comme *"Je préfère la montagne"* pourrait être enrichi par d'autres contraintes telles que *"Je souhaite découvrir la culture amazigh de montagne »* ou « *J'aimerais trouver un hôtel proche des montagnes pour me permettre de comprendre la culture amazigh*". Plus que cela, certains voyageurs exigent la planification de leurs vacances sans passer par les tour-opérateurs ou les agences de voyage (Mouhim, 2011). D'autres souhaiteraient interroger les moteurs avec des

expressions en langue naturelle ou encore de rechercher des informations en utilisant des « entités » ou des « concepts ».

La suite de cette communication est organisée en quatre parties. La partie 2 présentera le contexte et le cadre de notre étude. La partie 3 traitera quelques aspects de la recherche sémantique en présentant les critères retenus ainsi qu'une interrogation des quelques moteurs sémantiques choisis. La partie 4 montrera l'ontologie que nous avons développé pour améliorer les capacités de recherche de ses systèmes en fonction dans le cadre restreint du secteur touristique. Enfin, une conclusion viendra faire le point sur les résultats obtenus, en indiquant quelques perspectives à ce travail.

2. Contexte de l'étude

Sans vouloir rentrer dans les détails de toutes les approches sémantiques et des technologies sous-jacentes aux moteurs de recherche, ce papier se donne comme objectif primordial d'étudier la question de la recherche sémantique en rapport avec le tourisme, le tout appliqué à la découverte de la culture amazigh.

Nous essayerons d'interroger et d'examiner quelques moteurs de recherche sémantiques, parmi les plus connus et les plus documentés dans la littérature. Le but n'est pas d'effectuer une étude comparative de ces moteurs, mais d'essayer de montrer les possibilités existantes en matière de recherche sémantique, de clarifier les approches utilisées et d'examiner quelques potentialités précisées plus hauts (variations morphologiques, synonymie, généralisations et concepts, langages naturel, etc.). Nous essayerons également de déterminer quelques limitations liées au manque de relations sémantiques entre les concepts. Dans ce contexte, l'usage des ontologies s'avère important et permet d'améliorer les résultats de la recherche. Nous proposerons dans ce cas une ontologie qui tient compte du patrimoine amazighe pour une éventuelle amélioration du rendement des moteurs sémantiques.

Nous sommes convaincus que notre démarche ne considère pas toutes les approches et ne traite pas tous les outils existants, mais nous sommes cependant persuadés de capturer l'essentiel des caractéristiques communes et principales approches utilisées. Comme les recherches dans ce domaine sont très récentes contrairement aux moteurs traditionnels tels que Google, Yahoo, etc. (Dong, 2008), notre choix a surtout porté sur quatre moteurs de recherche sémantiques parmi les plus référencés dans la littérature et les plus testés.

3. La recherche sémantique

La recherche sémantique, comme application immédiate du web sémantique (Berners-Lee, 1994), a montré un potentiel important concernant le fonctionnement et l'amélioration de la performance de la recherche d'information. Par rapport aux

moteurs de recherche traditionnels qui se basent sur la fréquence d'apparition mot, les moteurs de recherche sémantiques sont plus susceptibles d'essayer de comprendre le sens caché dans les requêtes des utilisateurs et les documents récupérés. Les moteurs de recherche sémantiques utilisent des extensions sémantiques variées provenant du TALN[2].

Afin d'illustrer cela, nous proposons cinq caractéristiques principales qui définissent plus ou moins la recherche sémantique. Selon (Whatis, 2010), ces caractéristiques peuvent être résumées ainsi :

a- Les variations morphologiques. Un moteur de recherche sémantique est censé manipuler toutes les variations morphologiques (le temps, le pluriel, etc.) sur une base consistante. En d'autres mots, les résultats ne doivent pas changer si nous saisissons *« apprendre, apprentissage pour les temps, ou amazigh, amazighs, pour le pluriel etc. »*, ou alors en anglais « amazigh, tamazight, learn, learning, learned, etc. ». L'exemple de la question « apprendre amazigh » doit illustrer la qualité morphologique d'un moteur sémantique.

b- La synonymie. Un moteur de recherche sémantique doit être capable de manipuler les synonymes (amazigh, tamazight, berbère, kabil, etc.), dans le bon contexte et avec le bon sens des mots.

c-La généralisation. Un moteur de recherche manipule les généralisations (**bijoux** comme concept général, bracelet, gourmette, collier, *khalkhal,* etc.), quand la requête utilisateur est exprimée sous une forme générale et que les résultats attendus sont spécifiques. L'exemple de la question « Quel sont les bijoux des berbères ? » doit normalement donner lieu à une liste exhaustive de réponses, à savoir : bracelet, gourmette, etc.

d- Extension sur des concepts reliés – suggestions de recherche. Un moteur de recherche sémantique doit fournir des découpages, segmentations, relations faites entre les sujets. Ce niveau est généralement issu d'une ontologie. Il est composé des liens sémantiques entre des concepts (travaille_avec, est_relié_à, interagit_avec, etc.). Ce type d'extension sémantique peut être utilisé pour suggérer une extension automatique des recherches en utilisant de manière pertinente les liens sémantiques.

e-Le langage Naturel et les questions. Un moteur de recherche sémantique est prévu pour répondre judicieusement lorsqu'une requête est posée sous forme de question (quoi, où, comment, pourquoi, etc.). La tâche principale d'un moteur de recherche est de classer les résultats de recherche de la manière la plus logique de telle sorte que la réponse à la question constitue une seule entité. Le moteur sémantique pouvoir répondre directement et ne pas fournir une liste de pages, mais surtout une réponse précise à la question posée. Le moteur De

[2] TALN : Traitement Automatique du Langage Naturel.

plus, l'utilisateur d'un moteur sémantique doit poser sa question en langage naturel.

Ces types d'extension répondent en grande partie aux besoins d'un touriste en tant qu'usager des moteurs de recherche. Dans la suite de cette communication, nous montrons à travers un certain nombre d'exemples comment ces extensions sont pris en compte par les moteurs sémantiques choisis. Nous notons que parmi ces systèmes, nous avons intégré le moteur Google pour argumenter notre discussion. Notons que Google a souvent été vue comme non sémantique.

3.1. Google et la sémantique

Google a souvent été vu comme un outil de recherche traditionnel qui interroge son moteur à l'aide du principe des mots-clés en utilisant le fameux algorithme du « PageRank » (Brin, 1998 ; Rogers, 2002). Cet algorithme se base sur le principe de vote pour une page. Ainsi, plus une page reçoit des votes (des liens rentrants), plus elle sera considérée comme pertinente pour Google. Mais l'importance considérée est sans lien direct avec l'intérêt et la pertinence de celle-ci. De même, le choix en fonction des mots-clés, limite la vision de Google, dans le sens ou les extensions sémantiques sont absentes.

Il est vrai que Google n'a pas souvent été associé à la recherche sémantique. Cela était surement vrai il y a quelques années, mais si l'on regarde de plus près, certains traitements sémantiques sont déjà gérés (Charton et al., 2009). Ainsi, contrairement aux idées reçus, Google s'est mis au Web sémantique en structurant des données non structurés. Il a également amélioré son algorithme de recherche et a proposé une description des résultats plus longue. Nous pouvons aussi ajouter à cela le fait qu'il élargit le champ sémantique et inclut les concepts liés au mot clé. Il s'est également intéressé aux réseaux sociaux tels que Twitter, pour la recherche en temps réel (Passant, 2009). Google semble aussi répondre aux questions des utilisateurs, au lieu d'utiliser uniquement des mots clés. Ainsi, il est vrai que, depuis peu, quelques réponses de Google curieuses, notamment sur les dates de naissances, sur d'autres faits, ou sur les liens familiaux, Google répond directement à la question au lieu de proposer une liste de page. Cela fonctionne pour l'instant plus en anglais.

Prenons l'exemple la question "`who is the king of Morocco?`" Cette question va nous renvoyer comme premier résultat (Fig.1) :

Figure 1 : Réponses pour la requête « who is the king of Morocco?".

D'autres questions tels que : "Is amazigh a language?" aura comme réponse :

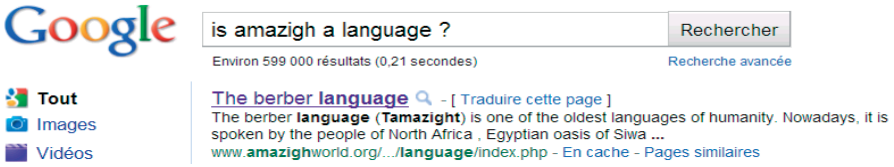


Figure 2 : Réponses pour la requête « Is amazigh a language?".

Nous notons dans ces réponses plusieurs éléments importants :

- C'est le tout premier résultat qui constitue la réponse à nos questions.
- Google cherche la réponse parfois dans le Wikipédia. On pourrait penser que Google extrait les deux informations importantes de la question « King of Morroco », mais notre question ne contient pas la réponse « Mohamed IV ». Pour la deuxième question, s'il extrait « amazigh » et « language », sa réponse contient le mot « Berber » qui est un synonyme de « amazigh ».

De même, si l'on poursuit nos interrogations, les deux questions *"Is amazigh arab?"* ou au pluriel "Is amazigh arabs?" ne change pas le résultat. On peut donc penser qu'il utilise ainsi les variations morphologiques. Il apparaît donc que les réponses récupérées, Google effectue des traitements sur les requêtes.

Nous pouvons noter cependant que la réponse à des requêtes complexes telles que : "Je souhaite découvrir la culture amazigh de montagne" - *donne des réponses parfois mitigées avec comme première réponse* « au cœur des *montagnes* et des paysages sahraouis *amazighs* ». Dans le texte de la ressource trouvée, un guide de montagne propose de visiter «*les montagnes de Matmata au sud de la Tunisie* ». Si l'on simplifie la requête en demandant " découvrir la culture berbère de montagne", nous parvenons à obtenir plus d'informations sur les sites les plus classés, nous arrivons à avoir plus loin dans les réponses « Location de mules, initiation à la langue berbère, et circuits de découverte, etc.», ce sont bien sûr des liens sponsorisés.

3.2. Les moteurs de recherche sémantiques

3.2.1. Hakia

Hakia est un moteur de recherche qui vise à fournir des résultats de recherche en se basant sur le sens du contenu plutôt que leur popularité de la page (ou PageRank) (Hakia, 2010 ; Shaikh**, 2010**). Durant la phase d'indexation, le moteur de recherche met l'accent sur *l'âge du contenu Web* ainsi que sur la *crédibilité* de la source.

Hakia utilise un algorithme de classement sémantique proposé par *OntlogicalSemantics* (Hakia, 2010), qui s'appuie fortement sur la sémantique ontologique et la linguistique computationnelle. Cet algorithme se base sur une base de concepts appelée OntoSem, il s'agit d'une base de données linguistique où les mots sont catégorisés selon leurs différentes significations. L'algorithme de classement sémantique permet la décomposition sémantique des phrases, qui s'apparente à une analyse morpho-syntaxique.

Pendant l'indexation à l'aide de l'algorithme QDEX (Query Indexing Technique) (Bröker, 2008), chaque page est analysée et l'algorithme extrait toutes les requêtes menant à cette page. Une autre fonctionnalité de Hakia provient du système de classement qui se base sur l'algorithme SemanticRank. Une autre nouveauté de Hakia concerne les Résultats de la recherche qui sont organisés en onglets: résultats Web, sites crédibles, images, vidéos et nouveautés.

Après cette caractérisation du fonctionnement et des outils utilisés par Hakia, essayons maintenant d'effectuer quelques tests en comparaison avec Google. Si nous reprenons les requêtes posées à Google, commençons par la question ″Who is the king of Morroco?″, le résultat est presque identique à celui de Google (Fig. 3), à savoir :

Figure 3 : Réponses pour la requête « Who is the King of Morocco?".

Le premier résultat est identique à celui d Google, mais les résultats qui suivent sont plus récentes et présentent la constitution marocaine en cours de changement, ainsi que le lien avec le Roi Mohamed IV. Les réponses qui suivent sont assez pertinentes puisqu'elles présentent la famille royale avec lien vers le roi Hassan II qui est le père, des liens avec les réformes actuelles, etc. Les liens présentés sont récents, utilisent des liens sémantiques (lien de parenté avec le roi Hassan II). De plus, le classement des ressources s'effectue en fonction des catégories : Web, Crédible, News, Blogs, etc. Le nombre de réponses n'est pas très grand et ne contient que cinq pages. Par contre, le temps mis pour répondre est assez grand comparé à Google. Notons également que la première réponse se base sur le Wikipédia.

Pour la question "Is amazigh a language?", le résultat est complètement différent de celui de Google. Hakia présente un résultat détaillant la terminologie de la langue amazigh, qui est un article de recherche. Le second résultat provient du Wikipédia et traite : l'amazigh comme étant un langage du peuple indigène de l'Afrique du Nord. En sixième position, il présente l'alphabet berbère ou « Tifinagh », alors que le mot amazigh n'existe pas dans la source, seul le mot « Tamazight » figure. Là, nous voyons bien que la synonymie est traitée. Dans la partie News, quelques liens sémantiques apparaissent en liant la langue amazighe avec la culture. Google se concentre plus sur les expressions « Amazigh language » ou « Berber Language » bien qu'il donne la réponse à la question en premier lieu.

Concernant la question : "Is amazigh arab?", la réponse est classée en premier lieu et elle est identique à celle de Google. La réponse à cette question tient sur une seule page, et les liens sémantiques. La question "Is amazigh arabs?" ne donne cependant pas le même résultat que la précédente. Pouvons nous conclure que l'analyse morphologique a moins d'importance que la sémantique ou la réponse à la question.

La requête "discover the amazigh culture", les réponses sont plus pertinentes. En premier lieu, un document intitulé "Where can I find information about Amazigh culture?" est présenté à partir de la source Yahoo! Answers. Concernant les requêtes complexes telles que "discover the amazigh culture in mountains" donne des résultats impressionnants e présentant un hôtel au sud du Maroc dans les montagnes de la région de Zagora, avec une incitation parfaite à découvrir la culture berbère. Les réponse qui suivent sont aussi pertinentes, car elles présentent : la culture Kabyle et les montagnes, la découverte de la musique, les attractions tels que le festival d'Imilchil.

3.2.2. Kngine

Kngine utilise une base de connaissances appelée 'Kngine live Objects' composée de plus de 1 milliard d'informations avec 7 millions de concepts (Kngine, 2010). Cette base de connaissances est utilisée pour déterminer : les synonymes, les relations entre concepts, le sens des concepts, la classification des documents, l'analyse basée sur le contexte et la recherche floue. Selon la même source qu'avant (Kngine, 2010), des travaux en cours portent sur : la technologie d'indexation Appelé «Snippet Recherche». Ce mécanisme consiste à proposer à

l'utilisateur des paragraphes riches montrant le contenu des ressources sans être dans l'obligation de les ouvrir.

Kngine propose des fonctionnalités intéressantes que nous avons explorés en fonction des requêtes proposées avant. Tout d'abord, nous pouvons citer la désambiguïsation. Par exemple, pour la requête "`Amazigh`", Kngine propose deux acceptions sous deux onglets différents : « Berber Poeople » et « Berber Language ». A partir de là, nous pouvons d'abord déduire qu'il manipule aussi la synonymie, en proposant le terme «Berber ». Il manipule aussi les variations morphologiques, en effet, le résultats est le même si l'on saisit « amazighs » ou « amazigh ». Kngine propose directement les réponses aux questions (Fig. 4), ainsi pour la question "`Who is the king of Morroco ?`", la réponse est placée en première ligne, avec éventuellement une photo, une carte du Maroc et des informations supplémentaires (nombre d'habitants du Maroc, langue officielle, etc.).

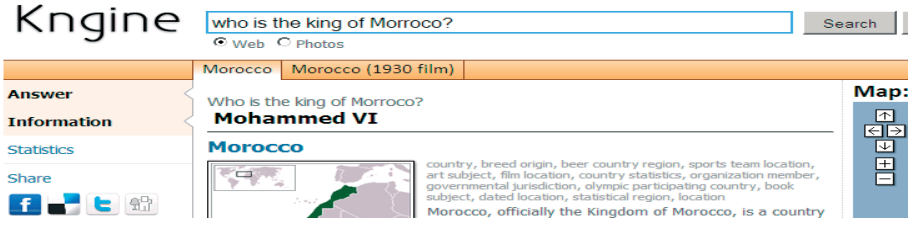


Figure 4 : Réponses pour la requête « Who is the King of Morocco?".

3.2.3. SenseBot

SenseBot présente une interface intéressante dans la mesure ou plusieurs moteurs peuvent être interrogés tels que : Google, Yahoo, Bing, senseBot (Fig. 5). De plusieurs, la recherche est proposée dans plusieurs langues (SenseBot, 2011).

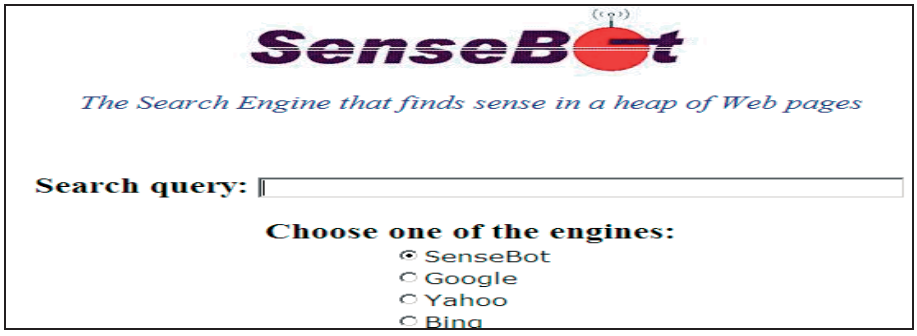


Figure 5 : Réponses pour la requête « Who is the King of Morocco?".

Pour le mot clé ″amazigh″, nous pouvons constater l'utilisation de la synonymie en proposant « berbère ». Par contre l'usage du pluriel change complètement la donne, et les résultats ne sont pas les mêmes. La réponse à la requête ″King of Morroco″ donne lieu à une réponse générique présentent la liste des rois du monde, sans présenter une réponse directe. En utilisant le Français, nous retrouvons la réponse à la requête en première position, bien sûr basée sur le Wikipédia. D'autres réponses qui suivent sont aussi intéressantes et actuelles (visite du Ministre de Tunisie au Maroc, etc.). D'autres requêtes tels que ″les amazighs sont des arabes″, donne lieu à des sites intéressants, en proposant également une liste de thèmes de recherche importants. Pour la requête ″connaître la culture amazigh″, un ensemble de liens important traitant des événements liés à la culture amazigh, tels que la musique, les festivals, etc. Dans les réponses, nous pouvons constater la référence aux forums. Nous pouvons cependant observer l'absence des traditions berbères.

3.3. Discussion

A travers nos interrogations, nous pouvons constater l'alignement de la plupart des moteurs présentés (y compris Google) par rapport aux critères sémantiques choisis. A part quelques imperfections, la plupart des critères sont vérifiés. Nous pouvons noter également quelques nouveautés importantes comme des interfaces d'interrogations qui sont plus ou moins ouvertes et sophistiquées (utilisant des liens vers Google, Yahoo, Bing, etc.), la catégorisation des réponses (Web, sites crédibles, images, vidéos et nouveautés, etc.), la présentation des informations récentes, l'usage du Wikipédia, etc. Toutes ces fonctionnalités combinées donneront certaines un nouvel élan à la recherche sémantique. De plus, comme nous le présenterons ci-dessous, la spécialisation à un domaine précis est probablement importante.

4. Vers une ontologie Touristique

Dans cette partie, nous présentons brièvement une ontologie touristique OTM (Ontologie du Tourisme Marocain) qui fait partie des travaux que nous menons dans le cadre du projet GECO-WES[3], Le contenu principal de cette ontologie provient d'une extraction manuelle de la connaissance à partir d'un certain nombre de sites Web officiels. Nous présentons ci-dessous les éléments qui nous ont permis de la concevoir et de la faire évoluer pour prendre en compte, par exemple des informations sur la culture amazighe.

[3] GECO-WES : Gestion de Connaissances et Web sémantique.

Conceptualisation

Pour élaborer un vocabulaire commun du tourisme, nous avons consulté diverses ressources sur le web. On peut citer en particulier le thesaurus de l'organisation mondiale du tourisme ainsi que le thésaurus de l'UNESCO (Unesco, 2008), qui sont des normes internationales. Nous avons aussi consulté divers portails web officiels tels que (ParisInfo, 2009), ce qui nous a permis de collecter un bon nombre de concepts ainsi que leur catégorisation. Par exemple, le concept "Hôtels et hébergement" comprend les sous-catégories: hôtels, résidences, meublés, campings, handicapés, centres de réservation. Pour le concept «Restaurants et cafés," nous avons trouvé les sous-concepts : la haute cuisine, les restaurants classiques, Fast_Food, cafétérias universitaires, traiteurs, etc.

D'autres plates-formes sociales ont été examinées. Par exemple les FAQs contiennent certaines questions, les plus posées par les touristes. Ces questions ont enrichi notre ontologie avec des relations sémantiques telles que (est proche_de, est_synonyme_de, est_une_sorte_de, est_analogue_à, etc.). D'autres concepts liés à la culture amazigh on été introduit au fur et à mesure pour capter les traditions berbères et les coutumes (cuisine, habit traditionnel, bijoux, langue, etc.) (Fig. 6, 7).

Nous avons également pris en considération les normes et règlements régis par la loi de ce secteur au Maroc. Après avoir recueilli les informations nécessaires, la prochaine étape consistait à la création des classes et des concepts pour former la structure de l'ontologie, ainsi que les propriétés, individus et annotations.

Les concepts de MTO

Dans cette étape de notre travail, l'ontologie que nous avons développé est composée d'un ensemble de 118 concepts et 75 propriétés. L'étude, l'analyse et la classification des concepts recueillis de différentes sources de données (portails, blogs, ontologies, etc.) nous a permis de les diviser en sept classes mère (Fig.6, Fig.7):

- **Activité** : cette classe contient toutes les activités que peuvent être pratiquées ;
- **Attraction** : classe qui définit tous les lieux que peut visiter un touriste au Maroc. Il nous a été difficile de collecter tous les concepts pouvant s'intégrer dans cette catégorie vue la richesse du patrimoine marocain et la diversifié des paysages marins, montagnards et désertiques ;
- **Etablissement touristique** : classe qui regroupe toutes les catégories offertes pour loger le touriste. La hiérarchisation de cette classe suit les normes régies par la loi marocaine ;
- **Transport** : classe qui regroupe tous les moyens de transport présents aux Maroc ;
- **Restaurant** : classe qui définit tous les services de restauration qui peuvent exister ;

- **Culture**: classe qui décrit les héritages culturels du Maroc.

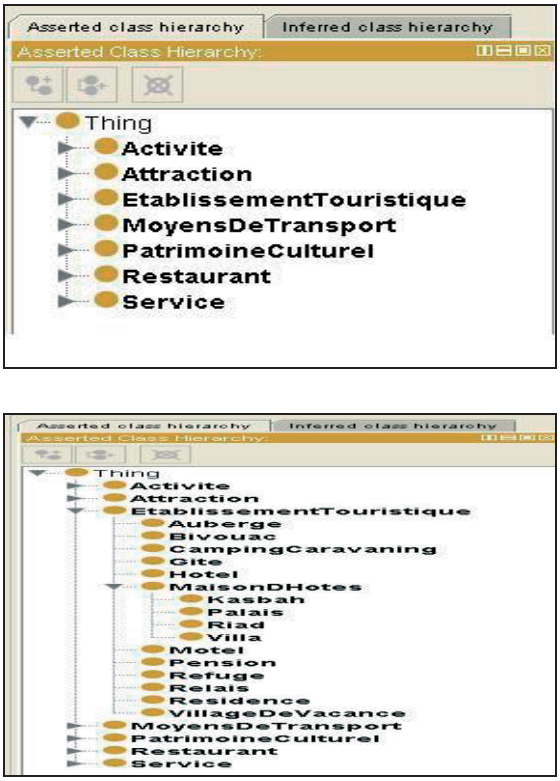


Figure 6 : Classes de l'ontologie OTM.

Figure 7 : Classe « Patrimoine Culturel » et Individus.

5. Conclusion

La recherche sémantique est une technologie qui commence juste à faire ses preuves. Ainsi, comme nous l'avons pu le montrer les moteurs sémantiques ne sont pas si nombreux que ça. Parmi ceux que nous avons sélectionnés, l'objectif était d'essayer de montrer les possibilités existantes en matière de recherche sémantique, de clarifier les approches utilisées et d'examiner quelques potentialités précisées plus hauts (*variations morphologiques, synonymie, généralisations et concepts, langages naturel, etc*.). Les exemples de requêtes que nous avons examinés étaient relatifs à un champ particulier qui est celui de la culture amazigh, en relation avec le tourisme. Il est vrai que certains moteurs répondent mieux que d'autres sur certains points particuliers. Nous n'avons pas vraiment pris position sur les réponses et surtout nous avons évité de dresser un tableau comparatif, bien que cela ait été fait implicitement.

Nous pouvons discuter longtemps et tester la manière qu'ont les moteurs de recherche sémantiques de nous répondre, mais les conclusions ne peuvent être que des suppositions, puisque les technologies et les algorithmes utilisés sont confidentiels. De plus, il est difficile, voire impossible de se faire une idée exacte sur un nouvel outil en testant seulement quelques requêtes.

Tout cela nous fait penser que ces systèmes ont des capacités complémentaires, et parfois sources d'émerveillement. Nous pensons que l'intégration des ontologies telles que l'OTM peut améliorer les recherches dans le tourisme ou plus particulièrement les recherches dans des champs liés à la culture amazighe.

La suite de ce travail se décline alors par le choix des moteurs sémantiques utilisant des ontologies intégrables, et dans ce contexte, nous pouvons tester notre ontologie, pour l'améliorer, voire pour construire un moteur sémantique spécifique au tourisme au Maroc.

Références

Brin S., Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine, retrieved en 2010, http://infolab.stanford.edu/~backrub/google.html

Bröcker, J., Van Ahee, G.J (2008). Semantics & Search Engine Optimisation, Consulté en 2010, http://www.yes2web.nl/files/ebooks/semantic-web-seo.pdf

Charton, E., Deveaud, R., Bonnefoy, L. (2009). *Interrogations de moteurs de recherche par des requêtes formulées en langage naturel*, Mémoire de recherche, 2009.

Cousin, C. (2008). *Tout sur le Web 2.0*, Editions Broché.

Dong, H., Hussain, F.K, Chang E., (2008). A Survey in semantic Search Technologies, In the IEEE International conference on digital EcoSystems and Technologies.

Engine, *ICIET2010*, Karachi.

Hakia, (2010). http://www.hakia.com, http://www.sirgroane.net/google-page-rank/.

Ielpa S. M., liritano S., Leone N., Ricca F., (2009). *An ASP-based System for e-tourism, in the e-book entitled Logic Programming Reasonning*, Springer Ed.

Kngine (2011). Semantic Serach engine, retrieved at October 2010, http://www.kngine.com/.

Media, (2009). Tirer le meilleur d'Internet - Recherche efficace en ligne, consulté en 2009 ; http://www.media-awareness.ca/ .

Mondeca, (2007). L'extension des capacités des moteurs de recherche, trouvé en 2010, http://mondeca.wordpress.com/.

Mouhim, S., El Aoufi, A., Cherkaoui, C., Megder, EL, and Mammass, D. (2011). An ontology-based knowledge Management System for Tourism, Proceeding of The Conference *SIIE'201, Marrakech.*

ParisInfo, (2009). Site officiel de Paris Info, consulté en 2009, http://www.parisinfo.com/.

Passant, A, (2009). *Technologies du Web Sémantique pour l'Entreprise 2.0*, Thèse d'Université Paris IV – Sorbone.

Rogers, I. (2002). The Google Pagerank Algorithm and How It Works, accedé en 2011;

SenseBot, (2010). The Search Engine that finds sense in a heap of Web pages, retrieved December 2010, http://www.sensebot.net/.

Shaikh, F., Siddiqui, U.A, Shahzadi, I. (2010). SWISE: Semantic Web based Intelligent Search

UNESCO, (2008). Site officiel du thésaurus de l'UNESCO, consulté en 2008, http://www.ulcc.ac.uk/unesco/.

Whatis, (2010). What is Semantic Search?, Retrieved Octobner 2011, http://company.hakia.com/new/whatis.html.

Étude contrastive des locutions en amazighe et en français en vue de la constitution d'une base de données lexicale[1]

Malika CHAKIRI
Paris-Descartes-Sorbonne/Laboratoire LARLANCO-Université Ibn Zohr-Agadir
malika.chakiri@parisdescartes.fr
chakirimalika@yahoo.fr

RESUME

Il arrive souvent que certaines possibilités de dire dans une langue ne se présentent pas dans une autre langue. Mais si cette langue ne connaît pas la même structure que la première, elle en possède très certainement d'autres qui la remplacent.
Partant de cette hypothèse, nous nous proposons de mener une étude contrastive, c'est-à-dire une réflexion méthodologique sur un ensemble de phénomènes liés aux contrastes entre l'amazighe marocain et le français afin d'établir des rapports entre elles. Nous centrons notre intervention sur le figement lexical et notamment les locutions verbales. Notre objectif est de chercher les points de ressemblance et de dissemblance entre l'image et le sens véhiculé par chaque locution dans ces deux langues. De ce fait, seront étudiées:

1. Des locutions dont le sens et l'image sont identiques (dans les deux langues)
2. Des locutions dont l'image est différente mais le sens est identique
3. Des locutions dont l'image est identique mais le sens est différent.

ABSTRACT

It often happens that some options of saying in a language do not appear in another language. But if this language does not know the same structure as the first, it possesses very certainly the others which replace it.
Taking in consideration this assumption, we suggest leading a contrastive study that is a methodological reflection on a set of phenomena linked to the contrasts between the Amazighe of Morocco and the French to establish relations between

[1] Un grand merci à Monsieur El Mehdi IAZZI qui m'a éclairée de ses lumières.

them. We focus our intervention on the domain of lexical solidifying in particular the verbal phrases. Our objective is to look for the points of similarity and dissimilarity between the image and the meaning conveyed by each phrase in these two languages. Therefore, will be studied:

1. Phrases whose meaning and image are identical (in both languages).
2. Phrases whose image is different but the meaning is identical.
3. Phrases whose image is identical but the meaning is different.

Mots clés*: figement, locution, étude contrastive, amazighe (berbère), français*

Key words*: solidifying, locution, contrastive analysis, amazighe (Berber), French language*

> « Les étrangers engendrent des séries incorrectes, d'abord parce qu'ils s'imaginent que les mots d'un groupe ont une existence indépendante et peuvent être remplacés par leurs synonymes : ainsi on dira *regagner sa liberté* au lieu de *recouvrer sa liberté*, parce qu'il n'y a pas de différence appréciable entre *regagne*r et *recouvrer,* ce qui n'empêche pas ici que *regagner* est une faute de français » (FONAGY 1994 : 73).

INTRODUCTION

Dans ce travail, nous proposons une analyse contrastive des locutions verbales entre l'amazighe (parler des Ayt Wirra – Maroc central) et le français. Nous entendons par locution verbale, toute suite polylexicale construite à partir de plusieurs unités non soudées, non autonome au niveau syntaxique et dont le noyau verbal forme avec l'un des actants (le nominale ou le pronom) un bloc figé inanalysable sémantiquement et s'opposant à la plupart des transformations syntaxiques.

Il est généralement admis que la compréhension et l'usage des unités lexicales figées, vu la particularité de leur statut qui est différent de la syntaxe ordinaire, nécessite un très bon maniement de la langue et non pas un apprentissage mécanique. D'où la difficulté de la traduction de ce type d'énoncés. Tous les traducteurs et les linguistes comparatistes l'affirment. Les synonymes sont rares d'une langue à l'autre car au-delà du signifiant, chaque unité lexicale véhicule une histoire, une culture…

Il sera question ici de relever les points de ressemblance et de dissemblance entre l'image et le sens véhiculé par chaque locution dans les deux langues. De ce fait seront étudiées :

1. des locutions dont le sens et l'image sont identiques (dans les deux langues)

2. des locutions dont l'image est différente mais le sens est identique

3. des locutions dont l'image est identique mais le sens est différent.

Cette étude est organisée comme suit : nous analysons, en premier lieu, les propriétés morphosyntaxiques de la locution verbale. Cette analyse sera ensuite suivie de l'étude comparative.

Pour la notation des données amazighes, nous utilisons le protocole suivant : - voyelles : *a*, *i*, *u* et *ə* pour noter le schwa. Semi-voyelle : *w*, *y*. Consonnes : *p*, *b*, *t*, *d*, *k*, *g*, *l*, *m*, *n*, *s*, *z*, *š*, *ž*, *ḥ*/*c* notent la fricatives pharyngales sourde et sonore, *x*/*ġ* les fricatives vélaires sourde et sonore, *h* la spirante, *q* l'occlusive dorso-uvulaire, *r* la vibrante apicale. Le point sous la lettre indique l'emphase, le *w* en exposant note la labiovélarisation, le trait sous la lettre note la spirantisation, le dédoublement de la consonne indique la gémination.

Par ailleurs, les signes et abréviations adoptés sont : A : aoriste, AI : aoriste intensif, EA : Etat d'annexion, loc. : locution, m.p. : monème du passif, N : nom, P : Prétérit, p.m.r : particule de mise en relief, PL : pluriel, PN : prétérit négatif, P.O. : particule d'orientation, Par. : participe, SV : syntagme verbal, V : verbe, enfin le symbole « * » renvoie au défigement ou à des séquences non attestées dans le parler étudié.

1. Propriétés morphosyntaxique des locutions verbales

Nombreux sont les chercheurs qui ont remarqué le rapport ''intime'' entre les constituants des expressions figées, faisant ainsi un barrage devant toute tentative de modifications. Mais les expressions figées ont-elles le même comportement au regard des manipulations admises par la syntaxe libre ? Telle est la question à laquelle nous essayons de répondre dans ce qui suit.

1.1. Opposition aspectuelle

L'amazighe est une langue à aspect. Elle distingue, de ce fait, l'aoriste intensif (l'aspect inaccompli) et le prétérit (l'aspect accompli). Il importe donc de tester ce qu'il en est du changement d'aspect dans les locutions. Considérons d'abord cette transformation dans un syntagme ordinaire :

Accompli	**Inaccompli**
i-swa bba il boire + P mon père « Mon père a bu »	*da i - ssa bba* réel il boire + AI mon père « Mon père boit »

Nous remarquons que le changement d'aspect dans un énoncé libre n'atteint pas le sens global du syntagme. Comme nous allons le voir, il en va autrement dans la locution :

Accompli	**Inaccompli**
i - raħə nn iḏarr n izzg̱ar il arriver + P p. or. pieds de bovins « Il est arrivé aux pieds des bovins » 1. « Sa situation est lamentable » 2. « Il a subi une grande chute »	**da- nn - i tt- raħ iḏar n izzg̱ar* réel p. or. il arriver + AI pieds de bovins « Il arrive aux pieds des bovins »

Inaccompli	**Accompli**
da - i -srurud ur i-rbi réel il faire des va et vient + AI nég. il avoir l'enfant sur le dos + PN « Il fait des va et vient sans avoir le bébé sur le dos » « Il est très inquiet »	**i - srurd ur i – rbi* il faire des va et vient + P nég. il avoir l'enfant sur dos + PN « Il a fait des va et vient sans avoir le bébé sur le dos »

Le changement d'aspect peut soit détruire le figement soit donner lieu à des séquences inacceptables. Dans les locutions étudiées, d'une manière générale, les verbes considérés comme imperfectifs (*srurd* « faire des va et vient » *sawəl* « parler », etc), sont davantage compatibles avec l'aoriste intensif (inaccompli), alors que les verbes perfectifs (*k̲šəm* « entrer », *ffeġ* « sortir », *kk* « passer par », *amə ẓ* « attraper », etc.) acceptent mieux le prétérit (l'accompli). C'est pour cette raison que la manipulation de l'aspect est réduite.

1. 2. Opposition singulier / pluriel

Ce passage disloque la locution et génère des énoncés non attestés dans la langue, comme étant le cas dans les exemples suivants :

t̲-tfar t̲ t̲mara elle suivre + P le misère (EA) « La misère l'a suivi » « Il est misérable »	**t̲far- ənt̲ t̲ t̲mariwin* suivre + P elles le misère (EA) + pl. « Les misères l'ont suivi »

Cependant, certaines locutions autorisent ce passage :

xwa- n as ***ifassən*** être vide + P ils à lui main + pl. « Ses mains sont vides » « Il ne possède plus rien »	*i- xwa- as* ***ufus*** il être vide + P à lui main (EA) « Sa mains est vide » « Il ne possède plus rien »

1.3. Réduction des expansions

Dans *swi-x assnatt aman* « j'ai bu l'eau, hier », l'effacement des segments *assnatt* « hier » ou/et *aman* « eau » n'affecte pas la validité syntaxique de l'énoncé. En revanche, aucun des éléments de *swi-x* « j'ai bu » ne peut être supprimé puisque l'indice de personne *x* « je » et la base verbale *swi*, en tant que syntagme verbal sont le prédicat de l'énoncé. Ce qui revient à dire que tout ce qui n'est pas prédicat est à considéré comme une expansion de ce prédicat.

Généralement, on distingue les expansions directes (sans marque fonctionnelle) et les expansions indirectes (à marque fonctionnelle). La première est liée aux verbes transitifs, la seconde au prédicat verbal par un indicateur de fonction. Par ailleurs, il est fréquent que le SV soit accompagné d'une expansion d'un type particulier dite complément explicatif représentant l'indice personnel qui est toujours associé au radical verbal. Placé en tête de l'énoncé, ce nominal assume la fonction d'un « indicateur de thème ».

Enfin, l'expansion peut-être liée à l'énoncé entier ou à l'un de ses constituants. La première est dite totale, la deuxième partielle. Ainsi, dans *ar iqqar ləkṯab s žžhəd assnatt* « il lisait le livre à haute voix, hier », *s žžhəd* « à haute voix » est une expansion de « lire » *i-e* elle détermine le noyau verbal, tandis que *assnatt* « hier » porte sur tout l'énoncé.

1.3.1. Effacement des compléments

Le complément explicatif, le complément d'objet et le complément de nom font partie des expansions supprimables, excepté dans le cas des verbes dont la valence en exige la présence.

a. Le complément explicatif

Ce complément est utilisé pour expliciter l'indice personnel ou pour étoffer l'énoncé, il s'agit d'une forme d'expressivité, d'insistance ou d'explicitation, comme dans l'exemple suivant *i-dda bba* « mon père est parti ». Il peut être nécessaire selon la situation de communication de savoir à qui renvoie *i* « il ». Son effacement ne touche pas à la validité syntaxique de l'énoncé : *i-dda* « il est parti ». Au contraire, dans la locution, l'effacement du complément explicatif aboutit au défigement. Ce comportement va dans le même sens que ce que remarque Boons (1976) qui constate que la métaphore se caractérise par la présence obligatoire du complément. Ainsi dans la locution, *ffğ-ən ṯ idammən* « il est pâle », la suppression du complément explicatif *idammən* « sang » détruit le figement et donne lieu à un énoncé libre : *ffğ-ən ṯ* « ils l'ont quitté ». Dans ce cas, le pronom *ṯ* « le » ne peut renvoyer qu'à un locatif alors que dans la locution, il renvoie à un [+ humain].

b. Le complément d'objet

Ce qui est dit du complément explicatif est valable pour le complément d'objet : *i-bna tasqqurt i bba* (il construire + P piège à mon père) « Il a piégé mon père ». La suppression de *tasqqurt* (complément d'objet), affecte le figement et donne naissance à un syntagme ordinaire : *i-bna i bba* « il a construit [une maison] à mon père ».

c. Le complément de nom

Soit l'exemple suivant où la suppression du complément de nom détruit le figement.

i-gga as iġəṣ n tmidžža il faire à lui os de glotte « Il lui est insupportable »	* *i-gga as iġəṣ* il faire + P à lui os

1.4. Adjonction des expansions

Même si elles ne sont que des modifieurs de référence, les expansions adverbiales et adjectivales ne sont pas acceptées par les locutions, excepté celles qui portent sur la totalité de la locution et non sur les éléments constituants.

a. Adjonction d'un adverbe

i-kšəm t ufus il entrer + P le main (EA) « La main l'a pénétré » « Il est perturbé »	*i-kšəm t ufus bzzaf* il entrer + P le main (EA) beaucoup « La main l'a beaucoup pénétré » « Il est très perturbé »

b. Adjonction d'un adjectif

i-ffəġ asən afus il sortir + P à eux main « Il leur est sorti de la main » « Il est devenu libre »	**i-ffəġ asən afus agffas* il sortir + P à eux main droite « Il leur est sorti de la main droite »

1.5. Transformation « passive »/ « active »

La transformation passive se fait dans le parler étudié à l'aide de *ttu* (ou *tw* devant un lexème commençant par la voyelle a) (Berntolila 1983 et Galand 2002) : *iwət muna* « il a frappé Mouna » / ***ttu-****wt muna* « Mouna a été frappée ».

La transformation passive est l'un des tests les plus opératoires dans la distinction entre les syntagmes libres et les syntagmes ordinaires. Pour que l'on puisse parler du processus de passivation, il faut que le nominal contenu dans la phrase ait le statut d'un complément d'objet direct sauf dans certains cas « où l'objet désigne une propriété inaliénable (partie du corps, qualité de l'âme, etc.) du référent du sujet, [qui] n'admettent pas, ou admettent mal, d'être mises au passif. Plus

généralement si l'objet comprend un possessif renvoyant au sujet, la phrase passive est rarement naturelle » (Ruwet, 1983). Seront donc éliminées de ce test toutes les structures qui ne comportent pas de verbes transitifs. Exemple de transformation passive :

i – swa ***aman***	*ttuswa-n* ***waman***
sj. V COD	m.p. boire + P ils eau (EA)
il boire + P eau	« L'eau a été bue »
« Il a bu de l'eau »	

Il est à remarquer que contrairement au français, le complément d'agent n'est pas exprimé, et que l'objet occupe la fonction du complément explicatif.
Si nous procédons de même pour les locutions, nous aboutissons à la destruction du figement, comme étant le cas dans l'exemple suivant :

t- tfar *tmara* ***mina***	**t-**ttu**tfar* *mina*
elle suivre + P misère Mina	elle m. lié p. suivre + P Mina
« La misère a suivi Mina »	« Mina a été suivie »
« Mina est misérable »	

Bien que les verbes des exemples étudiés soient réversibles, la passivation est réfutée parce que les actants ne respectent pas les traits sémantiques du verbe. En effet, les verbes *tfur* « suivre », *tš* « manger », *awəy* « ramener », exigent un sujet [+ animé] et un complément [- animé]. Or cette distribution n'est pas respectée. Le figement a conféré aux compléments des propriétés qui ne répondent pas aux contraintes du verbe car des actants abstraits assument la fonction des animés. Le figement a fait de l'abstraction une personnification (cf. § 5).

1.6. Thématisation

La thématisation consiste dans l'anticipation de l'un des actants : complément explicatif ou complément d'objet. Ainsi au lieu de dire *i-zrəy anas muna* « Anas a quitté Mouna », nous dirons ***anas,*** *i-zrəy muna* « Anas, il a quitté Mouna » ou ***muna,*** *i-zrəy tt anas* « Mouna, Anas l'a quittée ». Dans les syntagmes ordinaires, les nominaux thématisés sont détachés du reste de l'énoncé. Ils constituent une sorte d'expansion et peuvent donc être supprimés « sans que la phrase cesse de former un tout grammatical acceptable » (Galand, 2002). Cela étant, dans les locutions, il est impossible de les supprimer du fait de leur relation étroite avec le verbe. Par conséquent, ils ne peuvent pas être considérés comme de simples expansions.
Cette opération, ainsi que celle de la focalisation que nous verrons dans le paragraphe suivant, nous permettra de voir si les constituants de la locution sont susceptibles d'être déplacés sans que leur statut ne soit détruit.

a. Thématisation du complément explicatif

i-mmut as ***wul***	****ul*** *i-mmut as*
il mourir + P à lui cœur (EA)	cœur il mourir + P à lui
« Son cœur est mort »	« Son cœur, il est mort »
« Il n'a pas de dignité »	

La thématisation du complément explicatif implique que *ul* « cœur » n'est plus à l'état d'annexion. Lionel Galand, en étudiant les rapports que peut entretenir le syntagme verbal avec le complément explicatif, a conclu qu'en amazighe « le sujet du verbe est après le verbe à l'état d'annexion et […], en base ce que l'on considère comme un sujet avant le verbe, à l'état libre n'est qu'une anticipation du sujet » (Galand, 1964). Une fois thématisé, le complément explicatif est dit *indicateur de thème* dans la terminologie de Galand et de Bentolila et *anticipation du sujet* dans celle de Basset.
Dans les locutions, la thématisation du complément explicatif n'est pas attesté, d'où le phénomène de défigement.

b. Thématisation du complément d'objet direct

i-ffeġ ixf- ənsi	*ixf- ənsi, i-ffeġ t*
il quitter +P tête son	tête son il sortir + P le
« Il a quitté sa tête »	« Sa tête, il l'a quittée »
« Il est devenu fou »	

Dans l'exemple thématisé, *t* reprend le nominal *ixəf* « tête » car en amazighe lorsque le nominal objet est anticipé, il est toujours repris par un pronom personnel objet. Il en va de même dans l'exemple qui suit, où le complément indirect thématisé *i-yifassən* est repris par le pronom *asən* « à eux ».

c. Thématisation du complément d'objet indirect

i-rẓ əm ***i yifassən***	**ifassən i- rẓ əm* ***asən***
il lâcher + P à mains (EA) + pl.	mains + pl il lâcher + P à eux
« Il a lâché les mains »	« Les mains, il les as lâchées »
« Il a renoncé à tout »	

Dans cet exemple, la locution est détruite cédant le passage ainsi à un syntagme ordinaire. Les structures obtenues après la thématisation ne sont pas acceptables en tant que groupements figés. D'une manière générale, les locutions étudiées admettent difficilement cette manipulation. Par conséquent, ce test peut être considéré comme définitoire dans la distinction entre les groupements figés et les syntagmes ordinaires.

1.7. Focalisation

En amazighe, la focalisation ou *la thématisation renforcée* dans la terminologie de Basset, se caractérise par la présence de *ay* et de la particule de prédication *d* ou de *ami* lorsque la focalisation porte sur le COI. Cette transformation sert à mettre en relief un élément de l'énoncé. Le syntagme thématisé a, selon Galand (1957), la fonction d'un prédicat.

Signalons d'emblée qu'après la mise en relief du complément explicatif, l'énoncé se transforme en une relative marquée par la présence du participe, comme dans cet exemple : ***muħa ay d*** *i-**tšan*** *ṯibawin* (Moha p .m. r. il mange + par. + P petits pois) « C'est moha qui a mangé la pomme ». Dans les locutions verbales, la focalisation est impossible, comme dans les locutions suivantes où le figement est affecté :

a. Focalisation du complément d'objet direct

y-iwey mšerḍul gg wul-əns il emporter + P souffrance dans cœur (EA) son « Il a emporté la souffrance dans son cœur » « Son cœur est brisé »	**mšerḍul ay d y-iwey gg wul- əns* souffrance p.m.r. il emporter+P dans cœur (EA) son « C'est la souffrance qu'elle a emportée dans son cœur »

b. Focalisation sur le complément d'objet indirect

i-rẓ əm i yifass- ən il lâcher + P à main + pl. (EA) « Il a lâché les mains » « Il a renoncé à tout »	**ifassən ami i- rẓ əm* main + pl p. m. rel. il lâcher + P « C'est les mains qu'il a lâchées »

1.8. Blocage des paradigmes synonymiques

*ṯ- tša ṯ **ṯisənṯ*** elle manger + P lui sel « Il est très charmant »	**ṯ-tša ṯ **ṯigni n nnɛəmṯ*** elle manger+P lui datte de la nourriture (sel) « Le sel l'a mangé »

ṯigni n nnɛəmṯ « sel » et *ṯisənṯ « sel »* sont des synonymes, pourtant ils ne sont pas interchangeables dans cette locution. D'où la disparition de la locution et l'apparition de l'énoncé, non admis dans le parler étudié.

1.9. Nominalisation

Si les syntagmes ordinaires acceptent facilement la nominalisation, les locutions la refusent de manière catégorique. Ainsi, en est-il de l'exemple suivant où la nominalisation aboutit à des énoncés ordinaires.

t- tša t tisənt elle manger + P lui sel « Le sel l'a mangé » « Il est très charmant »	**tutši n tisənt* manger(sub) de sel

1.10. Pronominalisation

La pronominalisation du complément d'objet n'est pas admise dans les locutions, contrairement à ce qui se passe dans les énoncés ordinaires car cette transformation défige la locution et donne naissance à un énoncé ordinaire, comme dans le cas ci-après :

i- gga ixəf il faire + P tête « Il a fait la tête » « Il est devenu raisonnable »	**i-gga t* il faire + P le « Il l'a fait »

1.11. Interrogation

Que l'interrogation porte sur le complément explicatif, le complément d'objet ou sur le complément nominal, elle ne porte jamais atteinte à l'intégrité sémantique d'un énoncé libre. Or, cette opération déstabilise la locution et la défige, comme le montrent les exemples suivants :

a. Interrogation portant sur le complément explicatif

t-swa- t tfunast elle boire + P lui vache « La vache l'a bu » « Il a tout perdu »	*mayd t iswan ?* qui lui boire + Part. « qui l'a bu ? »

b. Interrogation porte sur le COD

t-iwəy- d adis elle ramener + P p. or. ventre « Elle a ramené le ventre » « Elle est tombée enceinte »	**mayd di t- iwəy ?* quoi p. or. elle ramener + P « Qu'est ce qu'elle a ramené ? »

c. Interrogation porte sur le COI

i-rẓəm i ifass- ən elle laisser tomber + P à main + pl.	**mami i- rẓəm* à quoi il laisser tomber + P

« Il a laissé tomber ses mains » « Il a renoncé à tout »	« Qu'est ce qu'il a lâché ? »

d. Interrogation porte sur le CC

i-ffəġ ġər iħiħan il sortir + P à ------- « Il déraisonne »	*ma ġər i-ffəġ* quoi à il sortir + P « Où est-ce qu'il est sorti ? / Chez qui est-il sorti ? »

1.12. Coordination

La coordination permet de faire l'ellipse d'une unité commune à deux propositions. Dans les locutions, toute tentative de coordonner deux nominaux associés à un même verbe et occupant la même fonction est vouée à l'échec car elle détruit le sens de la locution. Exemple :

t- tša t tisənt t- tša t tafugt
elle manger + P lui sel
elle manger + P lui soleil
« Le sel l'a mangé »
« le soleil l'a mangé »
« Il est charmant »
« Il a pris un coup de soleil »

*t- tša-t tisənt d tafugt
elle manger + P le sel et soleil
« Le sel et le soleil l'ont mangé »

1.13. Portée du figement

Les locutions verbales renferment au moins un élément variable susceptible d'être remplacé ou modifié. Cet élément peut être : l'indice de personne, le complément d'objet direct ou le complément d'objet indirect. Ainsi l'indice de personne peut être remplacé par n'importe quel pronom personnel sujet : *t* « elle », *n* « ils », *nt* « elles », *x* « je », *t-m* « vous », *t-mt* « vous + fém. » et l'insertion d'un complément explicatif lui correspondant est possible. En revanche, quand le complément explicatif est exprimé dans la locution, l'indice de personne perd son autonomie et constitue avec son complément un bloc figé, comme dans l'exemple ci-dessus : ***t- tša t tisənt*** (elle manger + P lui sel) « Le sel l'a mangé » = « Il est charmant ».

1.14. En résumé

Le figement ne produit pas uniquement des séquences totalement figées, mais également des séquences favorables à la variation. En effet, contrairement aux locutions nominales, les locutions verbales renferment souvent un élément libre servant de support à la variation. En outre, si les locutions nominales rejettent, d'une manière presque systématique, les différentes manipulations admises par la syntaxe ordinaire, les locutions verbales, quant à elles, n'ont pas le même comportement vis-à-vis de ces tests : certaines locutions admettent les manipulations, tandis que d'autres les rejettent, et ce à des degrés variés : les locutions n'ont pas les mêmes caractéristiques ni les mêmes comportements.

La variation et la liberté des locutions ne sont pas tributaires des caractéristiques de la catégorie syntaxique à laquelle elles appartiennent mais plutôt du degré d'opacité. En effet, plus la locution est opaque, moins elle accepte le changement, et moins la locution est opaque, plus elle admet les manipulations. Autrement dit, les locutions dont le sens ne peut être déduit des éléments constituants, sont réfractaires aux différentes transformations. Pour plus de précision, nous interrogeons dans la partie suivante les constituants variables et les constituants invariables dans les locutions verbales.

2. Structure syntaxique et invariabilité des constituants

Nous avons évoqué à plusieurs reprises le caractère figé des locutions verbales. Nous entendons par « figé », tout élément refusant la variation. Il est néanmoins nécessaires de remarquer que certaines structures comportent un seul élément figé, d'autres plus d'un seul. Les éléments non figées peuvent être remplacés par quelconque élément pouvant assumer la même fonction syntaxique à condition, bien évidemment, qu'il respecte les traits sémantiques du verbe. Nous donnons ici l'inventaire des structures de locutions verbales en précisant leur degré de figement (les composants figés sont notés en gras).

1. Structure à SV **+ COD** : *i-**gga ixəf*** (il faire + P tête) « Il a fait la tête » = « Il est devenu raisonnable ». L'indice de personne *i* « il » est l'élément variable. Il n'est pas affecté par le figement. Par conséquent, il peut être remplacé par des éléments assumant « la fonction sujet » : *t-**gga ixəf*** (elle faire + P tête) « Elle a fait la tête » = « Elle est devenue raisonnable » ; ***gga-nt ixəf*** (faire + P elles tête) « Elles ont fait la tête » = « Elles sont devenues raisonnables ».

2. Structure à SV + COI+ **COD :** *i-**ša yas awal*** (il donner + P à lui parole) = « Il lui a fait une promesse ». L'indice de personne et le complément COI sont les éléments variables, ils peuvent être remplacés ainsi : *t-**ša yasən awal*** (elle donner + P à eux parole) « Elle leur a fait une promesse ».

3. Structure à SV + COI+ **CC :** *i-**ssa** yas **xəf lwxa*** (il caus. vider + P à lui sur vide) = « Il l'a trahi ».

4. Structure à **SV + CE :** ***i- qqur ubriḏ*** (il ê. sec + P bouc (EA) = « Il n'y a rien à espérer ». Contrairement aux autres structures dégagées, cette structure présente un figement absolu. En effet, tous les éléments doivent être présents pour que la locution continue d'exister.

5. Structure à **SV** + COI + **COD** + poss. **:** ***i-ša** yas **wul-**əns* (il donner + P à lui cœur (EA) son) = « Il a osé faire… ».

6. Structure à SV **+ Prép. +** CC **:** *i-**dur diys*** (il tourner + P dans lui) **=** « Il lui a crié dessus »

7. Structure à **SV** + COD + **CE :** ***i-tfar** t̲ **ṣṣ**xəḏ* (il suivre + P le malédiction) = « Il est maudit ».

8. Structure à SV **+ COD** + CC: ***i-bna kuši** ġifs* (Il construire + P tout sur lui) = « il compte trop sur lui ».

9. Structure à **SV** + COI + **CE + CC :** ***y-uly** as **igdi ġif t̲xamt̲*** (il monter + P à lui chien sur tante (EA)) = « Il est vil ».

10. Structure à **SV + CE + CC** + poss. **:** ***brrm-ənt šlaġəm gg uqmu**-ns* (tourner + P elles moustaches dans visage son) = « Il a grandi ».

11. Structure à SV **+ COI :** *i-**rzəm i yifassən*** (il lâcher + P à main (EA) + pl.), « Il a lâché les mains » = « Il a renoncé à tout ».

12. Structure à SV + COI + **COD + CC :** *i-**gga** as **iġəṣ gg t̲migga*** (il faire + P à lui os dans glotte (EA)) = « Il lui est insupportable ».

13. Structure à **SV** + COI + **CE+ CC :** ***i-ffəġ** as **rray afus*** (il sortir + P à lui décision main), « La décision lui est sortie de la main » **=** « Il n'a plus de pouvoir ».

14. Structure à **SV :** ***t̲-fukka*** (elle finir + P), « Elle est finie » = « Il n'y a plus rien à dire » (même commentaire que la structure 4).

15. Structure à SV **+ COD + CC+ COD+ CC :** *i-**gga afus dat̲ afus dart̲*** (il mettre + P main devant main derrière), « Il a mis la main devant la main derrière » **=** « Il ne possède plus rien ».

16. Structure à SV **+ COD + SV :** *i-**gra t̲ i-qqərs*** (il jeter + P le il coaguler + P**) =** « Il a tout perdu ».

17. Structure à **SV+ CE + CC** + poss. : ***i-bdda wazzar gg ixəf-**ns*** (il ê. debout + P cheveux (EA) dans tête son) = « Il a très peur »

18. Structure à SV+ SV **+ CC :** ***da** i-**ktšəm ar** i-**ttffəġ gg wawal*** (réel il entrer + AI réel il sortir + AI dans parole (EA)) = « Il dit n'importe quoi ».

Si nous devions reprendre le classement des structures selon le degré de figement externe, c'est-à-dire en fonction des éléments variables ou non variables, nous pouvons donner le classement suivant :

1. locutions totalement figées : celles dont aucun élément n'est libre (structures 4 et 14)

2. locutions contenant un seul élément libre (structure 1)

3. locutions contenant plus d'un élément libre. (cf. les 15 autres structures).

3. Elaboration de la base de données : éléments de méthodologie

3.1. Présentation du corpus

Le corpus établi et analysé comprend plus de 1000 locutions. Ces dernières sont classées selon l'ordre alphabétique de la racine d'où sont issus les verbes de la locution. Chaque racine est indiquée en lettres capitales notées en italique et en gras. Dans la première ligne qui suit la racine, sont présentées les formes de l'aoriste, de l'aoriste intensif, du prétérit et du prétérit négatif. Ces formes sont écrites en *italique*. Exemple :

- ***SW***

  səw, ssa, swa, swi

Pour éviter toute confusion, le prétérit est toujours énoncé même dans les cas où sa forme coïncide avec celle de l'aoriste. Exemple : *ffəġ* « sortir + A » et *i- ffəġ* « Il sorti + P »

Sur la deuxième ligne sont notées entre < > les différentes acceptions de chaque verbe. Les dérivés (passifs, causatifs) sont également classés sous la même entrée que leur forme simple. Exemple : *ssird* <laver> est classé sous l'entrée *rid* <ê. lavé>

Nous avons conservé les différentes particules accompagnant la base verbale et notamment les particules d'orientation car leur effacement peut entraîner une altération de la locution. Exemple : ***ffəġ-n ṯ id iġṣan*** (sortir + P ils le p. or. os + pl.), « Ses os sont sortis » = « Il est maigre ». Avec la suppression de la particule d'orientation, nous obtenons : ***ffəġ-n ṯ iġṣan*** (sortir + P ils le p. or. os + pl.) = « Ses os sont sortis ».

Chaque locution est suivie de trois traductions : (1) une traduction juxtalinéaire, (2) une traduction littérale, et (3) une traduction littéraire.

i-ffəġ *ṣṣurṯ-əns*

1. il sortir + P image sa
2. « Il est sorti de son image ».
3. « Il a changé »

3.2. Problèmes de traduction

> Une étude contrastive entre deux langues différentes plutôt non apparentées, telles que le français et l'amazighe pose beaucoup de problèmes notamment d'ordre « traductologique ». En effet, la rencontre de deux langues n'est pas toujours d'ordre linguistique mais aussi d'ordre socioculturel[2].

La traduction consiste en le passage d'un code linguistique, de la langue source, à un autre code de la langue cible. Il ne s'agit pas uniquement de traduire et transmettre mécaniquement des unités lexicales d'une langue à une autre, mais également de transmettre des connaissances et des données extralinguistiques relatives au domaine culturel où est née cette langue. De ce fait, le passage d'une langue source à une langue d'arrivée n'est pas une tâche aisée. C'est un travail qui exige du traducteur une bonne connaissance et une capacité à comprendre et à saisir les différents aspects des deux langues.

La traduction des énoncés figés est plus délicate que le reste de la langue car, en plus de la différence au niveau des structures, chaque langue à sa manière de présenter et de découper le monde et de renvoyer à une symbolique qui n'est pas de portée commune à toutes les langues.

Pour notre part, nous n'avons pas pris, à chaque fois, en considération les images exprimées par les locutions parce que le français, qui est notre langue d'arrivée, n'utilise pas nécessairement les mêmes images que l'amazighe. Notre intérêt était de tenir compte spécialement de l'effet du poncif qu'ils comportent car traduire littéralement une locution n'évoque pas le même sens et, parfois, la structure à laquelle nous aboutissons n'est pas attestée dans la langue française. Ex. *: i-k ✶em-✹ ufus* « la main l'a pénétré ! » = « il est paniqué ». En outre, certains termes n'apparaissent que dans les locutions ; ils ne sont pas connus en dehors du figement et ne correspondent exactement à aucune unité lexicale particulière du français, d'où la difficulté de les traduire de façon satisfaisante dans la langue cible. Ainsi, la

---

[2]L'amazighe et le français diffèrent l'une de l'autre concernant leurs origines, leurs structures et leurs évolutions. Le français appartient à la famille des langues indo-européennes, l'amazighe, quant à lui, appartient à l'arbre généalogique des langues afro-asiatiques. Le français est une langue à tradition écrite, l'amazighe a porté, durant de longues années, l'étiquette de langue à tradition orale. Le français est une langue à mots, l'amazighe est une langue à racine. Ces deux langues ont cohabité pendant la période coloniale, et continuent de vivre côte à côte grâce à l'immigration. Ce contact a laissé des traces notamment au niveau lexical. En effet, le lexique de la langue amazighe est taché de plusieurs mots empruntés à la langue française et vis-versa. Autrement dit, l'influence lexicale et bidirectionnelle. Lors de la sélection de notre corpus, nous avons constaté qu'un nombre assez important de locutions véhiculent le même sens et la même image que leurs correspondantes en français.

perte du sens littéral ainsi que l'effet connotatif et la valeur métaphorique ont été négligés au profit du contenu sémantique des locutions verbales. Pourtant, nous avons traduit mot-à-mot nos locutions quand la structure et les composants de celui-ci le permettent.

L'intérêt de cette méthode est de segmenter l'énoncé transcrit en alphabet phonétique internationale afin de faciliter la tâche aux apprenants en s'initiant également la structure de l'énoncé en amazighe.

Les locutions véhiculant la même image ne posent aucun problème de traduction. Avec ce type de locutions, nous sommes devant la stratégie de l'équivalence. En fait, elles dégagent, dans les deux langues, des similarités connotatives très étroites.

Cela dit, il semble très important, au niveau contrastive et didactique, que les apprenants soient conscients du fait que la signification de ce type d'expressions n'est pas le cumul des sens véhiculés par leurs constituants et que le contexte et la situation de communication permettant de concevoir le sens de ces énoncés.

Pour mieux comparer les expressions figées dans les deux langues, nous avons cherché des locutions verbales françaises ayant soit le même sens, soit la même image, soit les deux. Nous avons évoqué également des locutions dont l'image est identique et le sens est différent afin d'avertir l'apprenant du danger de la traduction littérale.

Nous nous sommes intéressés uniquement aux locutions qui sont très utilisées aussi bien dans la langue source que dans la langue d'origine. Des exemples qui nous ont servi d'illustration consacrée à l'analyse des critères d'identification des locutions, dans la première partie, peuvent ne pas apparaître dans les tableaux ci-après.

Nous avons fait abstraction de toutes les locutions en voie du figement comme les énoncés qui s'articulent autour d'une comparaison explicite : *amm wayur* « comme la lune = beau / belle ». En effet, ce type de locutions ne posent pas de problèmes de compréhension car leur sens s'affiche dans la première partie de l'énoncé. La deuxième partie ne sert donc qu'à renforcer et à emphatiser l'image véhiculée. Elle constitue une sorte de redondance.

Comme nous l'avons mentionné ci-dessus, puisque l'amazighe est une langue à racine, nous avons préféré classer nos locutions par racine. Seront classés sous cette entrée, les locutions comportant la même vedette formelle.

Pour mieux décoder les locutions verbales amazighes, nous avons présenté notre corpus comme suit : Nous avons présenté les locutions dans des tableaux comportant trois colonnes : la première contient les locutions amazighes, la deuxième leurs correspondantes en français et la troisième leur signification. En revanche, les locutions verbales dont l'image est identique mais le sens est différent sont présentées dans des tableaux à quatre colonnes. La première est

réservée aux locutions verbales amazighes, la deuxième à sa signification, la troisième aux locutions françaises et la quatrième à la signification de celles-ci.

Les exemples amazighes sont transcrits en alphabet phonétique international, en gras et en italique. Ils ne sont pas mis entre crochets.

5. Illustrations de la base de données

5.1. Locutions dont le sens et l'image sont identiques

♦ ***AMN***
amən, *ttəmən*, *umən*, *umin*
< croire >

Locution amazighe	Sa correspondante en français	Signification
ur y-umin allən-əns nég. il croire + P yeux ses Il ne croit pas ses yeux	Ne pas en croire ses yeux	**Avoir du mal à admettre la vérité**

♦ ***AMẒ***
amәẓ, *ttamәẓ*, *umәẓ*, *umiẓ*
< tenir, saisir >

Locution amazighe	Sa correspondante en français	Signification
y-umẓ ṣṣrima il tenir + P bride Il tient la bride	Tenir la bride	**Maintenir**

♦***ANF***
anf, *ttanəf*, *unəf*, *unif*
< ouvrir>

Locution amazighe	Sa correspondante en français	Signification
y-unəf imllann Il ouvrir + P oreilles Il a ouvert les oreilles	Ouvrir les oreilles	**Ecouter attentivement**

♦***Ašk***
a✶k, *tta ✶ka*, *u ✶ka*, *u ✶ki*
< perdre, se perdre >

Locution amazighe	Sa correspondante en français	Signification
y-u ✶ka yas wawal **(EA)** il perdre + P à lui parole Il a perdu ses mots	Perdre sa langue	**Se taire**

♦ ***BY***

bbəy, *ttbbəy*, *bbəy*, *bbiy*

< couper, traverser>

Locution amazighe	Sa correspondante en français	Signification
da i-ttbbəy ul réel il AI déchirer cœur Il coupe le cœur	Déchirer le cœur	**Attrister**

♦ ***FFĠ***

ffəġ, *ttffəġ*, *ffəġ*, *ffiġ*

< sortir, quitter, abandonner >

Locution amazighe	Sa correspondante en français	Signification
i-ssu fə + d uxsan il caus. Sortir + P vers ici molaires Il a fait sortir les molaires	Montrer les dents	**Avoir une attitude hostile**

♦ ***ĦMU***

ħmu, *ttəħmu*, *ħma*, *ħmi*

< être chaud >

Locution amazighe	Sa correspondante en français	Signification
ħma- n as idammən être chaud ils à lui sang Son sang est chaud	Avoir le sang chaud	**Etre coléreux**

♦ ***KK***

kk, *ttkka*, *kka*, *kki*

< passer >

Locution amazighe	Sa correspondante en français	Signification
i-kka ddaw yifər-əns Il passer + P sous aile son Il est passé sous son aile	Etre sous l'aile de quelqu'un	*Etre sous sa protection*

♦ ***LH***

liha, *ttliha*, *liha*, *liha*

< jouer >

Locution amazighe	Sa correspondante en français	Signification
***da i- ttliha s wafa*(EA)** réel il AI jouer avec feu Il joue avec le feu	Jouer avec le feu	**Jouer avec le danger**

♦ ***LL***

li, *ttili*, *lla*, *lli*

< être, exister >

Locution amazighe	Sa correspondante en français	Signification
lla-nṯ +ars ✷i+ʷər…in être + P elles chez lui épaules Il a des épaules	Avoir les épaules larges	**Etre fort pour, être capable pour**

♦ ***MMT̲***

mmə✷, *ttəm✷a✷*, *mmu✷*, *mmu✷*

< mourir >

Locution amazighe	Sa correspondante en français	Signification
i-mmu✷ s ✷at✸a il mourir + P avec rire **Il meurt de rire**	Mourir de rire	Rire aux éclats

♦ ***MNQR***

mənqqər, *ttmənqqar*, *mənqqer*, *mənqqir*

< claquer >

Locution amazighe	Sa correspondante en français	Signification
da-ttmənqar- ən✹ ✹uġmas- ənsréel claquer + P elles dents ses Ses dents claquent	Claquer des dents	**Avoir très froid**

♦ ***RẒ***

rəẒ, *rẒẒa*, *rẒa*, *rẓi*

< casser >

Locution amazighe	Sa correspondante en français	Signification
rẓa- n✹ ✹inzar- ənscasser + P elles narines ses Ses narines sont cassées	Se casser le nez sur quelque chose	**Echouer**

♦ ***RẒM***

rẒəm, rẒẒəm, rẒəm, *rẒim*

< libérer, laisser tomber, détacher >

Locution amazighe	Sa correspondante en français	Signification
i-rẓəm i yifassən il laisser tomber + P à mains Il a laissé tomber ses mains	Baisser les bras	**Renoncer**

♦ ***SWL***

sawəl, *sawal*, *siwəl*, *siwil*

< parler >

Locution amazighe	Sa correspondante en français	Signification
da i-sawal z ggul- əns réel il parler + P de cœur son Il parle de son coeur	Parler à cœur ouvert	**Parler avec franchise**

♦ ***TŠ***

tš, *tetta*, *tš*, *tši*

< manger >

Locution amazighe	Sa correspondante en français	Signification
***i-t✶a ✹ s wallən* (EA)** il manger + P le avec yeux Il le mange des yeux	Manger des yeux	**Regarder avidement**

♦ ***XW***

xwu, *ttəxwu*, *xwa*, *xwi*

< vider >

Locution amazighe	Sa correspondante en français	Signification
i-ssxwa ulə ns il vider + P cœur son Il a vidé son cœur	Vider son cœur	**Révéler ses sentiments**

♦ ***W***

awə…, *ttawə…*, *iwə…*, *iwi…*

< arriver, masser>

Locution amazighe	Sa correspondante en français	Signification
y-iwə… as ✝ər ul Il arriver + P à lui jusqu'à cœur Il lui est arrivé au cœur	Gagner le cœur de quelqu'un	**Le séduire**

♦ ***WSX***

wssəx, *ssusəx*, *wssəx*, *wssix*

< salir >

Locution amazighe	Sa correspondante en français	Signification
i-ssusəx ifassən-əns il salir + P mains ses Il a sali ses mains	Avoir les mains sales	Participer à une affaire louche

♦ ***YR***

yər, *ggar*, *yra*, *yri*

< jeter, ajouter, verser >

Locution amazighe	Sa correspondante en français	Signification
i-yəra zzi✹ i wafa **(EA)** il mettre + P huile à feu Il a mis l'huile sur feu	Jeter de l'huile sur le feu	Exciter davantage

♦ ***YY***

yy, *ttgga*, *yya*, *yyi*

< mettre, poser >

Locution amazighe	Sa correspondante en français	Signification
i-yya as lmus g uyər… **(EA)** il mettre + P à lui couteau dans gorge Il lui a mis le couteau sous la gorge	Mettre le couteau sous la gorge	**Menacer**

5.2. Locutions dont le sens est identique mais l'image est différente

♦ ***AŠK***

a ✶*k*, *tta* ✶*ka*, *u* ✶*ka*, *u* ✶*ki*

< perdre, se perdre >

Locution amazighe	Sa correspondante en français	signification
u ✶***ka- n✹ as lf✹la✹*** perdre elles à lui mots Il a perdu ses mots	Il a perdu le nord	**Etre dans la confusion**

♦ ***BDD***

bədd, *ttbədda*, *bdda*, *bddi*

< être debout, arriver >

Locution amazighe	Sa correspondante en français	signification
i-bədda yas d wass il arriver + P à lui vers ici jour son jour est arrivé	- casser sa pipe. - passer l'arme à gauche. -avaler son bulletin de naissance. - Manger les pissenlits par les racines. - Fermer son parapluie	**Mourir**

♦ ***FFĠ***

ffə+, *ttffə+*, *ffə+*, *ffi+*

< sortir, quitter >

Locution amazighe	Sa correspondante en français	signification
i- ffə+ d ul- əns il sortir + P vers ici cœur son Son cœur est sorti	Ne pas porter quelqu'un dans son cœur	**Le détester**

♦ ***FẒẒ***

fəẓẓ, *ttfẓẓa*, *fẓẓa*, *fẓẓi*

< mâcher, mordre >

Locution amazighe	Sa correspondante en français	signification
i- fẓẓa illəs *il mordre + P langue* *Il a mordu sa langue*.	casser sa pipe. (cf. ***BDD***)	**Mourir**

♦***GG D***

ggəɛd, *ttggɛad*, *ggəɛd*, *ggɛid*

< monter >

Locution amazighe	Sa correspondante en français	signification
✹ ***ggəɛd as ttəx*** ✹ elle monter + P à lui lie La lie lui a monté	avoir la moutarde qui monte au nez	**Etre irrité**

♦ ***ḤQQ***

ḥqqa, *ttḥqqa*, *ḥqqa*, *ḥqqi*

< voir, regarder, contempler, méditer >

Locution amazighe	Sa correspondante en français	Signification
da ✹ i-ttħqqa s i✹rəm n ✹i✹✹ réel le il AI regarder avec coin de œil Il le regarde du coin de l'œil	Regarder quelqu'un de travers	**Lui témoigner de l'hostilité.**

♦ ***X☒R***

xi ✹ər, *ttxi ✹ir*, *xa ✹ər*, *xa ✹ər*

< grandir, devenir grand >

Locution amazighe	Sa correspondante en français	Signification
i-sxa✹r ixf- əns Il agrandir + P tête sa	Avoir la grosse tête	**Eprouver une fierté excessive**

♦ ***K̲ŠM***

k̲ ✶əm, *k̲ ✶ ✶əm*, *k̲ ✶əm*, *k̲ ✶im*

< entrer, pénétrer >

Locution amazighe	Sa correspondante en français	Signification
i-k ✶əm ⌊uq n ixf-əns Il entrer + P souk de tête sa	Se mêler de ses affaires	**Ne pas s'occuper des problèmes d'autrui**

♦ ***L***

li, *ttili*, *lla*, *lli*

< être, exister >

Locution amazighe	Sa correspondante en français	Signification
i-lla dgər wagmar ttagmar✹ il être + P entre cheval jument Il est entre le cheval et la jument	Il est entre le marteau et l'enclume	**Etre dans une situation difficile**

♦ ***SN***

isin, *ttisin*, *ssən*, *ssin*

< savoir, connaître >

Locution amazighe	Sa correspondante en français	Signification
i-ssən abrid il connaître + P chemin Il connaît le chemin	- connaître la musique	**Savoir comment faire**

♦***SR***

ssara, *ssara*, *ssara*, *ssara*

< chercher, se promener, se balader >

Locution amazighe	Sa correspondante en français	Signification
da i- ssara ⵜif wawal **(EA)** réel il AI chercher sur parole Il cherche la parole	chercher la petite bête	**Chercher le petit défaut pour déprécier.**

♦ ***SW***

səw, *ssa*, *swa*, *swi*

< boire >

Locution amazighe	Sa correspondante en français	Signification
✹- swa ✹ ✹funas✹ **(EA)** elle boire + P le vache La vache l'a bu	Boire la tasse	**Echouer, avoir tout perdu**

♦ ***ULY***

aləy, *ttaləy*, *uləy*, *uliy*

< monter, grimper >

Locution amazighe	Sa correspondante en français	Signification
y-uly as lⵜ^{w}mam ⵜif wallən **(EA)** Il monter + P à lui brouillard sur yeux Le brouillard lui est monté sur les yeux	Avoir un bandeau sur les yeux	**Manquer de clairvoyance**

♦ ***U☒***

u✸, *kka✸*, *u✸*, *wi✸*

< frapper, taper >

Locution amazighe	Sa correspondante en français	Signification
i-wə✸ dduni✸ s u…ar il frapper + P vie avec pied Il a frappé la vie avec un coup de pied	Se la couler douce	**Mener une vie insouciante**

♦ ***YR***

yər, *ggar*, *yra*, *yri*

< jeter, lancer >

Locution amazighe	Sa correspondante en français	Signification
✸- yra ✸ yrutt elle jeter + P le grenouille La grenouille l'a jeté	avoir la chair de poule	**Avoir froid**

5.3. Les locutions dont l'image est identique mais le sens est différent

Nous avons relevé dans notre corpus trois locutions ayant la même image dans les deux langues mais deux sens différents

♦ ***DDU***

ddu, *ttddu*, *dda*, *ddi*

< partir, marcher >

Locution amazighe	signification	locution française	signification
da i-ttddu ġif ✸əylay réel il AI marche sur œufs Il marche sur les œufs.	*Marcher lentement*	Marcher sur des œufs	*Agir avec précautions*

♦ ***FS***

fsis, *ttəfsis*, *fssus*, *fssus*

< être léger >

Locution amazighe	signification	locution française	signification
i-fssus as ufus **(EA)** il être léger + P à lui main Sa main est légère	**C'est un voleur**	Avoir la main légère	**Agir avec douceur**

♦ ***LL***

ili, *ttili*, *lla*, *lli*

< exister, être, posséder >

locution amazighe	signification	locution française	signification
lla- n✹ ġars ✹iyənzar exister + P elles chez lui nez Il a le nez	**Il a sa fierté**	Avoir du nez	**Etre très perspicace**

Conclusion

Comme le montre ces données, les deux langues jouissent d'un nombre assez important de locutions qui se rapprochent aussi bien au niveau de l'image que du sens. Cela explique d'une part le fruit du contact des langues, d'autre part les universaux linguistiques et culturels. Par contre, les locutions verbales dont les images sont différentes reflètent ce que nous pouvons désigner de particularisme de l'expérience humaine. En effet, et comme nous l'avons signalé plus haut, l'amazighe et le français sont deux langues et deux cultures différentes. D'où l'écart au niveau de quelques images métaphoriques. Tandis que l'amazighe se sert d'une image relative au domaine bestial pour décrire quelqu'un qui est dans une situation difficile : *illa d ger wagmar ttagmart* <il est entre le cheval et la jument>, la langue française, quant à elle, recourt au domaine de l'industrie : *être entre le marteau et l'enclume.*

Ces images évoquées rapprochent le chercheur de la réalité de chaque langue. C'est pour cette raison qu'il faut essayer, lors de la traduction, de garder ces concepts culturels même si le sens ne peut jamais être rendu d'une manière fidèle. Un autre problème à souligner, est celui du troisième tableau où sont classées les locutions ayant la même image mais de sens différent. C'est ce type d'énoncés qui induit l'apprenant en erreur. D'où la nécessité de connaître, outre les bases linguistiques de la langue étrangère étudiée, des notions culturelles de cette langue.

Nous avons opté pour un classement par racine car cette méthode de classement aide l'apprenant à collecter un nombre assez important de locutions se rapportant à chaque racine. Une autre méthode d'ordre thématique est appliquée dans le domaine de la didactique. Elle consiste à regrouper toutes les locutions exprimant le même thème bien qu'ils se rapportent aux différentes racines. Ce type d'analyse sera développé dans un travail ultérieur.

Références

BENTOLILA F. (1969), « Les modalités d'orientation du procès en berbère », *La linguistique*, Paris, fasc. 2, pp. 91-111.

BENTOLILA F. (1981), *Grammaire fonctionnelle d'un parler berbère, Aït Seghrouchen d'Oumjeniba (Maroc),* Paris, SELAF.

BOUCHARD C. (2002), « La locution : problèmes de traduction », *Le moyen français*, Paris-Louvain, pp. 19-27.

CHAKIRI M. (2002), « Locutions et verbes de mouvement en berbère marocain (Moyen-Atlas) », *Actes de la Journée d'étude de la Formation doctorale du Département de linguistique générale et appliquée*, n°7, Université René Descartes-Paris5-Sorbonne, pp. 9-20.

CHAKIRI M. (2004), "Expressions figées et fonctions discursives", *Langages et Politiques Linguistiques Actes du XXVIème Colloque international de linguistique fonctionnelle*, Guadeloupe (Gosier), Peter Lang, pp. 255-228.

CHAKIRI M (2007) « Les particules d'orientation en berbère », *SEMEION. Travaux de sémiologie*, n°5, Université Paris Descartes, Laboratoire DYNALANG-SEM, Sorbonne, Paris, pp. 35-39.

Galand P. et Galand L., (1993), *Comptes rendus du groupe linguistique d'études chamito-sémitiques, GLECS*, Suppléant 15, Paris, Geuthner, pp. 161-175.

FONAGY I. (1997), « Figement et changements sémantiques », *La locution entre langue et usages*, Coll. Signes, ENS. Editions Fontenay/Saint-Cloud, Ophrys, pp. 31-164.

GAATONE D. (1981), - « La locution verbale : pourquoi faire ? », *Revue Romaine*, n°16-4, 1981, pp. 49-73.

GALAND L. (1964), « L'énoncé verbal en berbère. Étude de fonctions », *Cahiers Fernand de Saussure*, n° 21, pp. 33-53.

GROSS G. (1996), *Les expressions figées en français*, Paris. Ophrys.

TAIFI M. (1991), *Dictionnaire tamazight-français* (Parler du Maroc central), Paris, L'Harmattan-Awal.

L'apport fondamental des ontologies pour le Web intelligent : Web 3.0

Hammou Fadili[1, 2]
[1]Laboratoire CEDRIC du Conservatoire National des Arts et Métiers de Paris
192, rue Saint Martin, 75141, Paris cedex 3, France
hammou.fadili(at)cnam.fr
[2]Pôle scientifique de la Maison des Sciences de l'Homme de Paris
190 avenue de France 75648 Paris Cedex 13, France
hammou.fadili(at)msh-paris.fr

Résumé

La notion « d'intelligence » sera au cœur des systèmes d'information futurs. Toutes les technologies (protocoles, normes, outils, etc.) mises en œuvre pour sa formalisation sont entrain d'être standardisées autour de l'internet 3ème génération ou le WEB 3.0. Cet aspect important caractérisant cette version du Web repose essentiellement sur la fusion et l'évolution du WEB 2.0 et du WEB sémantique. Il correspond à la résultante de l'intelligence collective humaine représentée dans les systèmes actuels, interactifs et coopératifs comme les réseaux sociaux et de l'intelligence artificielle gérée par les machines et représentée dans le WEB sémantique et la notion d'agents intelligents.

En parallèle, la notion d'ontologies évolue et devient fondamentale et incontournable dans la définition de la structure et de l'architecture des nouveaux systèmes, au niveau de la modélisation, conceptualisation et représentation des connaissances partagées, mais aussi au niveau du raisonnement, actions et interactions entre les différents agents humains et logiciels.

La première partie de cet article sera consacrée à la présentation des différents concepts ainsi qu'à la description de leur évolution et leur convergence vers un nouveau standard cohérent qui est le WEB 3.0. La deuxième partie sera consacrée à la notion d'ontologies et le rôle qu'elle peut jouer dans le partage et l'utilisation consensuels des concepts et des connaissances entre les différents acteurs humains et logiciels.

Keywords - Mots Clés

Web 2.0, Web sémantique, agents intelligents, ontologies, Web 3.0
Web 2.0, Semantic Web, Intelligent Agents, Ontologies, WEB 3.0

1. Introduction

Les technologies, issues des premières versions de l'Internet, sont arrivées à maturité et se sont spécialisées 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 des applications logicielles connectés dans le réseau mondial, comme les programmes interactifs d'aide à la création et à la diffusion de contenus. D'autres technologies commencent à émerger pour introduire et incorporer la notion de la sémantique dans les contenus afin qu'elle soit prise en charge par les machines à travers les technologies du Web sémantique. Cette version du Web combinée au WEB sémantique est amenée à évoluer. On parle déjà du WEB 3.0 pour la future version qui aura pour rôle de concilier l'internet, la sémantique et l'intelligence artificielle.

La notion d'intelligence sera au cœur des systèmes d'information futurs. Toutes les technologies (protocoles, normes, outils, etc.) mises en œuvre pour sa formalisation sont entrain d'être standardisées autour de l'internet 3ème génération, appelé le WEB 3.0. Cet aspect important caractérisé par la notion « d'intelligence » repose essentiellement sur la fusion et l'évolution du WEB 2.0 et du WEB sémantique dans un contexte multi-agents. Il correspond à la résultante de l'intelligence collective humaine représentée dans les systèmes actuels, interactifs et coopératifs comme les réseaux sociaux et l'intelligence artificielle véhiculée et gérée par les machines représentée dans le WEB sémantique et la notion d'agents intelligents. Dans tous ces systèmes, la notion d'ontologies est fondamentale dans la définition de leur structure et leur architecture, que ce soit au niveau de la modélisation, conceptualisation et représentation des connaissances partagées, qu'au niveau du raisonnement, actions et interactions entre les différents agents humains et logiciels.

La première partie de cet article sera consacrée à la présentation des différentes approches constituantes du WEB 3.0, ainsi qu'à la description de leurs évolutions et de leur convergence vers ce nouveau concept émergent. La deuxième partie sera consacrée à la définition et la description de la notion du WEB 3.0 ainsi qu'aux liens et rapports qu'entretiennent les différents éléments le constituant. La troisième partie décrit le rôle fondamental que les ontologies peuvent jouer dans le partage et l'utilisation consensuels des connaissances entres les différents acteurs humains et logiciels ; nécessaires à leur fonctionnement.

2. Eléments principaux du Web 3.0

Dans ce qui suit, nous allons présenter quelques éléments, constituants essentiels de la technologie WEB 3.0. Ce qui nous permettra de mettre en avant la définition et l'architecture de chacun de ces éléments, d'aider à définir la nouvelle technologie du WEB 3.0 et aider à comprendre l'apport fondamental de la notion d'ontologie pour ce nouveau concept.

2.1. Ontologies

D'une manière générale, les ontologies reposent sur des outils de modélisation et de représentation des connaissances permettant à des communautés d'experts humains et logiciels d'un domaine donné de partager, d'une manière consensuelle, un vocabulaire et de parler un même langage, i.e. communiquer et se comprendre. Définir une ontologie d'une manière précise est une tâche très difficile. En effet, la notion d'ontologies est très vaste et difficile à définir, d'ailleurs, plusieurs définitions existent suivant le contexte d'utilisation. Celle qui fait autorité aujourd'hui, au sein de la communauté scientifique, est celle de Gruber qui la définit une comme suit :

« Une ontologie est la spécification d'une conceptualisation d'un domaine de connaissance ». En d'autres termes, une ontologie est un modèle de représentation de la formalisation d'une conception d'un domaine.

La modélisation d'un domaine repose essentiellement sur deux aspects importants, la représentation des connaissances et le raisonnement qui peut leur être associé. La partie description permet de décrire et de formaliser le domaine en utilisant la notion de : classes, instances, attributs, relations, fonctions (ensembles de relations : simplification), restrictions (conditions sur certains éléments), etc. La partie raisonnement, quant à elle, décrit les aspects dynamiques liés à une ontologie, à travers des règles (règles mettant en valeur les éléments de l'ontologie), et des événements (événements modifiant certains attributs ou relations), etc.

Il existe plusieurs types d'ontologies. En se basant sur certains travaux effectués dans le domaine de la classification d'ontologie [GómezPérez99], [Guarino97b], [Mizoguchi98], [MizoguchiIkeda96], [VanHeijstAl97], [VanwelkenhuysenAl94], [VanwelkenhuysenAl95], [WieliingaSchreiber93], etc., on peut déduire la classification d'ontologies (dépendante des besoins d'utilisation) suivante :

- L'ontologie de haut niveau : définit et représente les concepts de haut niveau qui sont des concepts de l'univers de modélisation commun aux ontologies des niveaux inférieurs, comme la notion de temps, l'espace, etc. Son rôle est de réduire les ambigüités des termes entre les différentes ontologies et utilisations.

- L'ontologie générique : située d'un point de vue abstraction entre l'ontologie de haut niveau et l'ontologie du domaine. Ses concepts sont moins abstraits que l'ontologie de haut niveau et plus génériques par rapport à ceux de plusieurs ontologies de domaines, réutilisables par plusieurs domaines.

- L'ontologie de domaine : représente et décrit les concepts se rapportant à un domaine et/ou à une utilisation donnés.

- L'ontologie de tâches : représente et décrit le vocabulaire propre à une activité décrivant une structure de résolution d'un problème.

- L'ontologie d'application : représente et décrit les concepts propres à une application. Dans la plupart des cas, ces concepts proviennent de l'ontologie du domaine et de l'ontologie des tâches.

- L'ontologie de représentation : représentent et décrit les primitives des langages de représentation et de formalisation de connaissances : RDF, OWL, réseau sémantique, LD, LPO, etc.

2.1.1. Les langages d'ontologies

Les langages constituent une partie très importante des systèmes à base d'ontologies, car ils permettent la gestion de la base de connaissances et de raisonnements qui peuvent leur être associés. Les langages présentés dans ce paragraphe sont des recommandations du W3C, basés sur des normes, standards, etc. ayant pour but de permettre des traitements automatiques sur des contenus par des machines et applications différentes. Nous verrons un peu loin que ceci, fait de la notion d'ontologie un principe fondamental pour le partage des connaissances, la communication et la coopération au sein des systèmes distribués. Ci-après quelques éléments de description de ces langages, extraits de la pyramide du Web sémantique de Berners-Lee.

URI/IRI : les URIs (Uniform Resource Identifier) permettent d'identifier d'une manière unique les ressources sur le WEB et les IRIs (Internationalized Resource Identifiers) permettent aux personnes d'identifier des ressources Web dans leur propre langue.

XML : Extended Markup Language (XML) est un langage de description et d'échanges de documents et des données ne permettant pas leur présentation, appelé aussi format d'échanges standardisé. Il permet aux applications reconnaissant ce format d'échanger tous les types de données décrits dans ce langage, utilisé souvent pour assurer la compatibilité des données entre les applications hétérogènes. Exemple, on peut décrire la voiture 123 de la marque Renault et de couleur rouge.

<Voiture id='123'>
<Marque>Renault</Marque>
<Couleur>Rouge</Couleur>
</Voiture>

RDF : Resource Description Framework (RDF) est un métalangage qui sert à décrire les ressources, leurs propriétés et les valeurs des propriétés sous forme d'un graphe (ressource, propriété, valeur). Il est considéré comme un modèle standardisé de description des métadonnées qu'on peut définir et associer à des documents ou de description des annotations qu'on peut définir et associer à des éléments d'un contenu. Ces annotations et métadonnées permettent d'associer du sens à des contenus qui peuvent être traitées d'une manière automatique par les agents logiciels. Pour la gestion sémantique des données, les métadonnées RDF peuvent être des informations sémantiques associées à des mots du texte. Ces éléments peuvent être ensuite analysés et interprétés pour l'extraction du sens global. A noter que RDF peut être exprimé dans plusieurs langages, mais c'est XML qui est souvent utilisé, une version XML de RDF appelée RDF/XML a été même créée. L'exemple précédent peut être représenté en RDF/XML comme suivant :

<rdf :RDF>
<rdf :Description about='123'>
<rdf :Property about='Marque'>
Renault
</rdf :Property>
<rdf :Property about='Couleur'>
Rouge
</rdf :Property>
</rdf :Description>
</rdf :RDF>

RDF peut être aussi utilisé pour décrire des situations et des utilisations particulières avec un vocabulaire bien précis en utilisant la notion de RDF Schéma (RDFS). RDFS consiste à adapter RDF à des domaines modélisés particuliers décrivant des utilisations particulières au sein d'une communauté. On peut associer une ontologie de métadonnées ou d'annotations partagées définissant le contexte d'utilisation et échangeables entre les différents agents humains et logiciels. La définition d'un schéma RDF consiste en une activité de typage et de classification des ressources, des propriétés et des relations sous forme de classes définies dans des « espaces de noms » servant principalement à désambigüiser les mêmes éléments d'un vocabulaire définis dans des utilisations ou espaces de noms différents. Ci-après quelques notions définies dans RDFS :

- rdfs : Class : définissant la notion de classe, l équivalent de la notion de concept dans l'ontologie.
- rdfs : subClassOf : définissant la notion de sous-classe d'une classe.
- rdfs : Type : définissant les instances d'une classe.
- Etc.

OWL : est une extension de RDF enrichie avec des propriétés sémantiques, de contraintes, de comparaisons, de cardinalités, etc. pour décrire et manipuler les ontologies. C'est un langage recommandé par le W3C basé comme RDF sur XML qui permet à des moteurs d'inférences d'agents l'interprétation et le raisonnement automatique sur les ontologies. En effet, OWL est basé sur les logiques de description utilisées dans les systèmes de représentation de connaissances et offrant de fortes possibilités de manipulation de prédicats de classes, de rôles et d'individus, donc d'ontologies, contrairement aux logiques de premier ordre classiques ne manipulant que des objets de même type.

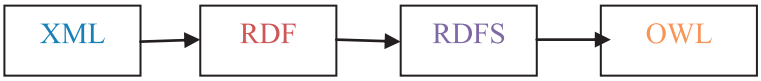


Figure 1. Evolutions des langages vers OWL

Ils existent trois versions d'OWL :

- OWL Full : c'est la version la plus complète, elle combine toutes les primitives d'OWL avec RDF/S sans respecter aucune contrainte si ce n'est celles de RDF, cette version n'est pas décidable.

- OWL DL : c'est un sous ensemble d'OWL Full basé sur la logique de description n'autorisant que des constructions garantissant la décidabilité des inférences.

- OWL Lite : est la version la plus simple, sert principalement à la création de hiérarchies de classes. Elle est dotée seulement de quelques propriétés de classes comme la comparaison, la restriction, la cardinalité (0 ou1).

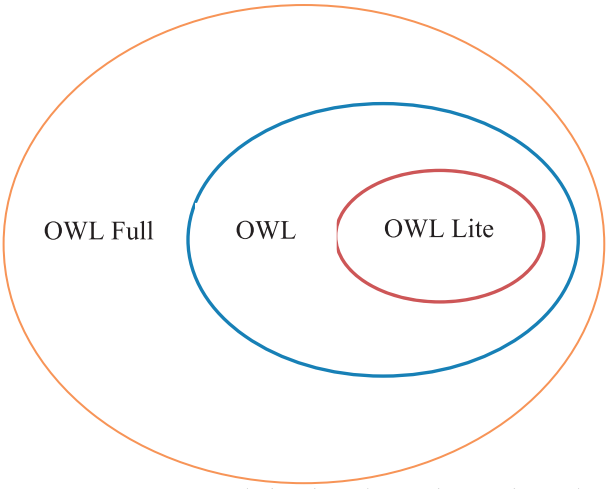


Figure 2. Imbrication des trois versions d'OWL

SPARQL (Protocol and RDF Query Language) : Langage d'interrogation des ontologies représentées sous forme de graphes RDF/S. Il est pour les bases de connaissances RDF ce que SQL est pour les bases de données relationnelles. Exemple :

SELECT ?x, ?y, ?z FROM URI/IRI WHERE {Conditions1, condition2, etc.}

RIF (Rule Interchange Format) : est format d'échange de règles standardisé. Il permet de faciliter l'échange de règles utilisables par des systèmes distribués sur le WEB en assurant l'interopérabilité et la portabilité entre divers langages et moteurs de règles.

SWRL (Semantic Web Rule Language) : est un langage de raisonnement sémantique à base de règles sur les ontologies. C'est une combinaison d'OWL-DL et RuleML language crée principalement pour le Web sémantique pour pouvoir développer des règles sémantiques au niveau des agents. Il permet contrairement à OWL de manipuler les instances par des variables, de définir des fonctions mathématiques, des types de données, etc. nécessaires pour la programmation Orientée agents. Exemple :

avec OWL on ne peut définir la relation oncle que comme suivant :
intersectionOf(SubClassOf(Homme), estfrereDe(Pere)
avec SWRL on peut la définir au niveau des instances représentées par des variables x,y,z comme suivant :
Personne(?x) ^ Personne(?y) ^ Personne(?z) ^ pere(?x,?y) ^ frere(?x,?z) → oncle(?z,?y)

Ce qui permet de définir qui est l'oncle de qui, impossible à définir avec OWL.

SWRL se différencie, en plus, d'OWL par la fait qu'il ne peut pas créer ni de nouveaux concepts ni de nouvelles relations sauf ceux créés par la manipulation des variables et par la satisfaction des règles d'inférences.

2.2. Le WEB sémantique

Le WEB sémantique est une notion très importante de l'Internet moderne, permettant non seulement le stockage et la diffusion des informations, mais également leur compréhension en effectuant des raisonnements sur leurs sens par des machines ou agents logiciels. Le web sémantique est basé sur RDF et OWL pour structurer et annoter des contenus sur lequel on défini et on construit des technologies permettant aux machines d'effectuer des raisonnements et des traitements automatiques (dans certains cas difficiles pour l'homme, par exemple : indexation, compréhension automatique et recherche sémantique) en s'appuyant sur les concepts comme, l'expression du sens, La représentation et gestion des connaissances, Les ontologies, les agents, etc. (Tim Berners-Lee, co-inventeur avec Robert Cailliau du World Wide Web). Ci-après quelques éléments sur les couches constituantes du Web sémantique. Il repose sur ce qu'on appelle la pile du WEB

sémantique représentée dans la pyramide de Berners-Lee, ci-dessous, composée d'une hiérarchie de langages normalisés par le W3C.

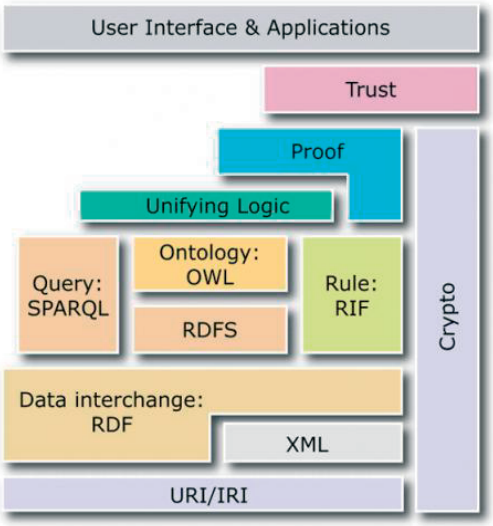


Figure 3. Pyramide du Web sémantique, Berners-Lee

Certains langages de cette pyramide peuvent être regroupés pour former des couches cohérentes correspondantes à des niveaux de traitements et d'utilisations donnés :

- localisation des données, URI/IRI,
- format description des données et de résolution des espaces de noms, XML/S,
- Données ou description des données, RDF,
- Définition et description des schémas et vocabulaires des ontologies : RDFS, OWL, RIF, SPARQ,
- Logique, déduction de nouvelles règles, non prévues au départ, SWRL,
- Preuve, définition des outils de description des étapes du raisonnement logique,
- Authentification, définition des outils et des services d'authentification des données,
- Utilisateur, définition des interfaces et applications utilisateur.

2.3. Agents sémantiques intelligents et systèmes multi-agents

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. D'une manière générale, un agent intelligent est constitué :

- D'une base de connaissance prédéfinie ou base de faits,
- D'un moteur d'inférence, lui permettant d'effectuer des déductions sur sa base de connaissances,
- D'un mécanisme d'apprentissage de connaissances lui permettant d'enrichir en permanence sa base de connaissances.

En plus, dans son fonctionnement, un agent intelligent doit être doté de compétences et d'une représentation de son environnement lui permettant de percevoir son environnement et d'agir en toute autonomie pour faire face à des situations imprévues. Il existe plusieurs types d'agents :

- Réactif qui ne fait que réagir aux stimuli qu'il peut d'une manière mécanique. Le comportement du système émerge des réactions simples des agents,
- cognitif disposant de capacités de raisonnement sur sa représentation du monde où il évolue. L'une des architectures les plus connues pour ce modèle est l'architecture BDI (Belif-Desir-Intention) où les agents sont constitués principalement des modules suivants :
    - 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, ayant des compétences complémentaires et coopérant afin d'atteindre des objectifs partagés. L'une des architectures les plus connues des systèmes multi-agents est l'architecture Système Multi-Agents Adaptatif (AMAS) qui 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 imprévues. 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.

La communication entre agents utilise un ACL (Agent Communication Language) basé sur les actes de langage afin d'échanger, générer, interpréter et séquencer les messages entre des agents intentionnels indépendamment des machines. Les actes de langage représentent les briques de bases dans les interactions afin d'accomplir une conversation régie par un protocole d'interaction. Les protocoles d'interaction représentent des séquences types de messages entre agents permettant :

- Les règles de communication entre agents,
- Les enchaînements conversationnels entre agents,
- De préciser qui peut dire quoi à qui et les réactions possibles.

Le langage le plus connu est l'ACL-FIPA (l'ACL proposé par the Foundation for Intelligent Physical Agents) doté d'une sémantique beaucoup moins ambiguë et utilisant peu de performatifs : Request, Clarify, Confirm, Respond, Accept Proposal, Call for Proposal, Inform, etc. Le format d'une interaction est le suivant :

request : *performatif (énoncé accompagné d'une action) ;*
Sender : *agent émetteur ;*
Receiver : *agent récepteur ;*
Base de connaissances : *définitions partagées* ;
Base d'inférences : *définitions partagées* ;
Parameter : *paramètres d'appel;*
Language : *langages utilisés*
Content : *contenu du message ;*

Avec les récentes évolutions du WEB, une nouvelle génération d'agents commence à émerger, appelés : « agents BDI sémantiques ». Ils permettent une implémentation des systèmes BDI dans les nouveaux systèmes du WEB sémantique. Ce sont des agents BDI classiques qui utilisent dans leur modèle de conception, à savoir la représentation de leurs connaissances internes, leurs systèmes d'inférences ou encore la représentation du monde où ils évoluent et du système de communication et de coopération avec d'autres agents, des technologies de gestion de la sémantique basées principalement sur les ontologies du WEB sémantique. C'est une évolution majeure des systèmes BDI au service du WEB moderne, sémantique et intelligent combinant les caractéristiques d'un agent BDI et les techniques du WEB sémantique. Pour leur fonctionnement, on peut par exemple utiliser OWL pour la représentation de la base de connaissances et SWRL pour la représentation des inférences. OWL et SWRL peuvent être aussi encapsulés dans un langage ACL utilisé au sein d'un système multi-agent pour assurer la partie communication et coopération entre les agents.

2.4. WEB 2.0

Le web 2.0 peut être défini comme un ensemble d'outils, normes, protocoles, etc. 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 entre les applications par échanges de données basées sur des protocoles normalisés : Web services, les flux RSS, OAI, etc. 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,
- la création des systèmes émergents dont le comportement global est la résultante des comportements élémentaires des services constituants,
- etc.

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 profitables aux prochaines versions du WEB.

D'une manière générale, une application de type Web 2.0 est 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 normalisés comme 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) et aussi pour permettre les échanges entre les applications. Les solutions utilisées sont souvent légères, moins couteuses et accessibles à tous. Leur utilisation ne nécessite pas de formation particulière. Nous proposons ci-après un résumé des fonctionnalités que

pourraient offrir les applications de type WEB 2.0, sous forme d'une liste non exhaustive suivante. Ils permettent :

- 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 de ces outils comme les Wikis, 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,
- 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,
- etc.

3. WEB 3.0

3.1. Définition

Le WEB 3.0 est le nom de la prochaine version de l'internet. C'est un concept émergeant et qui s'articule autour des ontologies, de l'intelligence artificielle distribuée, du WEB 2.0 et du WEB sémantique. Il est considéré comme l'internet 3ème génération, version évoluée du WEB ou WEB intelligent. Il aura pour mission de faire cohabiter les technologies évolutives du WEB et celles de l'Intelligence Artificielle Distribuée (IAD) en utilisant un ensemble, d'outils,

protocoles, normes, standards, etc. permettant à des machines ou à des « agents web intelligents » ou « agents sémantiques » d'effectuer, des raisonnements et des traitements en ligne d'une manière coopérative et automatique sur des contenus WEB. Ce principe 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 totale du réseau mondiale. En effet 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, que les problèmes de saturation des traitements peuvent être résolus, concernant la prise en charge de la totalité des données présentes sur Internet, constituant ainsi la base de connaissances sur laquelle s'effectuent les raisonnements. Ceci pourrait être rendu possible en structurant et en annotant intelligemment les données par l'utilisation des technologies issues du web sémantique et de la représentation du sens (sémantique), qui intègrent nativement des technologies avancées basées sur les ontologies et de l'analyse et de la compréhension automatique du contenu telles que : xml, métadonnées, RDF/S, OWL, SWRL, etc. et permettre ainsi à des outils intelligents de prendre en charge beaucoup de tâches automatisables et difficiles à effectuer par un utilisateur humain.

3.2. Une 1ère vue de l'architecture du WEB 3.0

Cette version du Web repose essentiellement sur la fusion des évolutions du WEB 2.0 et du WEB sémantique, dont l'intelligence correspond à la résultante de l'intelligence collective humaine représentée dans les systèmes actuels, interactifs et coopératifs comme les réseaux sociaux et de l'intelligence artificielle gérée par les machines et représentée dans le WEB sémantique et la notion d'agents intelligents.

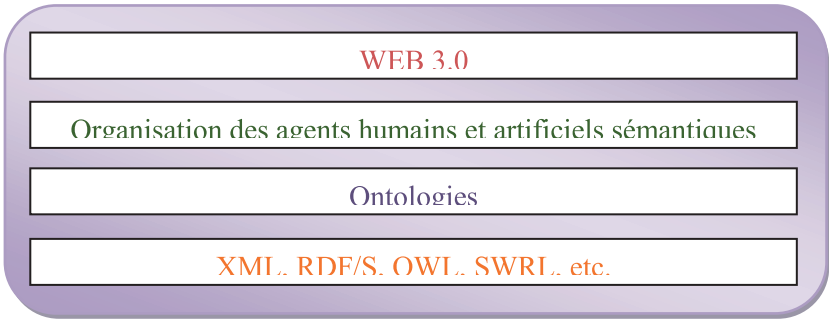


Figure 4. Une 1ère vue de l'architecture du WEB 3.0

Le WEB 3.0 est une agrégation d'agents humains et logiciels créateurs et gestionnaires de contenus. Ces agents reposent sur tous les langages de représentation, d'annotation, de traitements et d'échanges pour effectuer leurs tâches dans un contexte coopératif. Toute cette mécanique est basée sur une notion essentielle, incontournable qui est la notion d'ontologies. La définition de tous les concepts mis en place, les langages, les agents, les systèmes de coopération, etc. repose tous sur les différentes variantes et langages d'ontologies.

Les aspects liés à l'interactivité et à la coopération feront toujours partie du WEB futur. 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. Apport des ontologies pour le WEB 3.0

Dans ce qui suit, nous décrivons le rôle important joué par les ontologies dans les divers outils et technologies constituants le WEB sémantique, les WEB 2.0, des systèmes multi-agents intelligents et donc du WEB 3.0 pour la création, le partage et la diffusion des contenus ainsi que pour le fonctionnement de tels systèmes.

Dans le cas du WEB 2.0, la création et la diffusion des contenus se fait par intéressement et constitution de groupes et de communautés d'intérêts communs. Dans le cas du WEB sémantique et des systèmes multi-agents, la création et la diffusion des contenus se fait par des machines ou des agents intelligents permettant de générer et de produire de grandes quantités de contenus d'une façon automatique. Dans ces deux cas, la notion d'ontologie est nécessaire, car elle permet des traitements sur des connaissances partagées entre les différents agents. En plus, cette notion continue d'évoluer suivant les évolutions des différents composants du WEB 3.0 et devient donc fondamentales et incontournables dans la définition de la structure et de l'architecture des nouveaux systèmes, au niveau de la modélisation, conceptualisation et représentation consensuelles des connaissances partagées, mais aussi au niveau du raisonnement, actions et interactions entre les différents agents humains et/ou logiciels. Cette évolution d'ontologies prend en compte aussi l'expression et la gestion de la sémantique dans un contexte multi-agents que ce soit au niveau de la représentation de

connaissances qu'au niveau raisonnement et coopération des agents dans leur environnement.

4.1. Niveau représentation de connaissances

Au niveau de la représentation de connaissances, on a montré tout au long de cet article que cette notion est importante et constitue une couche au dessus des langages de représentations des connaissances capables de prendre en charge les aspects liés à la représentation des connaissances sémantiques dans un contexte distribué (cf. paragraphe 4.4).

4.2. Niveau inférences

De nouveaux langages toujours basés sur les ontologies sont mis en place pour assurer tous les aspects liés aux raisonnements, déductions, etc. Ce qui augmentera de l'importance des ontologies dans les futures versions du WEB. Les caractéristiques des agents peuvent être formalisées avec OWL et les ontologies, la partie raisonnement avec SWRL et les règles d'interactions.

4.3. Niveau communication entre agents

Les échanges entre agents peuvent se faire par l'intégration des ontologies dans les actes de communication. Comme mentionné précédemment, le but principal des ontologies est de permettre aux différents agents (humains et logiciels) de parler un même langage par rapport à un domaine donné. Les agents peuvent alors utiliser les mêmes formalismes pour représenter leurs concepts qu'ils peuvent percevoir de la même manière et éviter ainsi toute ambigüité au niveau de leur interprétation. Dans le cas des systèmes multi-agents classiques, les langages adaptés pour décrire les bases de connaissances et les inférences sont des langages issus de la Logique du Premier Ordre. Les langages de représentation et d'inférences d'ontologies sont aux aussi issus de ces mêmes types de logiques. Alors les langages de type OWL et SWRL peut être utilisée au sein du protocole de communication inter-agents pour contribuer à une bonne expression de la sémantique et un bon partage de connaissances au sein d'une communauté d'agents.

4.4. Agents BDI sémantiques et ontologies

Pour récapituler et se rendre compte de l'importance d'ontologies, on peut se baser sur cette présentation type de la structure d'un agent DBI sémantique. La notion d'ontologie peut être introduite dans les différents niveaux : définitions de la structure et du comportement d'un tel type d'agents.

Agent : *description de l'agent participant*
Base de connaissance : *représente les informations sur l'état de l'agent et sur son environnement exprimées en OWL, SWRL, etc. ;*
Actions : *actions/instructions implémentées dans l'agent ;*
Interface : *ensemble de services exposés et services exigés par l'agent;*

Comportement : *explicite les comportements possibles de l'agent ;*
Plan(1,*) : *défini et séquence l'ensemble des action et des services*
mis en jeu pour atteindre un but ;
Nom : *nom et identité du plan ;*
Contexte : *contexte du déclenchement ;*
Service(1,*) : **Request** : *performatif (énoncé accompagné d'une action) ;*
Sender : *agent émetteur ;*
Receiver : *agent récepteur ;*
Ontology : *ontologie exprimée en OWL, SWRL, etc.* ;
Parameter : *paramètres d'appel;*
Language : ACL (encapsulation des langages d'ontologies : OWL, SWRL, etc.) ;
Content : *contenu du message peut être exprimé en langages Ontologiques (RDF/S, OWL, SWRL, etc.) ;*

4.5. Une 2ème vue de l'architecture WEB 3.0

Le WEB 3.0 peut être vu comme une agrégation du WEB 2.0, le WEB sémantique, des agents intelligents et les technologies développées autour des ontologies. Les réseaux sociaux actuels peuvent être augmentés d'agents artificiels formant des communautés virtuelles interactives pouvant partager les mêmes intérêts et les mêmes données représentées sous forme d'ontologies pour assurer leur communication et leur coopération.

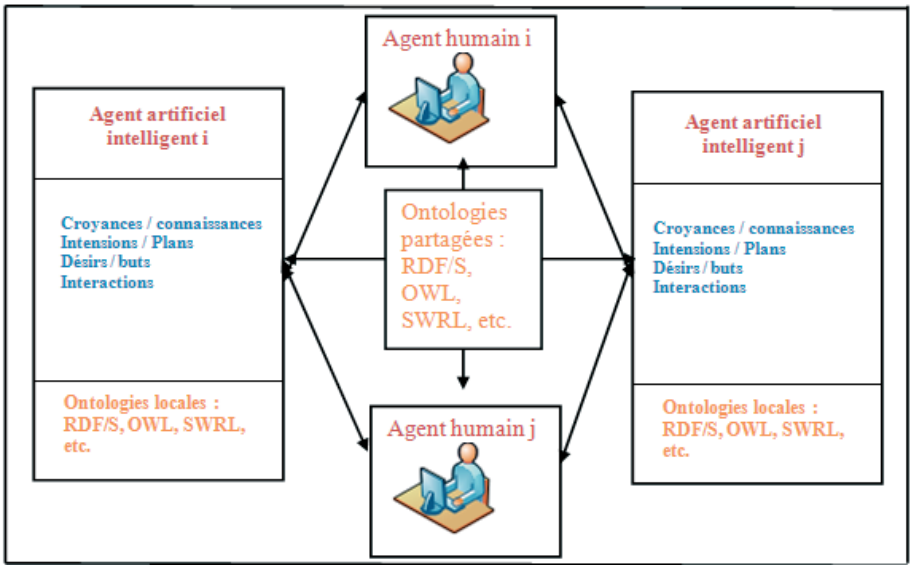


Figure 5. Une deuxième vue de l'architecture du WEB 3.0

Le WEB 3.0 est un ensemble de communautés d'agents humains et artificiels, constituées autour des connaissances ontologiques partagées.

5. Conclusion

Les ontologies constituent la base de tout système prenant en compte les aspects liés à la gestion des connaissances sémantiques dans des environnements hétérogènes où les agents humains et logiciels doivent en permanence interagir et adapter leur comportement face à des situations données pour interpréter des données par rapport à leur contexte. On peut déduire à travers les différentes analyses, que les ontologies constituent la brique de base de tout système hétérogène, coopératif et physiquement distribué : elles permettent une représentation des connaissances partagées et consensuelles permettant à ces systèmes de parler le même langage et donc de mieux communiquer et de coopérer. Elles permettent également l'interopérabilité entre les systèmes du fait qu'elles sont basées sur des langages de représentation et de raisonnement normalisés. En conclusion, les ontologies sont une construction très importante pour le WEB 3.0, sémantique, intelligent, interactif et multi-agents humains et artificiels.

Références

[1] A core ontology for requirements I. Jureta, J. Mylopoulos, S. Faulkner Applied Ontology 4(3-4):169-244, 2009.

[2] Bachimont B., Modélisation linguistique et modélisation logique : l'apport de l'ontologie formelle Actes de la Conférence Ingénierie des Connaissances (IC'2001), Grenoble, p349-351, Grenoble, Juin 2001.

[3] BERNERS-LEE Tim, HENDLER James and LASILLA Ora, The Semantic Web, Scientific American, May 2001.

[4] Brickley, D., Guha, R.V., Resource Description Framework (RDF) Schema Specification 1.0 World Wide Web Consortium, http://www.w3.org/TR/rdfschema/

[5] Cutkosky, M., Engelmore, R. S., Fikes, R. E., Gruber, T. R., Genesereth, M. R., Mark, W. S., Tenenbaum, J. M., & Weber, J. C. (1993). PACT: An experiment in integrating concurrent engineering systems. IEEE Computer, 26(1), 28-37.

[6] Fensel, Van Harmelen, Horrocks, McGuinness, Patel-Schneider. OIL: An ontology infrastructure for the semantic web. IEEE Intelligent Systems, 16(2):38-45, 2001.

[7] Gruber, T. R. (1991). The Role of Common Ontology in Achieving Sharable, Reusable Knowledge Bases. In J. A. Allen, R. Fikes, & E. Sandewall (Eds.), Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference, Cambridge, MA, pages 601-602, Morgan Kaufmann.

[8] JACK Intelligent Agents. http://www.agent-software.com/.

[9] M . E. Bratman. Intention, Plans and Practical Reason. Harvard University Press, 1987.

[10] Techne: Towards a New Generation of Requirements Modeling Languages with Goals, Preferences, and Inconsistency Handling I. Jureta, A. Borgida, N. Ernst, J. Mylopoulos IEEE International Requirements Engineering Conference (RE), 2010.

Sur la constitution de corpus de deux langues à tradition orale (le berbère tamazight et l'arabe marocain) parlées à Orléans

Samira MOUKRIM
LLL-Université d'Orléans
samiramoukrim@yahoo.fr

Résumé

Pour étudier l'expression du « présent » en français, berbère tamazight et arabe marocain, nous avons fait le choix de tavailler sur des données orales authentiques. Les données du français ont été extraites du corpus ESLO (Enquêtes Socio-Linguistiques à Orléans), piloté par le Laboratoire Ligérien de Linguistique de l'Université d'Orléans. Quant à celles de l'arabe marocain et du berbère tamazight, nous avons constitué nous-mêmes le corpus auprès de locuteurs marocains arabophones et berbérophones résidant à Orléans. Notre objectif était de recueillir un échantillon authentique de ces deux langues qui, depuis la signature de la Charte européenne des langues régionales et minoritaires en 1999, figurent parmi les 'langues de France non territoriales'.

Nous avons constitué un corpus de « données situées » : il contient, en plus des données primaires (les enregistrements de la parole), une riche documentation sur ces données et sur leur contexte de production. Nous avons tenu également à expliciter notre démarche, à documenter les conditions de constitution du corpus ainsi que nos choix théoriques et techniques. Toutes ces informations sont d'une grande importance pour rendre le corpus disponible.

L'hétérogénéité des trois corpus du français, de l'arabe marocain et du berbère tamazight se manifeste tant au niveau de la *langue*, différente d'un corpus à l'autre, qu'au niveau de sa *codification* graphique et du degré de son informatisation. Si le français dispose d'une tradition écrite ancienne et bien établie (standard), ce n'est le cas ni du berbère ni de l'arabe marocain. Le berbère est une langue à tradition orale, sans tradition écrite standard ni dialecte de référence. L'arabe marocain est une langue à tradition orale, pour laquelle il existe certes un standard, l'arabe classique, auquel il est apparenté, mais dont le système linguistique est très différent.

Nous avons essayé de voir en quoi la différence des corpus implique une spécification des méthodes et outils, en particulier si les outils et instruments

utilisés pour les corpus oraux de langues à grande diffusion (notamment celui d'ESLO) s'ajustent aux langues à tradition orale. Dans ce papier, il sera question des différentes étapes de la constitution des corpus de l'arabe marocain et du berbère tamazight, des choix opérés et de la démarche adoptée sur le terrain. Seront abordés également les aspects juridiques du corpus.

Mots clés :

constitution de corpus, langues à tradition orale, données situées, procédure de transcription spécifique, statut juridique du corpus

Introduction

Pour étudier l'expression du « présent » en français, berbère tamazight et arabe marocain, nous avons fait le choix de tavailler sur des données orales authentiques. C'est la diversité d'emplois de la forme du « *présent »* en français (présent actuel, habituel, gnomique, de vérité générale, narratif, historique, etc.)[1] qui a suscité notre intérêt et nous a poussé à explorer la question dans d'autres langues, notamment en arabe marocain et en berbère tamazight [2]. Nous nous sommes interrogée sur l'existence et l'expression de ce phénomène dans deux langues dont les systèmes verbaux reposent sur une opposition purement aspectuelle.

[1] Le présent de l'indicatif en français continue à poser problème malgré les nombreuses mises au point dont il a fait l'objet. Le débat sur ses valeurs et ses emplois est loin d'être clos. Les linguistes proposent des explications tout aussi variées selon la théorie adoptée. Pour les uns (Damourette et Pichon (1911-1926/1970), Mellet (1980-2000), Serbat (1980-1988), Chuquet (1994)), le présent est une forme neutre, capable d'inscrire le procès dans n'importe quelle époque. Pour d'autres (Benveniste 1970/ 1974, Guillaume 1929, Gosselin (2005), Haillet (2005)), le présent est une forme temporelle déictique, marque de coïncidence entre le moment d'énonciation et le moment du procès. Une troisième hypothèse est avancée par Wilmet (1997 : 341) qui définit le présent comme étant une « forme verbale qui affirme la concomitance d'un procès au repère de l'actualité ». Avis partagé par Beauzée (1782-1786/1986), Jaubert (2001), Revaz (2006). Cf. Moukrim (2010 : 295-309) pour plus de détails sur la diversité des traitements des emplois du présent de l'indicatif en français.

[2] L'étude de ce phénomène dans des langues différentes constitue un enjeu théorique pour la linguistique. Comme le précisent Fuchs & Robert (1997 : 1) : « il n'est sans doute pas exagéré de dire que *la* question centrale pour la linguistique est celle de la diversité des langues : c'est en effet à partir des langues –au pluriel– que la linguistique tente d'appréhender le langage –au singulier ». Dans ce travail, l'enjeu est de taille car les systèmes verbaux des langues étudiées sont très différents.

Pourquoi une énième étude sur le « présent de l'indicatif » en français ? Pour deux raisons : i) Explorer la *démarche comparative* ii) travailler sur des *données orales authentiques*. Nous supposons que la difficulté de rendre compte de la multiplicité d'emplois de la forme du « présent » en français pourrait être élucidée en examinant le fonctionnement de formes équivalentes dans d'autres langues. Nous supposons également que l'observation de la langue parlée pourrait apporter du nouveau sur la question, car la plupart des études fabriquent des exemples, ou empruntent des exemples à l'écrit[3]. Ce travail consistait donc en deux volets i) la constitution du corpus et ii) l'observation des données.

Dans ce papier, il sera question des différentes étapes de la constitution des corpus de l'arabe marocain (désormais AM) et du berbère tamazight (désormais BT), des choix opérés et de la démarche adoptée sur le terrain. Seront abordés également les aspects juridiques du corpus.

1. Contexte

Pour les trois langues à l'étude (AM, BT et français (désormais FR), notre corpus a été constitué à Orléans. Cette *unité* de lieu (Orléans) se justifie d'abord par des raisons pratiques : nous travaillons dans le Laboratoire Ligérien de Linguistique (LLL-Université d'Orléans et Tours), qui pilote le projet Enquêtes Socio-Linguistiques à Orléans (ESLO)[4], l'un des corpus les plus importants du français oral. C'est dans ce dernier que nous avons puisé les données du français.

Quant à celles de l'arabe marocain et du berbère tamazight[5], nous avons constitué le corpus auprès de locuteurs marocains arabophones et berbérophones résidant à Orléans. Notre objectif était de recueillir un échantillon authentique de ces deux langues pour les comparer avec le français parlé à Orléans. Nous avons voulu parallèlement à travers cette étude dessiner une image de l'arabe marocain et du

[3] Une étude basée sur des données empiriques orales permet non seulement d'observer les emplois de la forme du « présent » qui sont en rapport direct avec le moment de la parole mais aussi de déterminer les paramètres qui font que cette même forme énoncée dans l'actuel puisse renvoyer au non-actuel.

[4] cf. site d'ESLO : http://www.univ-orleans.fr/eslo pour plus de détails sur le programme ESLO.

[5] Il s'agit du berbère parlé par des témoins berbérophones originaires du Maroc central (dialecte *tamazight) et de* l'*arabe marocain* parlé par des témoins arabophones originaires de différentes villes du Maroc (Casablanca, Rabat, Berkane, Khémisat, Meknes, Khénifra…)

berbère tamazight parlés actuellement en France, à Orléans. Ces deux langues[6] qui, depuis la signature de la Charte européenne des langues régionales et minoritaires en 1999, figurent parmi les 'langues de France non territoriales'[7].

Cette étude s'inscrit également dans le cadre du programme Langues en Contact à Orléans (LCO), module d'ESLO, qui étudie la vie des langues en contact avec le français, d'où le choix des thématiques proposées (langues parlées / en contact, représentation des langues, culture, traditions...).

2. Méthodologie

Notre corpus a été constitué en trois étapes[8], qui représentent autant de moments de choix méthodologiques :

a. Le travail préparatoire à la phase d'enregistrement ;

b. L'enregistrement des données ;

c. La mise en forme des données à des fins d'exploitation (transcription et annotation)

Pour la constitution du corpus, nous avons pris en compte un certain nombre de critères pertinents pour définir la sélection des données et décider des types d'enregistrements, au nombre desquels : i) les objectifs du travail, sa finalité ii) le type de données requis pour satisfaire ces objectifs iii) les modalités de collecte.

Le corpus constitué a pour fin une étude des formes verbales du « présent » en AM et en BT parlés à Orléans. Pour répondre aux exigences de cette recherche, le corpus doit permettre d'attester les formes verbales du présent dans leurs différents usages. Se pose alors le problème de savoir quelles sont les données qu'on peut considérer comme représentatives de ce phénomène ?

La représentativité du corpus étant difficillement envisageable[9], nous avons essayé de l'améliorer, comme le précise à juste titre Habert (2000) : « Améliorer la

---

[6] Pour plus de détails sur la situation de l'arabe dialectal et du berbère en France, cf. Moukrim (2010 : 42-55).

[7] A signaler que la Charte des langues régionales et minoritaires, bien que signée le 7 mai 1999, n'a pas encore été ratifiée par la France. La Charte étant contraire à la Constitution, la France ne peut la ratifier sans engager une révision constitutionnelle qui rendrait les deux textes compatibles.

[8] Delais-Roussarie, E. (2003 : 91-92) a présenté les principales étapes de la constitution de corpus oraux en prenant appui sur la littérature traitant des problèmes de constitution des corpus écrits (cf. entre autres, Habert *et al.* 1997a et b).

représentativité d'un corpus consiste à préciser la production et la réception de chacun de ses composants, en lien avec les motifs qui ont conduit à la création du corpus, mais aussi à pouvoir déterminer sur des bases objectivales les différents emplois du langage auquel on s'intéresse ».

Dans ce travail, nous avons essayé de diversifier les situations enregistrées ainsi que les catégories de locuteurs en différenciant sociologiquement les témoins par l'âge, le sexe, le niveau scolaire, la profession et les langues parlées. Nous avons tenu également à documenter les données ainsi que leur contexte de production.

Partant du fait que les données recueillies sont toujours l'effet d'un conditionnement consécutif au procédé d'enquête, les modes de recueil de notre corpus ont été conduits dans l'esprit de ce que souligne Maurer B. (1999) : « Le choix du mode de recueil des données constitue même une étape fondamentale dans la démarche de recherche puisque l'objet effectivement étudié en dépend étroitement ».

Pour réunir le plus de données possible, nous avons eu recours à l'entretien semi-directif (face à face). Un guide d'entretien a été réalisé afin de faire parler les témoins, en ciblant les contextes propices à l'émergence des formes verbales au présent. Les questions que nous avons choisies portent d'une part, sur les langues utilisées par nos informateurs à Orléans, sur leur importance et sur ce qu'elles représentent pour eux ; d'autre part, sur la culture et les traditions transplantées du pays d'origine au pays d'accueil. L'entretien constitue une incitation à des réflexions sur la situation présente, passée et sur les perspectives. Ce qui va nous permettre de cerner le fonctionnement de la forme du présent sans pour autant nous restreindre à une époque (présente, passée ou future) particulière.

Et pour varier les situations enregistrées, figurent dans le corpus d'autres genres de parole comme les communications téléphoniques, les recettes de cuisine, le récit et le récit de vie, la conversation et le commentaire de photos.

	Corpus de l'arabe marocain (2008-2009)	**Corpus du berbère tamazight** (2008-2009)
Nombre d'heures	**7- 8 h**	**7- 8 h**

[9] Pour A. Mettouchi et A. Larcheret-Dujour (2006), un corpus de référence pour les langues à petite diffusion peut « se définir comme une base contenant non pas tous les genres ou types de données ou tous les types de locuteurs, ni tous les points d'enquête pour une langue donnée, mais comme un ensemble structuré et mutualisable de textes rattachés à leur enregistrement sonore, accompagné par des informations linguistiques pertinentes pour la langue en question, ainsi que par des métadonnées riches » (cf. http://crdo.risc.cnrs.fr/ecoles/elco/bilan-ELCO.pdf)

d'enregistrement		
<u>Situations de parole</u>	- Entretien face à face - recettes de cuisine - communications téléphoniques - récit/ récit de vie - conversation - commentaire photos	- Entretien face à face - recettes de cuisine - communications téléphoniques - récit/ récit de vie - conversation - commentaire photos
<u>Catégories des locuteurs :</u> - Age : - Sexe : - Niveau scolaire : - Profession/CSP :	- 10-20 ans - 20-30 - 30-50 - 50 et plus - Hommes: 36% - Femmes: 64% **Niveau d'études** - non scolarisé - primaire/collège - bac - supérieur **(professions)** - Directeur - Informaticien - Préparateur en pharmacie -Technicien -Commerçant/ Aide commerçant - Médiateur - ouvrier - Couturière - Femme de ménage - Etudiant/élève - Sans profession	- 20-30 ans - 30-50 - 50 et plus - Hommes: 40% - Femmes: 60 % **Niveau d'études** - non scolarisé - primaire/collège - bac - supérieur **(professions)** - Secrétaire - Préparateur en pharmacie - technicien - Commerçant/ Aide commerçant - ouvrier - Couturière - Sans profession

Concernant les langues parlées par nos témoins, nous sommes partie de leurs déclarations en sorte qu'il est difficile, à partir de cette enquête, de mesurer le degré de compétence réel. Deux témoins peuvent être considérés comme monolingues, sept comme bilingues et dix-huit trilingues ou plus :

3. L'enregistrement

La qualité de la constitution du corpus est un facteur de la relation de confiance établie entre nous, en tant que personne (de la même communauté

linguistique) et en tant que chercheur (qui peut, du point de vue des locuteurs, contribuer à la promotion de leur langue), et les enquêtés.

Avant de commencer l'enregistrement, nous essayons de nous familiariser[10] avec le témoin pour qu'il se sente parfaitement à l'aise et que la conversation enregistrée soit aussi naturelle que possible. En effet, nous avons pris systématiquement le temps de discuter avec le témoin pour avoir sa confiance. Tout cela a l'avantage d'atténuer « le paradoxe de l'observateur » et d'accéder au vernaculaire du locuteur. Comme le précise Bourdieu : « La proximité sociale et la familiarité assurent en effet deux des conditions principales d'une communication 'non violente' ». Bourdieu (1993 : 1395).

Notre relation privilégiée avec les témoins vient du fait que nous partageons avec eux un certain nombre de traits : même origine (pays d'origine), même langue(s), même situation en France (appartenance à la population immigrée), etc.). Nous sommes membre à part entière de la culture vernaculaire. Nous la connaissons de l'intérieur et nous la comprenons profondément. Cela permet, selon Labov (1978 : 10) de « réussir une pércée plus profonde ».

Les enregistrements ont été recueillis à Orléans entre 2008 et 2009. 8 enregistrements ont été effectués à ACM Formation[11] et le reste (38 enregistrements) chez les témoins, soit 46 enregistrements (audio) au total, d'une durée variant de 5 minutes (recettes de cuisine/communications téléphoniques) à 45 minutes[12].

---

[10] Les problèmes de proximité/distance du chercheur aux situations enquêtées « ont été traités en termes de paradoxe d'observateur-selon lequel le phénomène enquêté se dissout dès qu'il est observé (tel le vernaculaire pour Labov 1972)- aussi bien qu'en termes de violence symbolique entre l'enquêté et l'enquêteur (Bourdieu, 1993). Ils ont aussi été traités en termes de réflexivité- par des chercheurs intégrant leur présence et celle du dispositif d'enquête dans l'analyse de l'objet enquêté (en anthropologie notamment, Clifford & Marcus, 1986, Mondada, 1998) » (Baude 2006 : 52).

[11] ACM Formation (Accès aux Clés de la Modernité) est une association 1901. Son objet est d'intervenir, notamment auprès des publics les plus socialement défavorisés, au travers d'actions de formation et de pratiques culturelles.

[12] Qu'il s'agisse du français extrait du corpus ESLO1, de l'arabe marocain ou du berbère tamazight, chaque corpus représente un peu plus de sept heures d'enregistrement, soit environ vingt-deux heures au total. Une différence d'importance est à signaler : le corpus du français, tiré de l'ESLO, date de 1968-1970, alors que ceux de l'AM et du BT ont été constitués en 2008-2009 à Orléans. Le corpus du français analysé présente un seul genre de parole, des entretiens en face à face.

4. Des données situées

Depuis le développement de la linguistique du corpus, la documentation de ce dernier est devenue fondamentale. Celle-ci consiste à fournir des renseignements sur la situation de collecte et le profil des témoins. Dans ce travail, il a été question de l'observation et la description des usages authentiques[13] de la langue. D'où l'introduction du locuteur réel, d'une part, en tant que « voix » indissociable de la transcription (transcription alignée au son) et d'autre part, par la reconstitution de son profil sociologique. Ce qui peut « rendre à la linguistique la méthodologie d'une véritable science des *données attestées et situées*[14] » (Abouda & Baude 2006 : 9).

Nous avons constitué un corpus de « données situées » : il contient, en plus des données primaires (les enregistrements de la parole), une riche documentation sur ces données et sur leur contexte de production[15]. Ce qui n'est pas sans importance pour l'analyse du phénomène étudié : la plupart de nos résultats émergent grâce à ces données situées. Nous avons tenu également à expliciter notre démarche, à documenter les conditions de constitution du corpus ainsi que nos choix théoriques et techniques. Toutes ces informations sont d'une grande importance pour rendre le corpus disponible[16].

5. La transcription du corpus

La collecte des données de l'arabe marocain et du berbère tamazight s'est effectuée sous forme d'enregistrements. Ces données sonores brutes ne peuvent pas être analysées sans un travail préalable de transcription et de segmentation. Transcrire

---

[13] Concernant la question de *données authentiques*, quelques restrictions sont à signaler : *i)* le corpus est lui-même un *construit* car il résulte d'une sélection : « un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extra-linguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue. » (Habert, B. 2000) ; *ii)* le corpus n'est pas représentatif des différents usages de la langue.

[14] C'est nous qui soulignons.

[15] « La linguistique du corpus prend sens dans la réintroduction de la question de l'usage, elle amène à *situer*, c'est-à-dire à replacer les phénomènes observés et décrits dans un contexte. » (Jacques 2005 : 29).

[16] Pour chaque locuteur, il y a une fiche d'information récapitulant l'âge, le sexe, le niveau scolaire… complétée par des indications sur l'enregistrement (n°, type (situation de parole), participant(s), lieu, date et durée de l'enregistrement…), situation d'enregistrement...

des données sonores consiste à fournir une représentation symbolique du signal[17]. Mais la transcription en elle-même n'est pas une tâche anodine. Le processus de transcription soulève un certain nombre d'enjeux théoriques, méthodologiques et pratiques.

Transcrire des langues à tradition essentiellement orale soulève des problèmes spécifiques, surtout si l'on entend produire un corpus partageable. En effet, en plus des difficultés que pose toute transcription de l'oral, d'autres interrogations, d'une autre nature, surgissent lors de la phase préparatoire à la transcription de ces deux langues 'non latines' et pour lesquelles il existe une concurrence entre plusieurs alphabets et donc plusieurs traditions orthographiques, des problèmes liés à la fois au système graphique, au mode, aux conventions et aux outils de transcription à adopter...

Si le français dispose d'une tradition écrite ancienne et bien établie (standard), ce n'est le cas ni du berbère ni de l'arabe marocain. Le berbère est une langue à tradition orale, sans tradition écrite standardisée. L'arabe marocain est une langue à tradition orale, pour laquelle il existe certes un standard de référence, l'arabe classique, avec laquelle il a un lien de parenté, mais dont le système linguistique est très différent[18].

5.1. Système(s) de notation

Plusieurs systèmes graphiques sont en usage pour écrire la langue berbère : graphie tifinaghe[19], graphie arabe[20], graphie latine et graphie hébraïque[21]. Il y a eu deux tentatives de normalisation du système graphique du berbère. La première est

---

[17] Delais-Roussarie, E. (2003 : 98)

[18] En se basant sur le fait que les systèmes linguistiques de l'arabe marocain et de l'arabe classique présentent des différences sur les plans phonique, morphosyntaxique, voire lexical, et en tenant compte de l'absence de compréhension immédiate, les linguistes leur accordent le statut de deux langues différentes.

[19] La majorité des berbérisants postule que les tifinaghe sont d'origine phénicienne. Ainsi pour Chaker (1984), « l'alphabet est très certainement d'origine phénicienne - punique, comme la quasi totalité des systèmes existants ». Pour Hachid (2001), *« Cette écriture est soit un emprunt à l'alphabet phénicien, soit une invention locale, ou encore un emprunt à un prototype fort ancien que l'on ne connaît pas encore.»* (cité dans Bouhjar (2003).

[20] L'alphabet arabe a été utilisé en particulier chez les Chleuhs (sud du Maroc). Cette pratique est encore très vivante chez les écrivains contemporains chleuhs dont la majorité utilise l'alphabet arabe pour écrire leurs oeuvres (poésie, nouvelle, manuels grammaticaux...) : Moustaoui, Akhiat... Et aussi chez Chafik.

[21] Les communautés juives ont employé l'alphabet hébreu pour transcrire le berbère comme en témoigne une version de la Haggadah de Pessah (Cf. Galand-Pernet et H. Zafrani, une version berbère de la Haggâdâh de Pesah : texte de Tinrhir du Todrha (Maroc), Paris, 2. Vol., 1970 (Comptes rendus du GLECS, Supplément I).

pilotée depuis 2001 par une institution publique au Maroc, l'Institut Royal de la Culture Amazighe (IRCAM) qui a adopté un système graphique à base tifinaghe. La deuxième a commencé au début du XXe siècle, en dehors de toute intervention institutionnelle ou étatique, par des universitaires et militants associatifs, avec des publications écrites en caractère latin, et se poursuit jusqu'à présent à travers des ateliers et journées d'études organisés par les chercheurs[22].

Quant à l'arabe dialectal, il a été noté à l'aide de plusieurs systèmes graphiques. Avant la colonisation française, les musulmans ont utilisé la *graphie arabe* pour noter les parlers arabes maghrébins, et les juifs du Maghreb, la *graphie hébraïque*[23]. Vers la fin du XIXe et au début du XXe siècle, la *graphie latine* a été introduite par des linguistes comme William Marçais, Marcel Cohen, Georges S. Colin, Jean Cantineau, etc. qui ont publié des textes en appui de leurs descriptions de l'arabe dialectal.

Le choix de la graphie latine pour transcrire les corpus de l'arabe marocain et du berbère tamazight suppose la réalisation des phénomènes spécifiques à ces deux langues. Quel mode de transcription choisir ? Quelles conventions de transcription adopter et pourquoi ?

5.2. Mode de transcription du corpus

Choisir un mode de transcription implique des enjeux qui dépendent des objectifs de la recherche, mais également de la représentation[24] de la langue parlée, ou encore de la représentation que l'on veut en donner (Bilger, 2008 : 248). Et comme le souligne (Gadet 2008 : 37), « la transcription ne peut être regardée comme une opération banale, car on transcrit pour donner à voir quelque chose ».

Pour la transcription de l'arabe dialectal et du berbère, il existe plusieurs modes de notation : phonétique, phonologique, morphophonologique et orthographique (notation usuelle). C'est ce dernier mode de transcription (la notation usuelle) qui a été adopté pour la transcription du corpus. La *notation usuelle* n'est ni une transcription phonétique, ni une transcription phonologique stricte, mais une notation d'inspiration phonologique qui prend en compte la structure morpho-syntaxique des énoncés.

---

[22] Ces deux tentatives de normalisation sont présentées en détails dans Moukrim (2010 : 85-116).

[23] Dès leur insertion dans la civilisation musulmane, les Juifs du Mashreq et du Maghreb adoptèrent la langue arabe dans la communication courante comme dans leurs productions littéraires. La transcription de la langue se fait alors en caractères hébreux d'où son appellation de 'judéo-arabe' (Joseph Tedghi (2002 : 415).

[24] Représentation que le chercheur et le transcripteur se font de la langue.

Le choix de la notation usuelle s'explique par les raisons suivantes :

- le thème de recherche ne requiert pas le recours à une transcription phonétique (difficulté de décodage), ni à une transcription phonologique stricte.
- la volonté de fournir un corpus facilement lisible (décodable) par tout le monde et qui prend en compte également la structure morpho-syntaxique de l'énoncé.
- la possibilité d'accompagner la notation usuelle par une transcription phonétique et/ou phonologique pour une étude spécifique (phonétique, intonation...) : possibilité d'ajouter d'autres niveaux de transcription en fonction des objectifs de recherche et des outils de transcription utilisés.

Cependant, le choix d'une transcription 'orthographique' (usuelle) pose problème en l'absence d'un standard stabilisé. Une transcription orthographique *partageable* suppose l'acceptation d'un standard ou à défaut la reprise des pratiques orthographiques les plus courantes au sein de la communauté scientifique.

5.3. Les conventions de transcription

Les corpus de l'arabe marocain et du berbère tamazight ont été transcrit sous TRANSCRIBER[25], un logiciel d'aide à la transcription manuelle de fichiers audio qui permet de transcrire de nombreuses langues y compris non européennes.

Le choix des conventions de transcription n'est jamais neutre, ce qui explique une variété dans les pratiques (Bilger, 2008 : 34). Il dépend de différents facteurs : les objectifs de la recherche, la nature des données primaires (audio ou vidéo), la taille du corpus, la représentation que le chercheur se fait de la langue parlé ou la représentation qu'il veut en donner, la langue du corpus, etc.

Les conventions ne sont pas les mêmes pour une langue à tradition écrite ancienne et bien établie, comme le français, et pour le berbère ou pour l'arabe marocain (langues sans tradition orthographique solide). En effet, deux types de conventions de transcription ont été distingués : des conventions '*spécifiques'* à chaque langue ; et des conventions « *communes »* à tout corpus oral quelle que soit la langue. Pour la transcription des corpus de l'AM et du BT[26], on s'est conformé aux propositions des ateliers de l'INALCO (1996-1998) en tenant compte également des

[25] Télechargeable sur : http://www.ldc.upenn.edu/mirror.Transcriber/

[26] Le même système de notation a été adopté pour la transcription des deux corpus car il n'y pas une grande différence entre le système de notation proposé pour l'écriture du berbère et celui utilisé pour la notation de l'arabe dialectal en caractères latins.

propositions de L'IRCAM[27] concernant le traitement des problèmes spécifiques que posent les dialectes berbères marocains. Quant aux phénomènes liés à l'oralité, ont été retenues les conventions proposées par le LLL pour le corpus du français de l'ESLO[28].

En somme, travaillant dans une perspective de partage et de mutualisation des données et en l'absence d'une norme stabilisée pour chacune des deux langues, nous avons repris les pratiques majoritaires au sein de la communauté scientifique travaillant sur le berbère et l'arabe dialectal, ainsi que sur la langue parlée d'une manière générale.

6. Les aspects juridiques du corpus

La diffusion des corpus oraux suppose une attention accrue au cadre juridique dans lequel s'inscrivent la réalisation et l'exploitation des corpus. l'aspect juridique des corpus linguistiques est une question relativement nouvelle, à laquelle Baude (2006) a apporté des réponses consensuelles et intéressantes.

Les données recueillies dans ce travail se présentent sous forme d'enregistrements de locuteurs s'exprimant en berbère tamazight ou en arabe marocain (ou les deux) dans des situations de communication ordinaires. Se pose alors la question suivante : quel est le statut juridique de ce corpus ?

Les aspects juridiques des corpus oraux concernent trois grands domaines[29] : i) le droit d'auteur et la propriété intellectuelle ; ii) les données personnelles et le respect de la vie privée et ii) les responsabilités des « exploitants » et diffuseurs.

La définition du statut juridique d'un corpus « nécessite de décrire avec précision son contenu puis les conditions d'élaboration et d'exploitation de celui-ci. » (*Ibid.* : 25). Cela permet de « différencier les composantes susceptibles d'être protégées par le droit d'auteur ainsi que les composantes contenant éventuellement des données personnelles » (*Ibid.* : 26).

Pour qu'un corpus (ou une composante de ce corpus) soit considéré comme protégé par le droit d'auteur il faut qu'il remplisse trois conditions : qu'il corresponde à

[27] l'Institut Royal de la Culture Amazighe est une institution publique marocaine dévolue à la promotion du berbère et à son insertion dans l'enseignement, les médias et la vie publique en général.

[28] Les conventions de transcription que nous avons adoptées ne sont pas *individuelles*. Elles sont, comme le signale Mondada (2008 : 87) « le produit historique de discussions et souvent d'accords établis au sein de groupes théoriquement homogènes ».

[29] Baude (2008 : 24)

l'exigence d'une activité créatrice, qu'il ait une forme définie et que cette forme soit originale. Ayant pour objectif d'étudier la langue (BT ou AM) des témoins, on a traité de questions qui relèvent de la vie courante (ordinaire) et qui ne présentent aucune originalité créative. Notre corpus n'est donc pas concerné par les protections du droit d'auteur.

Quant aux données personnelles et le respect de la vie privée[30], leur présence dans un corpus implique une mise en conformité avec la loi informatique et liberté (licéité et loyauté, information préalable, obtention du consentement) ou une anonymisation irréversible de celle-ci (Baude 2008 : 27). L'anonymisation consiste d'abord à repérer les données permettant l'identification directe ou indirecte et celles qui pourraient porter préjudice (propos diffamatoires). Ces données primaires[31] (enregistrements) sont ensuite traitées au moyen d'opérations techniques (bippage, effacement, déformation, etc.).

Le recueil du consentement des personnes enregistrées est la meilleure solution éthique et juridique (Baude 2006, 2008). Toutefois, il faut établir un consentement *« éclairé »* qui démontre que le signataire est informé des finalités de la recherche et des conséquences qu'entraîne sa participation au projet : « […] sans informations préalables précises la demande d'autorisation n'a pas d'objet ni de sens.» (Baude 2006 : 60).

Avant de commencer l'enregistrement, le projet et les finalités de l'enquête sont clairement expliqués aux témoins. Une fois l'enregistrement terminé, ils sont invités à signer l'autorisation qui permettra de disposer des enregistrements et de leur transcription à des fins de recherche. La quasi-totalité des locuteurs ont préféré donner leur consentement oralement. Cela s'expliquerait par le fait que cette procédure de signature les inquiète, du fait qu'elle peut être liée à d'autres pratiques avec lesquelles elle pourrait être confondue comme la signature de chèques (Baude 2006 : 64). La seule personne qui a accepté de signer l'autorisation écrite a un niveau universitaire de troisième cycle. Par ailleurs, et malgré nos explications, certains de nos témoins trouvent contradictoire le fait d'écrire leur nom pour signer l'autorisation et 'l'anonymisation'.

---

[30] Parmi les données personnelles qui permettraient, dans un corpus, d'identifier directement un témoin on compte les formes nominatives, données personnelles, profession, statut, titre, activités sociales, parenté, réseaux, référence à des lieux, référence à des caractéristiques de la personne, caractéristiques physiques, etc. Il en va de même si le témoin peut être identifié par les possibilités de recoupement d'informations (*Ibid.* : 28).

[31] Les corpus oraux sont en général composés d'enregistrements audio ou vidéo (données primaires) et d'annotations de ces derniers (données secondaire) (baude 2006 : 45-46, pour plus de détails).

En écoutant les enregistrements, on a constaté que les témoins étaient tous très discrets sur tout ce qui pourrait les rendre identifiables, comme s'ils prenaient toutes les précautions pour ne pas être reconnaissables, en sorte que les données sont anonymisées à la « source ». Il n'est donc pas nécessaire d'anonymiser les données ni d'obtenir un consentement : « L'obligation d'obtenir un consentement préalable peut être levé si retrouver les personnes concernées s'avère difficile » (Baude 2006 : 109).

Les aspects juridiques des corpus oraux concernent également le domaine de la responsabilité de ceux qui auront à intervenir dans la *« vie du corpus »* : responsabilité des créateurs, responsabilité des exploitants, responsabilité des diffuseurs…

Collecteur, transcripteur et traducteur (les séquences contenant le phénomène linguistique étudié sont accompagnées d'une traduction), nous sommes aussi le premier exploiteur du corpus et on se doit de veiller à la qualité des données, aux respects des finalités indiquées, au respect du principe de licéité et aux conditions de conservation (*Ibid.* : 121).

a. Conclusion

Bien que le corpus de l'arabe marocain et du berbère tamazight ne soit pas sociologiquement représentatif, la diversité qui a présidé à sa constitution permet de le concevoir comme un réservoir dans lequel on peut rechercher des attestations concernant les distributions de faits de langue. Les échantillons retenus ne sauront être véritablement représentatifs de toute la communauté marocaine présente sur Orléans. Néanmoins, se voulant diversifiés, ils offrent également une catégorie particulière de locuteurs, rarement inclus dans les échantillons, des locuteurs en situation irrégulière[32] (des 'sans papiers').

Une fois le corpus constitué et transcrit, nous sommes passée à l'observation des données. Nos résultats proviennent, d'une part, de la comparaison de trois langues différentes (la diversité d'emplois de la forme du « présent » est appréhendée dans la perspective du fonctionnement du langage) et d'autre part, du fait que nous avons travaillé sur des données orales authentiques.

La démarche comparative a permis de dégager les propriétés spécifiques à chacune des langues étudiées. Et le fait de travailler sur des données orales réelles et situées a permis de mettre en lumière des fonctionnements linguistiques qui échappent à l'intuition. Bien que ce travail ne soit qu'une ébauche, la transformation des

[32] Un de nos témoins est en situation irrégulière en France, c'est un 'sans papiers' (il a été sollicité pour un récit de vie) ; deux l'ont été pendant cinq et six ans, ils viennent d'être régularisés.

méthodes d'analyse à partir de nouvelles données va faire progresser la description grammaticale d'une manière générale.

b. Bibliographie

Abouda, L. & Baude, O., (2005) : "Du français fondamental aux ESLO", colloque international Français fondamental, *corpus oraux, contenus d'enseignement*. 50 ans de travaux et d'enjeux, SIHFLES - Laboratoire ICAR, Lyon, 8, 9 et 10 décembre 2005.

Abouda, L. & Baude, O., (2006), « Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des ESLO », *in* F. Rastier, M. Ballabriga (dir.), *Corpus en Lettres et Sciences sociales — Des documents numériques à l'interprétation,* actes du XXVII colloque d'Albi, *Langages et signification,* publiés par C. Duteil-Mougel et B. Foulquié.

Baude, O. (coord.) (2006), *Corpus oraux, Guide des bonnes pratiques* . CNRS éditions et P.U.O.

Baude, O. (2008), « Le droit de la parole » dans Bilger, Mireille (éd.). *Données orales : les enjeux de la transcription*. Perpignan. PUP.pp. 24-33

Benjelloun, S. (2002), « Une double graphie, latine et arabe, pour enseigner l'arabe marocain », in : D. Caubet, S. Chaker, J. Sibille (éds), *Codification des langues de France*, pp. 331-340, L'Harmattan, Paris.

Bergounioux, G. (dir.) (1992), « Enquêtes, Corpus et Témoins », *Langue Française* 93.

Bergounioux G. et al. (1992), « L'Etude socio-linguistique sur Orléans (1966-1991), 25 ans d'histoire d'un corpus», *Langue française*, 93, p. 74-93.

Bilger, M. & Cappeau, P. (2004), L'oral ou la multiplication des styles. *Langage et Société* 109, 13-30.

Bilger, M. (2008), « Les enjeux des choix orthographiques » dans Bilger, Mireille (éd.) *Données orales – Les enjeux de la transcription*. Perpignan. PUP. pp. 248-257

Blanc & Biggs, (1971) « L'enquête socio-linguistique sur le français parlé à Orléans », *Le français dans le Monde*, 85, pp 16-25.

Blanche-Benveniste, C. & Jeanjean, C. (1987), *Le français parlé : transcription et édition*, Paris, Didier-Erudition.

Bouhjar, A. (2003), « Le système graphique tifinagh-Ircam », In *Standardisation de l'amazighe*, Actes du séminaire organisé par le Centre de l'Aménagement

Linguistique à Rabat, 8-9 décembre 2003, Publication de l'Institut Royal de la Culture Amazighe, Série : Colloques et séminaires n°3.

Boukous, A. (2003), « La standardisation de l'amazighe : quelques prémisses », *Standardisation de l'amazighe*, Actes du séminaire organisé par le Centre de l'Aménagement Linguistique à Rabat, 8-9 décembre 2003, Publication de l'Institut Royal de la Culture Amazighe, Série : Colloques et séminaires n°3.

Bourdieu P. (2003), (sous la direction de), *La misère du monde*, Paris, Seuil – Collection Point

Cappeau, P. (2008), « Perception et reconstruction » dans Bilger, Mireille (éd.). *Données orales : les enjeux de la transcription*. Perpignan. PUP. 235-247.

Caubet, D. (1999), « Arabe maghrébin : passage à l'écrit et institutions », In *Faits de Langues*, vol. 7, n° 13, pp. 235-244.

Caubet, D. (2002), « Arabe maghrébin, langue de France : entre deux graphies », in : D. Caubet, S. Chaker, J. Sibille (éds), *Codification des langues de France,* p.331-340, L'Harmattan, Paris, 2002

Cerquiglini, B. (1999), *Les langues de la France*, rapport aux ministres de l'Éducation nationale et de la Culture et de la Communication. (en ligne : http://www.dglf.culture.gouv.fr/lang-reg/rapport_cerquiglini/langues-france.html)

Chaker, S. (1996), « Propositions pour la notation usuelle à base latine du berbère » (Atelier du 24-25 juin 1996, INALCO-CRB. *Synthèse des travaux*).

Chaker, S., Achab, R. & Naït-Zerrad, K. (1998), « Aménagement linguistique de la langue berbère : atelier organisé du 5 au 9 octobre 1998 », *Synthèse* préparée par Chaker, S., Achab, R, Naït-Zerrad, K.

Chaker, S. (2002), « Variation dialectale et codification graphique en berbère. Une notation usuelle pan-berbère est-elle possible ? », in : D. Caubet, S. Chaker, J. Sibille (éds), *Codification des langues de France*, p.331-340, L'Harmattan, Paris, 2002

Gadet, F. (2000), « Derrière les problèmes méthodologiques du recueil des données », dans M. Bilger (dir.), *Linguistique sur corpus*, Presses Universitaires de Perpignan.

Gadet, F. (2008), « L'oreille et l'œil à l'écoute du social » dans Bilger, Mireille (éd.). *Données orales: les enjeux de la transcription*. Perpignan. PUP. 35-47.

Habert, B., Nazarenko, A. & Salem, A. (1997), *Les linguistiques de corpus*, Paris, A. Colin.

Habert B. (2000), « Des corpus représentatifs : de quoi, pour quoi, comment ? », dans M. Bilger (dir.), *Linguistique sur corpus*, Presses universitaires de Perpignan.

Hachid, M. (2000), *Les Berbères, aux origines de l'histoire*, Edisud-Inna yas, Aix-en-Provence, Alger

Labov, W. (1976), *Sociolinguistique*, Minuit.

Labov, W. (1978), *Le parler ordinaire*, Ed. de Minuit, Coll. Le sens commun, Paris.

Maurer, B. (1999), « Quelles méthodes d'enquête sont effectivement employées aujourd'hui en sociolinguistique », dans L.-J. Calvet et P. Dumont (dir.), *L'enquête sociolinguistique*, L'Harmattan.

Mondada, L. (2008), « La transcription dans la perspective de la linguistique interactionnelle » dans Bilger, Mireille (éd.). *Données orales : les enjeux de la transcription*. Perpignan. PUP. pp. 78-109

Moukrim, S. (2010), *Morphosyntaxe et sémantique du « présent » : une étude contrastive à partir de corpus oraux, arabe marocain, berbère tamazight et français (ESLO/LCO)*, Thèse de doctorat, Université d'Orléans

Tedghi, J. (2002), « Usage de la graphie hébraïque dans la transcription des parlers judéo-arabes modernes au Maghreb, in : D. Caubet, S. Chaker, J. Sibille (éds), *Codification des langues de France*, p.331-340, L'Harmattan, Paris, 2002

Le projet OLPC (One Laptop Per Child): développements récents et prochains (2010 - 2011)

Jean M. Thiéry

ModLibre.info
575 A chemin de Rastel, F-13510 EGUILLES
jean.thiery@olpc-france.org

Résumé – Abstract

Nous présentons les développements récents du projet sans but lucratif OLPC décrit lors de SITACAM'09. La plate-forme éducative libre Sugar est maintenant compatible avec les distributions GNU/Linux usuelles telles que Fedora, Mandriva et Ubuntu. Les portables XO récents sont plus puissants mais gardent leur conception robuste et économe en énergie. Les nouveaux XO prévus en 2012 seront des tablettes. Les tablettes commerciales actuelles offrent des applications pédagogiques parfois gratuites mais rarement libres. Les tablettes XO utiliseront la plate-forme éducative libre Sugar.

We present recent developments of the nonprofit OLPC project expounded during SITACAM'09. The free-libre Sugar educational platform is now compatible with common GNU/Linux distributions such as Fedora, Mandriva and Ubuntu. Recent XO laptops are more powerful but keep their original sturdy and low consumption design. New XO predicted in 2012 will be tablets. Present commercial tablets offer some educational applications sometimes free but seldom libre. The XO tablets will run the free-libre Sugar educational platform.

Keywords – Mots Clés

Culture, Éducation, Logiciel libre, Multilingue, OLPC, Ordinateur portable, Pédagogie, Sans but lucratif, Sugar, Tablette.

Culture, Education, Free-libre software, Laptop, Multilingual, Nonprofit, OLPC, Pedagogy, Sugar, Tablet.

1. Introduction

Le *projet OLPC* (*One Laptop Per Child* ou *un ordinateur portable par enfant*) avait été présenté lors de SITACAM'09. L'exposé avait rappelé l'histoire de ce projet basé sur plus de 30 ans de recherches pédagogiques. Depuis 2008, la fondation sans but lucratif *Sugar Labs* (http://www.sugarlabs.org/) a pris en charge le développement des logiciels basés sur l'environnement spécifique Sugar, tandis que la fondation sans but lucratif *OLPC* se consacre à la conception et à la promotion des ordinateurs portables XO. Dans ce document, l'expression *projet OLPC* désigne les projets pilotés par ces deux fondations et les organisations *OLPC* locales (associatives ou gouvernementales).

L'ordinateur XO fonctionne dans des conditions difficiles que ne supporteraient pas les ultra-portables usuels : froid ou chaleur, forte humidité ou poussières, etc. En position fermée, toutes les prises sont protégées. Son clavier est caoutchouté pour l'isoler des poussières et des liquides. Son écran fonctionne en modes couleur ou monochrome (économe en énergie et lisible en pleine lumière). L'alimentation électrique standard de 12 V permet d'utiliser de nombreuses sources de courant (dynamos, panneaux solaires, ...) avec un stockage externe dans des batteries usuelles.

Grâce à leur interface WiFi, les XO se reconnaissent entre eux et peuvent se connecter aux ordinateurs voisins (et aux éventuelles bornes WiFi) pour un travail collaboratif des élèves entre eux ou avec leur enseignant. On crée ainsi un réseau maillé local. Ce réseau peut inclure un serveur d'école XS optimisé pour les XO.

La plate-forme éducative Sugar respecte les quatre libertés informatiques des logiciels libres : libertés d'exécution, d'analyse, de redistribution et d'amélioration des programmes. Pour faciliter la prise en main par de jeunes enfants, Sugar n'affiche qu'une fenêtre à la fois parmi quatre fenêtres possibles. Toutes les activités sont gérées par un journal qui permet de les redémarrer dans l'état où on les avait quittées.

L'exposé de 2009 avait insisté sur l'importance culturelle de ce projet pédagogique qui devrait faciliter la transmission des cultures de générations en générations. Ce projet devrait aussi permettre à chaque culture de se faire connaître dans le monde entier.

2. Évolution de la plate-forme éducative Sugar

La fondation *Sugar Labs* avait été créée pour rendre Sugar compatible avec le maximum de matériels, en particulier, pour réutiliser à des fins pédagogiques des ordinateurs anciens mais toujours vaillants.

Fin 2009, on pouvait tester Sugar sur un CD vif en téléchargeant et en gravant une image *iso* (http://wiki.laptop.org/go/LiveCd). On pouvait l'installer sur une clé USB (http://wiki.sugarlabs.org/go/Downloads). On pouvait enfin l'exécuter dans un émulateur QEMU, VirtualBox ou VMware y compris sur des systèmes Mac ou Windows (http://wiki.sugarlabs.org/go/Supported_systems).

La plate-forme éducative Sugar a été réorganisée selon un modèle multicouche qui garantit sa portabilité à long terme (http://wiki.sugarlabs.org/go/Taxonomy) :
- *Honey* : les activités développées à l'extérieur des *Sugar Labs,*
- *Fructose* : un ensemble d'activités de démonstration testées par les *Sugar Labs,*
- *Glucose* : l'environnement graphique de base,
- *Ribose* : le système d'exploitation et son interface avec Glucose.

La plate-forme éducative Sugar est maintenant disponible dans les dépôts officiels de nombreuses distributions GNU/Linux (http://wiki.sugarlabs.org/go/Downloads). L'intégration dans Fedora est encore la plus complète dans la mesure où les premiers logiciels OLPC-Sugar avaient été développés avec cette distribution.

Voici les principales activités disponibles dans les XO-1 (X) et dans trois distributions grand public : Fedora 14 Laughlin (F), Mandriva 2010.2 (M) et Ubuntu 10.10 Maverick Meerkat (U).

Bureautique et communications

- Calculatrice : calculatrice scientifique (X F M U)
- Dessiner : activité de dessin (X F)
- Discuter : activité pour des discussions collaboratives (X F M U)
- Écrire : traitement de texte dérivé d'Abiword (X F M U)
- Enregistrer : activité gérant la webcam intégrée (X F U)
- Firefox : navigateur Mozilla Firefox (U)
- Help : mode d'emploi en anglais (X F)
- Naviguer : navigateur simplifié(X F M)

Outils pédagogiques

- Moon : informations sur les phases de la lune et sur les éclipses (X F)
- Physics : modélisation des interactions mécaniques entre de nombreux objets (F U)
- Speak : synthèse vocale appréciée par les enfants à partir de 3 ans (X)
- Wikipédia : sélection des meilleurs articles pour l'éducation (X)

Expérimentation

- Distance : distance entre deux XO mesurée avec le temps de propagation d'un son (X)
- Measure : oscilloscope numérique avec transformée de Fourier (X)
- Ruler : règles, grilles et rapporteurs pour mesurer des objets (X)

Programmation pour tous les âges

- Etoys : environnement de modélisation très puissant (X F M U)
- Pippy : initiation à Python (langage de la plupart des activités) (X F M U)
- Scratch : programmation multimédia (X)
- TamTam : création et reproduction de séquences musicales (X F)
- TurtleArt : adaptation du langage Logo de Seymour Papert et al. (X F M U)

Voyage au cœur du système

- Analyze : analyseur de communication (X F)
- Log : historique (X F M U)
- Terminal : console (X F M U)

3. La diffusion des XO

Les XO ne sont pas vendus directement au grand public. Les raisons sont complexes et difficiles à pondérer : en particulier la multiplicité des claviers fragmente le *marché* potentiel. Les deux ventes *G1G1* (*Give 1 Get 1*) permettant d'acheter deux XO (l'un pour une association et l'autre pour l'acheteur) n'ont pas été encourageantes.

La distribution normale des XO se fait donc dans le cadre de *marchés publics* entre une collectivité territoriale (ville, région ou pays) et le fabriquant actuel (la société taïwanaise Quanta) avec la mise en concurrence d'autres solutions éventuelles.

Avant de prendre de telles décisions, forcément coûteuses, il est souhaitable de lancer des opérations pilotes pour
- assurer la traduction de tous les logiciels indispensables en utilisant le site collaboratif http://translate.sugarlabs.org,
- proposer des documents pédagogiques conformes aux programmes nationaux,
- équiper des écoles représentatives et
- sensibiliser tous les décideurs potentiels.

La fondation OLPC peut fournir des XO pour ces opérations pilotes mais les demandes sont très nombreuses. Il est souhaitable que les dossiers soient bien préparés et bien coordonnés, en particulier avec l'aide des associations OLPC locales.

Plus de 2 millions de XO ont été diffusés dans le monde entier mais surtout au Pérou (870 000) et en Uruguay (380 000) (http://en.wikipedia.org/wiki/One_Laptop_per_Child). On peut obtenir des statistiques plus précises en consultant la carte du site officiel (http://one.laptop.org/map) ou la carte communautaire (http://www.olpcmap.net).

4. Le futur

La fondation OLPC étudie le cahier des charges des futurs ordinateurs XO. Elle suit la démarche classique des projets innovants qui proposent des concepts, puis des prototypes, avant de passer à la construction en série.

Le XO-1.75 devrait être disponible à la rentrée prochaine. Il sera très proche des XO actuels mais il sera optimisé avec des composants plus performants. En particulier il utilisera l'architecture ARM (http://fr.wikipedia.org/wiki/Architecture_ARM) qui prédomine dans l'informatique embarquée (téléphonie mobile, tablettes, …) grâce à sa faible consommation. Le changement d'architecture nécessitera de nombreuses modifications dans les couches profondes de Sugar (*Ribose* et *Glucose* : voir ci-dessus) mais devrait peu modifier les couches supérieures (*Fructose* et *Honey*) écrites majoritairement dans des langages de script (essentiellement Python).

Le projet XO-2 a été abandonné pour des raisons de coût.

Le XO-3 sera probablement une tablette avec un écran tactile. Il utilisera aussi une architecture ARM. Le clavier mécanique serait remplacé par un clavier virtuel adaptable facilement à tous les alphabets et toutes les dispositions de touches. Cela facilitera la diffusion des XO dans les pays qui possèdent plusieurs claviers officiels.

Les tablettes actuelles sont fragiles. La fondation OLPC souhaiterait des tablettes relativement souples pour mieux résister aux chocs. D'autres pistes sont explorées en attendant que cela soit possible. En particulier, la fondation OLPC et la société Marvell (http://en.wikipedia.org/wiki/One_Laptop_per_Child) mettent au point une tablette optimisée pour l'enseignement, qui sera diffusée dans un premier temps avec le système Androïd.

« Les produits comme les liseuses et les tablettes usuelles sont des plate-formes formidables pour la culture, l'audiovisuel et le jeu. Elles ne satisfont pas les exigences d'un modèle éducatif basé sur la construction et non seulement sur la consommation. Les environnements éducatifs actuels nécessitent des plate-formes robustes pour les calculs, la création de contenus et l'expérimentation. Et le tout à un prix très bas » comme le dit Dr Nicholas Negroponte, fondateur et président d'OLPC (http://www.tablets.com/tablets/one-laptop-per-childs-olpc/).

L'utilisation des tablettes actuelles confirme rapidement cette analyse. En particulier, les tablettes « haut de gamme » sont pratiquement fermées : connectique limitée, logiciels propriétaires inadaptables et passage obligatoire par des boutiques en ligne spécifiques. La plupart des autres tablettes sont basées sur le système Androïd optimisé pour les dispositifs mobiles (tablettes, téléphones, etc.). Ce système a été adopté par de très nombreux constructeurs séduits par son ouverture facilitant toutes les adaptations nécessaires. Androïd est globalement libre car il est basé sur GNU/Linux et Java. Son ouverture n'est pas garantie à long terme et chaque constructeur cherche à imposer sa boutique en ligne spécifique.

En mars 2011, il existait déjà plus de 150 000 applications Androïd dont environ 57 % étaient gratuites (http://fr.wikipedia.org/wiki/Android_Market). On peut les étudier en simulant un système Androïd dans une machine virtuelle comme VirtualBox (http://www.virtualbox.org/). Les applications classées dans la catégorie éducation sont très inégales. Certaines servent de vitrines pour des sites commerciaux ; d'autres ont un réel intérêt pédagogique. L'analyse de ces applications est importante car Androïd sera bientôt utilisé par des enfants, en particulier sur des tablettes ou des téléphones mobiles déclassés mais encore utilisables pour certaines applications.

L'utilisation des tablettes actuelles, avec leurs meilleures applications éducatives, permet d'anticiper la tablette XO-3 : l'intérêt pédagogique de l'écran tactile, la souplesse et les limites du clavier virtuel, etc. Cependant on n'obtiendra jamais une véritable plate-forme éducative en juxtaposant des applications disparates. La plate-forme éducative Sugar révolutionnera l'usage des tablettes pour l'enseignement.

5. Conclusion

Les nouveaux XO sont très attendus. On peut adapter ou développer leurs logiciels dès maintenant sur des ordinateurs personnels usuels et tester certaines applications sur des tablettes récentes.

Remerciements

Je remercie tous les membres d'OLPC-France qui m'ont transmis des références indispensables et en particulier Bastien Guerry et Lionel Laské. Ce document est diffusé sous licence *Creative Commons cc-by-sa*, l'une des licences recommandées par le projet OLPC (http://wiki.laptop.org/go/Licensing).

Références

Thiéry J. M. (2009), Le projet OLPC (One Laptop Per Child) : un atout pour toutes les cultures, A*ctes de SITACAM'09*, 15-31.

Contribution à la reconnaissance des caractères Tifinagh par utilisation des réseaux de neurones et la squelettisation

K. MORO(1), B.EL KESSAB(1), M.FAKIR(1), B.BOUIKHALENE(2), S.SAFI(2)
(1) Equipe de traitement de l'information et télécommunication
Faculté des Sciences et Techniques
Université Sultan Moulay Slimane, Béni Mellal, Maroc
Email : kamalmoro@hotmail.com, bade10@hotmail.fr, fakfad@yahoo.fr
(2) Equipe de traitement de l'information et télécommunication
Faculté Poly disciplinaire,
Université Sultan Moulay Slimane, Béni Mellal, Maroc
bbouikhalene@yahoo.fr, said_safi@gmail.fr

Résumé - Abstract:

Dans ce document, nous présentons un système de reconnaissance des caractères basé sur les réseaux de neurones. Nous avons appliqué l'égalisation d'histogramme, la correction d'inclinaison et la squelettisation au niveau du prétraitement, et nous avons montré l'avantage du traitement sur le squelette par rapport à la forme brute. Les résultats expérimentaux ont été appliqués sur la base des caractères Tifinagh.

In this paper, we present a recognition system of characters based on neural networks. We applied histogram equalization, tilt correction and thinning at the preprocessing and we have shown the benefit of treatment on the skeleton from the raw form. The experimental results were applied on the base on the characters Tifinagh.

Keywords - Mots clés :

Squelettisation ; Réseaux de neurones ; Perceptron multicouche ; Caractères Tifinagh.

Skeletonization; Neural Networks; Multilayer perceptron; Tifinagh characters.

1. Introduction

La reconnaissance des caractères est l'une des axes les plus actifs dans le domaine de la reconnaissance des formes, plusieurs travaux de recherche ont été appliqués sur les caractères latins (R.M. Bozinivik et al 1989, M.K. Brown 1983), arabe (M.Fakir et al 1993, S.I.Ibrahim et al 2005), Gujarati (M.K.Jndal et al 2007,

B.Anuradhasrinivas et al 2008, A.Dessa, 2010) et Tifinaghe (M. Amrouch et al 2009, B. El Kessab et al 2009). Dans ce document, nous avons choisi de travailler sur la base des caractères Tifinagh de l'IRCAM, cette base est parmi les plus utilisées par la communauté scientifique. La Fig.1 montre des échantillons de cette base.

Fig.1 Exemple des caractères Tifinagh

Dans ce document, nous avons utilisé le perceptron multicouche au niveau de la classification, mais avant, nous avons utilisé la normalisation des caractères, appliqué l'égalisation d'histogramme et la squelettisation au niveau du prétraitement. La Fig.2 montre le processus de reconnaissance. Le but de cet article est de montrer la valeur ajoutée à l'utilisation du squelette au lieu du traitement sur la forme brute afin de montrer l'avantage de l'utilisation de la squelettisation.

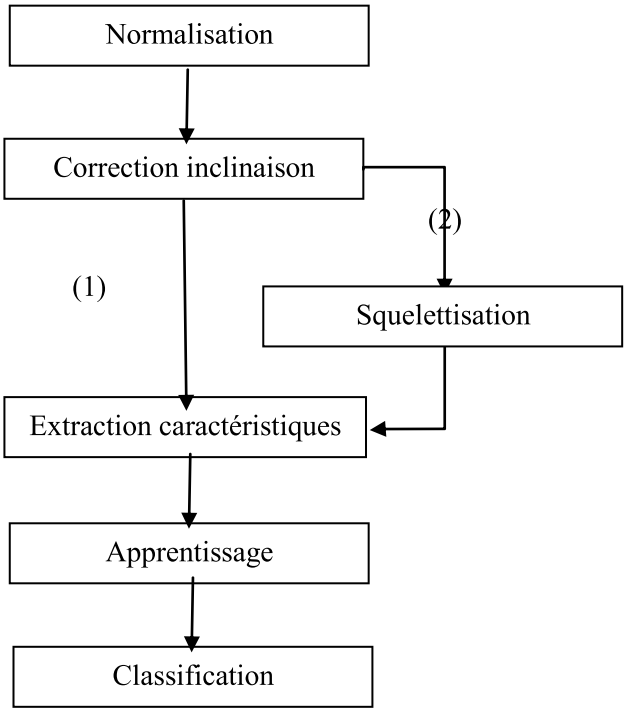


Fig.2. Processus de reconnaissance

2. Squelettisation

Un des problèmes fondamentaux en reconnaissance de formes est la représentation synthétique de celle-ci. Dans de nombreux cas, le travail sur la forme brute est laborieux et inutile. Il est beaucoup plus avantageux en terme de temps et de qualité de travailler sur une forme épurée. La notion de squelette a été introduite à cet effet.

Dans le plan continu, le squelette d'une forme est un ensemble de lignes passant en son milieu.

C'est la notion d'axe médian d'une forme continue introduite par (Blum 1967). Il existe actuellement une grande variété de méthodes permettant de construire des squelettes à partir de formes, parmi lesquelles l'amincissement topologique qui consiste à retirer au fur et à mesure les points du contour de la forme, tout en préservant ses caractéristiques topologiques. Dans ce document, nous avons choisi d'utiliser l'algorithme de (Guo_Hall 1989), celui-ci utilise l'approche parallèle d'amincissement, il préserve la topologie et la géométrie du squelette et il est parmi les plus utilisés par la communauté scientifique (K.MORO et al 2009).

3. Extraction des caractéristiques

L'extraction des caractéristiques est la phase la plus importante dans le domaine de la reconnaissance des caractères. (*R. El Yachi et al* 2010) ont utilisé les moments invariants pour la reconnaissance des caractères Tifinaghe. (*R. El Kessab et al 2009)* ont utilisé les vecteurs de cavité et l'ont appliqué aussi sur les caractères Tifinaghe. Dans ce document, nous avons choisi d'utiliser une méthode qui est à la fois simple et efficace, cette méthode consiste à faire la somme des valeurs des pixels au niveau horizontal, vertical et des deux diagonales. La Fig.4 montre la façon de sommer les pixels pour une image 3x3.

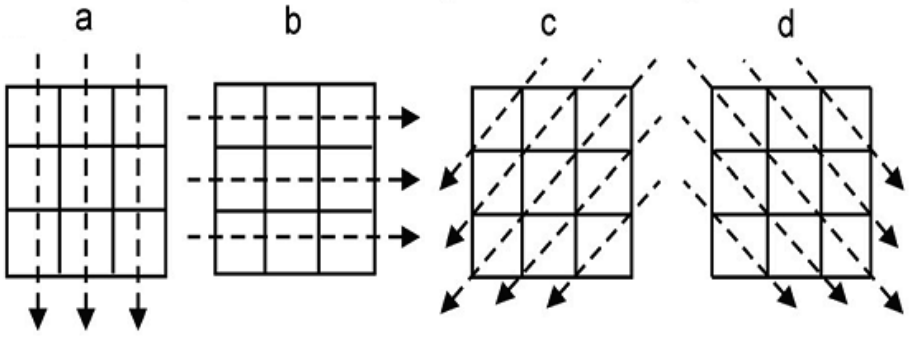


Fig.3. Extraction pour une image 3x3

Par exemple, en considérant la forme de la Fig.4, le Tab.1 montre le vecteur d'extraction en suivant les schémas de la Fig.3.

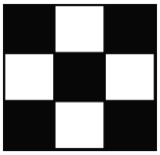


Fig.4 Forme 3x3 pixels

Fig.3	Vecteur d'extraction
a	(2,1,2)
b	(2,1,2)
c	(1,0,3,2,1)
d	(1,0,3,2,1)

Tab.1. Vecteurs d'extraction

4. Réseaux de neurones :

Plusieurs travaux de recherche ont utilisé les réseaux de neurones au niveau de la classification des caractéristiques (*B. El Kessab et al 2009*, *M. Amrouch et al 2009*, *A. Dessai* 2010). Dans ce document, nous allons choisir de même le perceptron multicouche comme type du réseau pour la classification. La méthode de rétro propagation du gradient est utilisée au niveau de l'apprentissage, la fonction sigmoïde $f(x) = \frac{1}{1+e^{-\alpha x}}$ comme fonction d'activation au niveau de la couche d'entrée et cachée, avec $\alpha = 0.1$, la fonction seuil au niveau de la couche de sortie et nous avons fixé la constante d'apprentissage à $\gamma = 0{,}1$. L'architecture multicouche du réseau consiste à utiliser 94, 50 et 10 neurones au niveau de la couche d'entrée, cachée et sortie respectivement.

5. Apprentissage :

Dans ce document, 90 caractères sont pris en considération pour l'apprentissage, celui-ci est effectué dans un premier temps sur 10 chiffres écrits sous la forme standard. Après, nous effectuons l'apprentissage sur 30 caractères, ceux-ci représentent les 10 classes de la base, 3 caractères pour chaque classe. De même, nous utilisons encore une fois un troisième apprentissage sur 90 caractères,

9 caractères pour chaque classe, les matrices des poids calculés sont ensuite utilisées pour la reconnaissance.

6. Résultats expérimentaux :

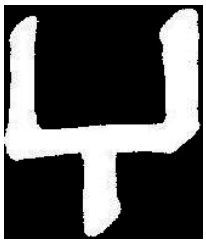
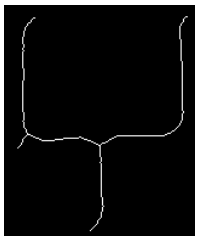
Dans la partie expérimentale, nous avons effectué la reconnaissance sur 2100 caractères Tifinagh, soit 210 caractères pour chaque classe. Dans un premier temps, nous effectuons la normalisation des caractères de la base, cette opération a pour but de faciliter le traitement afin que les vecteurs d'extraction des différents caractères de la base gardent la même dimension.

Après la normalisation, nous appliquons l'ajustement du contraste sur les caractères, cette opération a pour but d'éliminer les différentes intensités du gris des pixels de l'image pour la rendre binaire, pour cela, nous avons utilisé l'algorithme d'égalisation d'histogramme (CLAHE) pour cet effet, La Fig.5 montre le résultat de cette manipulation.

Avant égalisation d'histogramme Après égalisation d'histogramme

Fig.3 égalisation d'histogramme du caractère yaf

Après l'égalisation d'histogramme vient l'étape de la squelettisation, cette démarche a pour but de présenter la forme avec un minimum d'informations et pour réduire le temps d'exécution, la Fig.4 montre le résultat de l'algorithme de (Guo_Hall 1989) qui est utilisé dans ce document.

Caractère Squelette

Fig.4 Squelette du caractère yagh

Avant de passer à l'apprentissage, nous entamons l'étape de l'extraction des caractéristiques, comme expliqué dans la partie extraction des caractéristiques dans ce document, un vecteur de 94 éléments est construit, l'exemple d'un vecteur d'extraction des caractéristiques est le suivant :

(12 11 7 8 9 5 4 3 3 4 5 5 5 5 5 5 2 6 10 12 10 8 11 11 8 4 0 0 5
2 2 5 1 1 2 3 3 4 4 3 4 5 6 5 6 6 7 9 7 6 4 2 2 2 1 2 1 0 0 0 0 0 0
1 1 1 2 3 3 3 3 2 3 4 3 3 4 4 7 6 5 5 5 6 6 4 1 2 2 1 2 2 1 1).

Nous procédons ensuite à l'étape de l'apprentissage. Comme nous l'avons mentionné dans la partie introduction, nous avons effectué l'apprentissage sur 90 caractères, puis nous avons utilisé les poids calculés pour la reconnaissance, le Tab.2 montre l'avantage de l'utilisation du squelette par rapport à la forme brute, que ce soit au niveau de l'apprentissage ou la reconnaissance. D'après le tableau, on remarque que le taux des caractères reconnus lors de l'apprentissage est 100% dans tous les phases avec l'utilisation du squelette, alors que ces taux n'atteignent pas 100% dans tous les cas sans l'utilisation du squelette, aussi, le nombre des caractères reconnus après l'apprentissage est 1943 avec utilisation du squelette, soit 92.52% des caractères, alors que ce nombre est 1742 sans l'utilisation du squelette, soit 82.95.

On remarque d'après les tableaux Tab.5 et Tab.6 que le taux de reconnaissance augmente avec l'utilisation du squelette pour la plupart des caractères, aussi, les caractères yahh et yad et les caractères yag et yach sont assez confondu du fait de leurs ressemblances, avec et sans l'utilisation du squelette.. Dans le cas général, l'utilisation de la squelettisation avant la phase de l'extraction des caractéristiques fait augmenter le taux de reconnaissance des caractères de la base Tifinagh.

Nombre caractères	Type de test	Taux de reconnaissance (%)	
		Avec Squelette	Sans squelette
10	Apprentissage	100	100
30	Apprentissage	100	100
90	Apprentissage	100	93.33
2100	Reconnaissance	92.52	82.95

Tab 2 : Résultats expérimentaux

Char.	Sans squelette										
											Succès(%)
	186	1	0	18	0	0	0	4	1	0	88.57
	0	205	0	5	0	0	0	0	0	0	97.62
	0	68	136	1	5	0	0	0	0	0	64.76
	3	0	0	195	0	7	3	0	0	2	92.86
	0	0	4	2	149	0	21	2	0	32	70.95
	1	0	0	48	0	115	0	0	46	0	54.76
	0	0	0	8	0	0	201	0	1	0	95.71
	10	2	0	18	0	0	0	180	0	0	85.71
	0	23	0	0	0	1	0	0	167	19	79.52
	0	0	0	1	0	0	0	1	0	208	99.05

Tab.3 : Performance du réseau sans l'utilisation du squelette sur les caractères de reconnaissance

Char.	Avec squelette										
											Succès(%)
	178	0	4	16	0	2	10	0	0	0	84.76
	0	210	0	0	0	0	0	0	0	0	100
	0	1	206	0	0	2	0	0	0	1	98.10
	4	5	9	189	0	1	0	0	0	2	90
	1	0	0	2	190	7	0	2	0	8	90.48
	2	8	0	2	0	198	0	0	0	0	94.29
	1	0	4	2	1	1	198	2	0	1	94.29
	0	2	24	0	0	0	4	178	0	2	84.76
	0	2	0	0	0	2	0	8	198	0	94.29
	0	0	4	2	5	0	0	1	0	198	94.29

Tab.4 : Performance du réseau avec l'utilisation du squelette sur les caractères de reconnaissance

7. Conclusion

Dans ce document, nous avons utilisé le perceptron multicouche dans la phase de la classification des caractéristiques de la base des caractères Tifinagh. Plusieurs techniques ont été implémentées dans le processus de reconnaissance avant la phase de la classification, parmi lesquelles la squelettisation. Nous avons montré que celle-ci permet d'augmenter le taux de reconnaissance par rapport au traitement sur la forme brute du caractère. Elle nous a permis d'atteindre un taux de reconnaissance de 92.52%.

La performance de chaque méthode de classification est basée sur l'extraction des caractéristiques. Dans nos perspectives, nous comptons appliquer d'autres techniques d'extraction dans le processus de reconnaissance et utiliser les réseaux de Markov cachés et les réseaux bayesiens au niveau de la classification.

References

R. M. Bozinovic and S. N. Shihari (1989), Off Line Cursive Script Word Recognition, IEEE Trans.Pattern Anal. Mach. Intell. PAMI 11, pp. 68- 83.

M. K. Brown (1983), pre-processing techniques for cursive word recognition, Pattern Recognition, Vol.13, N°.5, pp: 447-451.

M. Fakir and C. Sodeyama (1993), Recognition of Arabic printed Scripts by Dynamic Programming Matching Method, IECICE Trans. Inf & Syst, Vol. E76- D, pp: 31-37.
Ibrahim S.I. Abuhaiba (2005), Arabic Font Recognition using Decision Trees Built from Common Words, Journal of computing and information technology- CIT 13, pp:211-223.
M.K. Jindal, R.K. Sharma, G.S. Lehal (2007), Segmentation of horizontally overlapping lines in printed Indian scripts, International Journal of Computational Intelligence Research 3 (4) 277–286.
B. Anuradhasrinivas, A. Agarwal, R. Rao (2008), An overview of OCR research in indian scripts, International Journal of Computer Science and Engineering System 2 141–153.
A. Desai (2010), Gujarati handwritten numeral optical character reorganization through neural network Pattern Recognition 43-2010, pp: 2582-2589
M.Amrouch et al (2009), Apprentissage Markovien et Neuronal: cas des caractères amazighes imprimés, Sitacam'09 Agadir Morocco 12-13, pp: 58-67.
B. El Kessab, B. Bouikhalene, M. Fakir, S. Safi (2009), Reconnaissance des caractéres Tifinaghe par l'utilisation des réseaux de neurones multicouches, Sitacam'09 Agadir Morocco, pp: 68-83
H. BLUM (1967), A transformation for extracting new descriptions of shape. *In Models for the Perception of Speech and Visual Form*, pp: 362–380. MIT Press.
Z. Guo and R.W. Hall (1989). Parallel thinning with two subiteration algorithms. *Comm. ACM*, 32(3) :359–373.
K. Moro, M. Fakir, B. Bouikhalene, S. Safi (2009), Skeletonization Methods Evaluation for the Recognition of Printed Tifinaghe Characters, Sitacam'09, pp:33-47.
R. El Yachi, K. Moro, M. Fakir, B. Bouikhalene (2010), Utilisation des moments invariants et la programmation dynamique pour la reconnaissance des caractéres Tifinagh, Journal of Theoretical and Applied Information Technology, pp : 61-66.

Tifinagh characters recognition using Self Organizing Map and Fuzzy k-Nearest Neighbor

Mohamed FAKIR (1), Belaid BOUIKHALENE (2), Said GOUNANE (1)

(1) Information Processing & Telecommunication Team
Dep. Of Computer Sciences – FST, University Sultan Moulay Slimane,
Beni Mellal, Morocco
fakfad@yahoo.fr,, gounane.said@gmail.com
(2)) Information Processing & Telecommunication Team
Faculty multidisciplinaire, University Sultan Moulay Slimane
Beni Mellal, Morocco,
bbouikhalene@yahoo.fr

Résumé – Abstract

Cet article présente une comparaison entre l'algorithme des cartes auto-organisatrices de Kohonen et celui des k-plus proche voisins flou et leur application dans la reconnaissance des caractères Tifinagh manuscrites. L'extraction des caractéristiques des caractères s'est basée sur la technique du codage rétinien décrite dans (D.Slezak, Junzhong, & T.Kim).Pour la démonstration, un ensemble de 200 exemples a été utilisé pour la phase d'entrainement. Les résultats ont montré que l'efficacité et la rapidité de l'algorithme des k-Plus Proches Voisins Flou.

In this paper we present a comparison between SOM (Self-Organization Map) neural network and Fuzzy K-nearest Neighbor algorithms and their application to handwriting Tifinagh character recognition. The Box approach proposed in (M.Hanmandlu, A.V.Nath, A.C.Mishra, & V.K.Madasu, 2007) is used for features extraction. Experimentation is carried out on a limited database of nearly 200 samples. The results showed that Fuzzy K-Nearest Neighbor had a very good performance.

Keywords – Mots Clés

Caractères Tifinagh, k-plus proches voisins flou, cartes auto- organisatrices, k-means Flou, extraction des caractéristiques.
Tifinagh characters, Fuzzy k-Nearest Neighbor, Self Organizing Map, Fuzzy k-means, Features extraction.

Introduction

The recognition of characters from scanned images of documents has been a problem that has received much attention in the fields of image processing, pattern recognition and artificial intelligence.

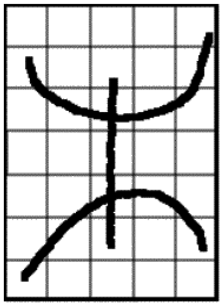
For many years, fuzzy logic and Artificial Neural Networks have been used in a wide range of problem domains: process control (where Fuzzy Controllers have been very popular), management and decision making, operations research, economics and pattern recognition and classification.

This paper presents an application of both SOM neural network and fuzzy k-Nearest Neighbor in recognition of handwritten Tifinagh characters. This paper is organized as follows: in section (M.Hanmandlu, A.V.Nath, A.C.Mishra, & V.K.Madasu, 2007) a features extraction method, which is an essential step prior to pattern recognition, is described. Section (haykin & Simon, 1999) describes the architecture and the learning mechanism of the SOM neural networks. Section (Mingoti & Lima, 2005) presents the k-Nearest Neighbor algorithm. Section (D.Slezak, Junzhong, & T.Kim) gives results of the application of both SOM and Fuzzy k-NN on Tifinagh handwriting character recognition.

Features extraction

Preprocessing techniques like thinning, slant correction and smoothening are applied. For extracting the features, the Box approach proposed in (M.Hanmandlu, A.V.Nath, A.C.Mishra, & V.K.Madasu, 2007) is used here. This approach requires the spatial division of the character image. The major advantage of this approach stems from its robustness to variation, ease of implementation and high recognition rate. Each character image is divided into L × H boxes so that the portions of character will be in some of these boxes. There could be boxes that are empty (Figure 5(a)). The choice of number of boxes is arrived at by experimentation. Elements of a Normalized Vector that describe the character is obtained by dividing the number of all black pixels present in this box by their total number for each box, (Figure 5 (b)).

One can easily see that this characterization is invariant of the character image dimensions. Hence an image of whatever size gets transformed into a vector of L × H predetermined dimensions.



0	0	0	0	0.1
0.26	0	0.08	0	0.3
0.1	0.35	0.44	0.34	0.12
0	0	0.36	0	0
0	0.1	0.45	0.34	0.12
0.12	0.3	0.24	0	0.27
0.21	0	0	0	0

(a) (b)

Figure 5 Features extraction

Self Organizing Map (SOM)

The Self-Organizing Maps, abbreviated SOM, was developed by Professor Teuvo Kohonen in the early 1980s. Self-organizing maps are a special class of artificial neural network, because those are based on competitive learning and the learning itself is unsupervised.

Also SOM is considered as a special case in data-mining, it can be applied to both clustering and projecting the data onto a lower dimensional display at the same time.

In a SOM network, there is an input layer and an output layer which is usually designed as two-dimensional arrangement of neurons that maps n dimensional input to two dimensional (Figure 6). It is basically a competitive network with the characteristic of self-organization providing a topology-preserving mapping from the input space to the clusters.

The algorithm proceeds first by initializing the synaptic weights in the map for the neurons. This can be done by assigning them small values picked from a random number generator; in so doing no prior order is imposed on the feature map. Once the map has been properly initialized, there are three essential processes involved in the formation of the self-organizing map: Competition, cooperation and synaptic adaptation.

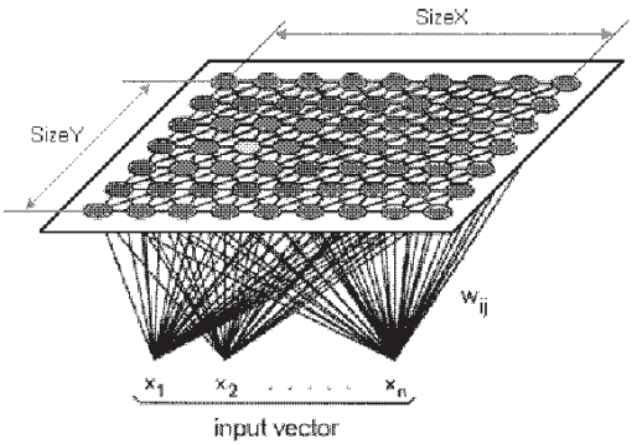


Figure 6: SOM architecture

Competition

Let $X = [x_1,..., x_n]$ be the input pattern and $W_j = [w_{j1},..., w_{jn}]$ the synaptic weight vector of each neuron j in the map. To find the best match of the input vector X with the synaptic weight vectors W_j, compare the inner products $W_j^T.X$ for j=1,2,…,l (l is the number of output neurons) and select the largest denoted $i(X)$. Maximizing $W_j^T.X$ is equivalent to minimizing $W_j - X$ (haykin & Simon, 1999) then one can have:

$$i(X) = \arg\min_j \left(W_j - X\right) \qquad (1)$$

The neuron number $i(X)$ is called the winning neuron.

Cooperation

In neurobiology, a neuron that is firing tends to excite neurons in its immediate neighborhood more than those farther away. The same with output neurons in the SOM, a topological neighborhood around the winning neuron i is made and it decay smoothly with lateral distance. Another unique feature of the SOM algorithm is that the size of the topological neighborhood shrinks with time. Let $h_{j,i}(k)$ denote the topological neighborhood at time k, centered on the winning neuron i, and j denote a typical neuron of a set of excited (cooperating) neurons around

winning neuron i. On can assume that the topological neighborhood $h_{j,i}$ is unimodal function of the lateral distance $d_{i,j}$, such that it satisfies:

1. $h_{j,i}$ is symmetric and attains its maximum value at the winning neuron i for which $d_{j,i}$=0.
2. The amplitude of $h_{j,i}$ decreases monotonically with increasing $d_{i,j}$, decaying to zero for $d_{i,j} \to \infty$.

A typical choice of $h_{j,i}$ is the Gaussian function

$$h_{j,i}(k) = e^{\left(\frac{d_{i,j}^2}{2\sigma^2(k)}\right)} \qquad (2)$$

$$\sigma(k) = \sigma_0 e^{\left(\frac{k}{\tau_1}\right)} \qquad (3)$$

Where τ_1 is the time constant of the algorithm.

Synaptic adaptation

By definition, for the network to be self-organizing (and unsupervised), the synaptic weight vector W_j of neuron j in the network is required to change in relation to the input vector X. This change is expressed by the equation as follows:

$$W_j(k+1) = W_j(k) + \alpha(k) h_{j,i(X)}(k)\left(X - W_j(k)\right) \qquad (4)$$

Where $\alpha(k)$ is the learning-rate.
The exact form of learning-rate is not important. It can be linear, exponential or inversely proportional. However it should be time varying. In particular, it should start at an initial value α_0 , and then decrease gradually with increasing time n. This requirement can be satisfied by using an exponential learning-rate, as shown by:

$$\alpha(k) = \alpha_0 e^{\left(-\frac{k}{\tau_2}\right)} \qquad (5)$$

Where τ_2 is another time constant of the SOM algorithm.

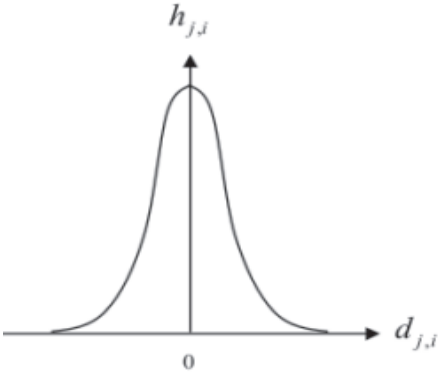


Figure 7 : example of topological neighbourhoods (Gaussian)

Fuzzy K-Nearest Neighbor

Fuzzy kNN is part of supervised learning that has been used in many applications in the field of data mining, statistical pattern recognition, image processing and many others. Some successful applications are including recognition of handwriting and satellite image. Fuzzy K nearest neighbor algorithm is very simple. It works based on an unsupervised clustering algorithm like fuzzy k-means algorithm to determine prototypes representing clusters. Then the minimum distance from the query instance to these prototypes is used to determine the K-nearest neighbors. After and basing on membership function we take the neighbor with the maximum membership to be the prediction of the query instance.

Fuzzy k-means

The Fuzzy k-means clustering algorithm is an improvement of the k-means algorithms. K-Means are the simplest methods of clustering data. The fuzzy K-means algorithm uses a set of *N* unlabeled feature vectors and classifies them into k classes, where k is given by the user.

From these *N* feature vectors, k of them are randomly selected as initial seeds. The feature vectors are assigned to the closest seeds depending on its distance from it and on a given membership function. The mean of features belonging to a class is taken as the new center. The features are reassigned; this process is repeated until convergence.

In a fuzzy clustering, a pattern vector can belong to all clusters with different degrees given by a membership function [figure 4]. One can prove that such a function exists (J.Sagau), and for each cluster C_i it is defined as follows:

$$f_i(X) = \frac{1/d^2(X,c_i)^{\frac{1}{m-1}}}{\sum_{i=1}^{N} 1/d^2(X,c_i)^{\frac{1}{m-1}}} \quad m \in \Re, \quad m \geq 1 \qquad (6)$$

Where c_i is the center of the class C_i, and The parameter m is defined by the user to adjust the fuzziness of the clustering. The hard clustering case is obtained by taking m = 1.

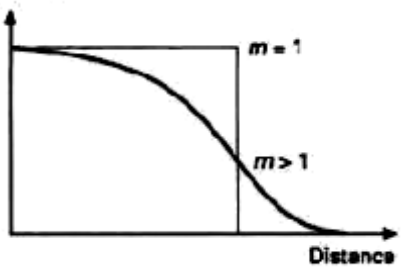


Figure 8: Membership function

Fuzzy K-means algorithm

7. Choose randomly the k prototypes c_i .

8. Compute the degree of membership of all feature vectors X_j in all clusters C_i : $\mu_{ij} = f_i(X_j)$.

9. Compute new cluster prototypes as follows:

$$c_i = \frac{\sum_{j=1}^{N} (\mu_{ij})^m x_j}{\sum_{j=1}^{N} (\mu_{ij})^m} \qquad (7)$$

10. Iterate back and force between (2) and (3) until the memberships or cluster centers for successive iteration differ by more than some prescribed value ε.

Fuzzy K-Nearest Neighbor

The Fuzzy k-NN classifiers consist on proximity measures. They are ideally suited for modeling the non parametric distribution on handwritten word recognition data.

For a given character X, the fuzzy classifier computes the membership of X in different classes $C_1, C_2, ...C_N$. The character X is allocated to the class for which the membership function yields the maximum value. By a learning process (Fuzzy k-mean, k-mean, LVQ ...) we assign a number of prototypes P_j for each class C_i. After generating the k-Nearest Neighbor prototypes P_j for a character (by distance similarity), the degree of membership of the vector X to the class C_i: $\mu_i(X)$, can be calculated as follow:

$$\mu_i(X) = \frac{\sum_{i=1}^{k} \mu_{ij} / d^2\left(X, P_j\right)^{\frac{1}{m-1}}}{\sum_{i=1}^{N} 1 / d^2\left(X, P_j\right)^{\frac{1}{m-1}}} \qquad (8)$$

Where μ_{ij} is the membership of the prototype P_j to the class C_i. To compute this value tow methods are proposed:
Using a crisp assignment of P_j to C_j :

$$\mu_{ij} = \begin{cases} 1 & P_j \in C_i \\ 0 & P_j \notin C_i \end{cases} \qquad (9)$$

Or by using the k-NN to determine k-nearest neighbor of P_j, then the number of neighbor belonging to the same class as P_j denoted n_j is used to compute μ_{ij} as follows:

$$\mu_{ij} = \begin{cases} 0.51 + 0.49 n_j / k & j = i \\ 0.49 n_j / k & j \neq i \end{cases} \qquad (10)$$

Results

A java application was developed in order to test and compare those two algorithms. To extract features each, character was divided into boxes, the input layer of the SOM neural network was made of 63 neurons and 35 for the output layer (number of character to recognize), for the fuzzy k-nearest neighbor the fuzziness factor m is equal to 2 and number of neighbors' k is equal to 3.

As there is no standard database available for handwritten Tifinagh characters, a database was created from samples of different writing styles with different sizes. The database also includes some complex samples that are even hard to recognize by careful inspection Figure 9: Samples used for training (figure 5). This database contains 200 samples (about 5 samples per character) divided into two disjoint sets, one for training both SOM and Fuzzy k-NN, and another for testing these two algorithms.

The features extraction method used here is scaling invariant, that's why there is always a recognition problem between Tifinagh character ⵄand ⵇ, where ⵄ's are mistaken as ⵇ's and vice versa. Some other recognition problems are with ⵣ and ⵙ, ⵇ and ⵔ but not persistent as the first one.
Experiments shows that the best results are given by the Fuzzy k-NN where in many cases it success to recognize handwritten characters that SOM algorithm fails to (figure 6).

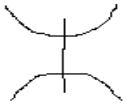
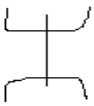
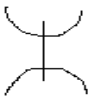
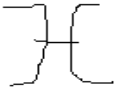
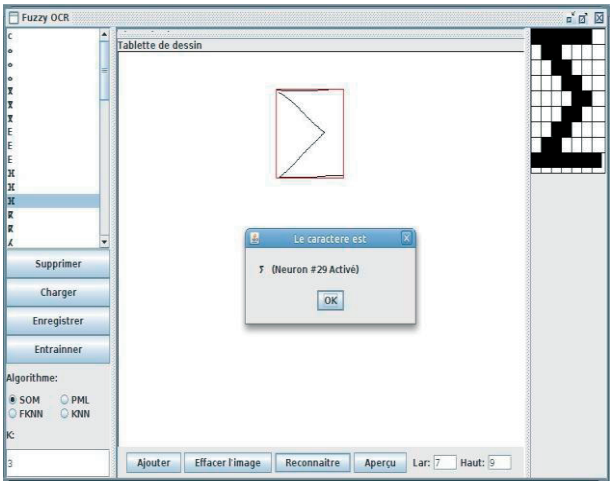
	Sample1	Sample2	Sample3	Sample4
ⵣ				
ⵥ				
ⵙ				
ⵛ				
ⵀ				
ⵅ				
ⵄ				

Figure 9: Samples used for training



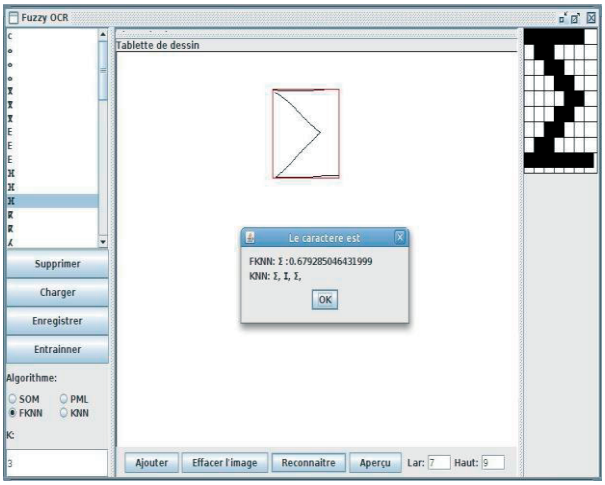


Figure 10: ⵣhandwritten character.SOM algorithm (left) fails (ⵖ) but not Fuzzy k-NN(right) (ⵣ)

Conclusion

In this paper two clustering algorithms are presented, Self Organizing Map neural network and the Fuzzy k-Nearest Neighbor. We applied these two algorithms to Tifinagh character recognition. The box approach was used to extract features from the character image. Fuzzy k-NN was more robust and fast than the SOM.

References

(J.C.), & BEZDEK. (1981). Pattern recognition with fuzzy objective function algorithms. Plenum Press.

A.Mingoti, S., & Lima, J. O. (2005). Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. European Journal of Operational Research .

A.Tellaeche, Burgos-Artizzu, X. P., G.Pajares, & A.Ribeiro. (2007). On Combining Support Vector Machines and Fuzzy K-Means in Vision-based Precision Agriculture. World Academy of Science, Engineering and Technology .

D.Slezak, B.Kang-Junzhong, T.Kim, & G.Kuroda. Segnal processing, Image processing and Pattern recognation. New York: Springer.

F.Zheng. (n.d.). K-Means-based fuzzy classifier.

haykin, & Simon. (1999). Neural Networks: a comprehensive fondation. new jersey: Prentice-hall inc.

J.Sagau. (n.d.). Logique flou en classification. Techniques de l'Ingénieur , H3638.

M.Hanmandlu, A.V.Nath, A.C.Mishra, & V.K.Madasu. (2007). Fuzzy Model Based Recognition of Handwritten Hindi Numerals using bacterial foraging. 6th IEEE/ACIS ICIS .
Nasser, S., Alkhaldi, R., & Vert, G. (n.d.). A Modified Fuzzy K-means Clustering using Expectation Maximization.
Nassery, P., & Faez, K. (n.d.). Signature pattern recognition using psudo zernike moments and fuzzy logic classifier.
O.Matan, R.K.Kiang, C.E.Stenard, & B.Boser. (1990). Handwritten character recognition using neural network architectures. 4th USPS advanced Technology Conference, Washington D.C , 1003-1011.
S.Nasser, R.Alkhaldi, & G.Vert. (n.d.). A Modified Fuzzy K-means Clustering using expectation Maximization.
S.Theodoridis, & Koutroumbas, K. (2009). Pattern recognation (fort edition). Elsevier.
V.Ganapathy, & Liew, K. L. (2008). Handwritten Character Recognition Using Multiscale Neural Network Training Technique. World Academy of Science, Engineering and Technology .

Contributions à la Reconnaissance Hors Ligne de l'Ecriture Amazighe

[1]Y. Es Saady, [2]A. Rachidi, [1]M. El Yassa, [1]D. Mammass

[1]IRF – SIC, Faculté des Sciences, B.P. 8106, Hay Dakhla, Université Ibn Zohr, Agadir, Maroc,
essaady2110@yahoo.fr, melyass@gmail.com, mammass@univ-ibnzohr.ac.ma.

[2]Ecole nationale de Commerce et de Gestion, B. P. 37/S Hay Salam, IRF – SIC, Faculté des sciences, Université Ibn Zohr, Agadir, Maroc,
rachidiali69@gmail.com.

Résumé – Abstract

La reconnaissance de **l'écriture** Amazighe a connu ces dernières années un intérêt important dans les travaux de recherche. En effet, quelques approches ont été développées pour reconnaître ces écritures. Elles sont aussi variées que celles utilisées pour les écritures des autres langues. Dans ce papier, nous décrivons les différents travaux publiés dans le domaine de la reconnaissance de l'écriture Amazighe avec nos contributions dans ce domaine, ainsi les bases de caractères Amazighes existantes. L'objectif est de faire une synthèse de ces travaux qui nous permettra de lancer des perspectives pour les travaux futurs.

Amazigh handwriting recognition has in recent years a significant interest in research work. Indeed, some approaches have been developed to recognize these scripts. They are as varied as those used for the scripts of other languages. In this paper, we describe the various published works in the field of Amazigh handwriting recognition with our contributions in this field and the existence of Amazigh characters database. The aim is to make a synthesis of this work will enable us to launch outlook for futures works.

Mots Clés – Keywords

Caractères Amazighes, Reconnaissance de caractères, Reconnaissance de l'écriture, Tifinagh.
Amazigh characters, Character recognition, Handwriting recognition, Tifinagh.

Introduction

La reconnaissance automatique de l'écriture manuscrite ou imprimée reste encore un sujet de recherche et d'expérimentation. Le problème n'est pas encore entièrement résolu bien que l'on sache atteindre des taux assez élevés dans certaines applications et pour certaines langues. Plusieurs recherches scientifiques ont été effectuées sur l'écriture latine, arabe, et autres. Ceci a permis le développement de plusieurs approches de reconnaissance automatique de ces écritures. Par contre, l'écriture Amazighe, appelée Tifinaghe, est peu traitée. Quelques tentatives ont été menées pour améliorer la situation actuelle. Elles sont regroupées généralement en grandes classes telles que les approches statistiques (Oulamara, 1988), (Djematen et al., 1997), Les réseaux de neurones (Ait Ouguengay, 2008), (El Yachi et al., 2009), (Bouikhalene et al., 2009) (Es Saady et al., 2011), l'approche syntaxique (Es Saady et al., 2008), (Es Saady et al., 2010), les Modèles de Markov cachés (Amrouch et al., 2009), (Amrouch et al., 2010) et la programmation dynamique (El Yachi et al., 2010). On s'intéresse dans ce papier de présenter une synthèse des travaux de recherches publiés dans le domaine de la reconnaissance automatique de l'écriture Amazighes imprimés et manuscrits.
Nous présentons en première partie de ce papier les principales caractéristiques de la langue Amazighe. La seconde partie est consacrée à la description des différents travaux effectués dans le domaine de la reconnaissance automatique de l'écriture Amazighe avec nos contributions dans ce domaine. Nous présentons dans la troisième partie les principales bases de données de caractères Amazighes existantes et enfin nous proposons certaines perspectives de développement sur le domaine.

Ecriture Amazighe

Le Tifinaghe est le système d'écriture de la langue Amazighe. Il tire son origine du vieil alphabet libyque et saharien, déjà utilisé depuis le VIème siècle avant l'ère chrétienne par les populations de l'Afrique du Nord, du Sahel et des Iles Canaries. Cet alphabet a subi des modifications et des variations depuis son origine jusqu'à nos jours (Rachidi et Mammass, 2007).

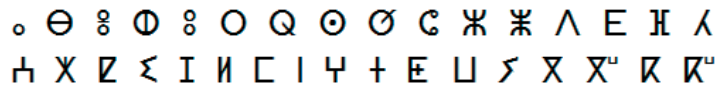


Figure 1: Alphabet Tifinaghe IRCAM.

La figure 1 ci-dessus présente les différents modèles de l'alphabet Amazighe (Tifinaghe-IRCAM). Il comporte trente trois lettres. A la différence des caractères latins et arabes, l'écriture Amazighe n'est jamais cursive, ce qui facilite toute opération de segmentation. La majorité des modèles graphiques des caractères est composée de points, de petits cercles, et/ou de segments. De plus, l'écriture Amazighe est écrite de gauche à droite, elle utilise des signes de ponctuation classique acceptés en alphabet latin. La figure 2 ci-dessous présente un exemple du texte Amazighe dans un manuel scolaire.

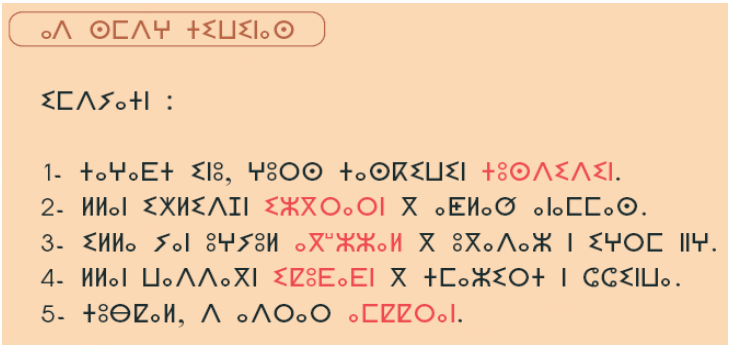


Figure 2: Exemple du texte Amazighe dans un manuel scolaire.

Différentes Approches Proposées pour la Reconnaissance de l'Alphabet Amazighe

Etat de l'art

Parmi les anciennes études qui portent sur la reconnaissance de caractères Amazighes ont été publiées dans les travaux d'Oulamara (Oulamara, 1988). La méthode proposée dans cette référence est une méthode statistique basée sur l'extraction de segments de droite par la transformée de Hough. L'analyse du caractère dans l'espace paramétrique, obtenu par la transformation de Hough, permet d'extraire les caractéristiques spécifiques en association avec un modèle de référence générateur de l'ensemble des caractères de l'alphabet. Un codage original

est déduit puis utilisé comme base de construction de la matrice de lecture représentant une forme codée de l'alphabet. L'auteur a obtenu des résultats qui semblent intéressants sur les caractères Amazighes imprimés d'une base locale.

Dans (Djematen et al., 1998), Les auteurs proposent une méthode statistique de reconnaissance de caractères berbères manuscrits basée sur la position des points caractéristiques dans le rectangle-enveloppe de l'image du caractère. Après des prétraitements (normalisation bidirectionnelle, le lissage, l'extraction des composantes connexes) sur le caractère, des primitives sont extraites sur chaque squelette, comme les extrémités, les points, les sommets (points de changements de direction) et les nœuds à 3 et 4 branches. En fin, La représentation du caractère fournit une description sous forme de lettres utilisant un codage prédéfini. Cette description code les positions des points caractéristiques du caractère dans le rectangle-enveloppe. La reconnaissance consiste à mesurer le degré de ressemblance entre le code élaboré et les codes de référence en utilisant la distance métrique. Selon les auteurs, cette approche a donné des bons résultats sur une base de caractères localement définie malgré quelques erreurs qui viennent du module du prétraitement.

Ait Ouguengay (Ait Ouguengay, 2008), a proposé un réseau de neurones artificiels (RNA) pour la reconnaissance de caractères Amazighes. Le réseau de neurones utilisé est un perceptron multicouche á une seule couche cachée. Ce dernier a été entrainé sur une base de données qui contient des patterns de la graphie Amazighe de différentes fontes et de tailles. La simulation du réseau de neurones a été réalisée par le logiciel libre JavaNNS (java neural networks simulator). Un vecteur de 8 primitives géométriques extrait sur chaque caractère alimente le réseau. L'approche a été expérimentée sur une base des patterns de la graphie Amazighe imprimés. D'après l'auteur, cette approche a donné des bons résultats sur l'ensemble des patterns d'entraînement. Cependant, les résultats de test sont encore loin d'être satisfaisante à cause de la base de test qui est très faible par rapport aux poids de RNA à déterminer.

Dans (Amrouch et al., 2010), les auteurs présentent un système automatique de reconnaissance de caractères Amazighes basé sur le Modèle de Markov cachée (HMM). Après des prétraitements sur l'image du caractère, la chaîne représentative du caractère a été construite à partir de la transformation de Hough. La chaîne obtenue est traduite en séquence d'observations qui est utilisée, lors de la phase d'apprentissage, par le HMM. Le classifieur Forword a été utilisé pour reconnaître le caractère. Selon les auteurs Les résultats obtenus sont prometteurs sur une base localement définie. Cependant, la discrimination de ces modèles n'est pas très bonne parce que chaque modèle de Markov caché utilise l'apprentissage d'un seul caractère. Le taux d'erreur a été enregistré principalement en raison de la mauvaise écriture et les données d'apprentissage.

Dans (El Yachi et al., 2010), les auteurs proposent un système de reconnaissance de l'écriture Tifinaghe basé sur les moments invariants et la transformée de Walsh utilisant la programmation dynamique. Le système proposé contient trois parties principales: le prétraitement, l'extraction de caractéristiques et la reconnaissance. Dans le processus de pré-traitement, l'image du document numérisé est nettoyée, puis segmentée en lignes par les techniques de l'histogramme vertical. Les lignes sont segmentées en mots puis en caractères à l'aide de l'histogramme horizontal. Dans le processus d'extraction de caractéristiques, les moments invariants et les coefficients de Walsh sont calculés sur les caractères segmentés. La programmation dynamique est adoptée dans l'étape de reconnaissance. Les tests ont été faits sur plusieurs images d'écriture Amazighe. D'après les auteurs, Les résultats expérimentaux montrent que la méthode de la reconnaissance utilisant des moments invariants donnent de meilleurs résultats par rapport à la méthode fondée sur la transformée de Walsh en termes de taux de reconnaissance, de taux d'erreur et de temps de calcul.

Nos Contributions

Dans cet axe de recherche, on a proposé plusieurs systemes de reconnaissance de caracteres et de textes amazighes. A travers ces systemes, on a réussi à améliorer les différents taux.

En effet, nous avons présenté, au départ, un système automatique de reconnaissance de caractères Amazighes imprimés, basé sur une approche syntaxique utilisant les automates finis (Es Saady et al., 2010). Après des prétraitements sur l'image du caractère, des algorithmes appropriés sur le squelette de caractère permettent de construire la chaîne représentative du caractère à partir du codage de Freeman. La chaîne reconstruite est utilisée à l'entrée d'un automate fini qui reconnaît les caractères Amazighes segmentés.

La représentation d'un caractère Amazighe par une grammaire régulière se fait en trois étapes. L'étape initiale est le prétraitement, où l'on construit le squelette du caractère. La seconde permet d'extraire des points caractéristiques du caractère étudié et enfin, la construction de la chaîne représentative du caractère est réalisée. En effet, nous avons utilisé l'algorithme de Zhang-Suen pour construire le squelette du caractère (Zhang and Suen, 1984). Une fois le squelette obtenu, nous le décomposons en un ensemble de segments élémentaires. Tout d'abord, trois types de points caractéristiques seront extraits du squelette: les points d'extrémité, les points de croisement et les points d'inflexion. Ensuite, un algorithme de suivi de squelette permet de construire les segments d'un caractère. Pour cela, on recherche uniquement les extrémités et les croisements et on les relie en passant par les inflexions. Enfin, chaque caractère Amazighe est alors représenté par un ensemble de primitives en utilisant les 8 directions de codage de Freeman. Chaque caractère

est représenté par une grammaire régulière, puis par un automate fini en utilisant les règles de correspondance entre les grammaires régulières et les automates finis. Une fois qu'on a les automates de tous les caractères, on génère un automate global qui reconnaît tous les caractères Amazighes segmentés. Cet automate global a été construit à partir des automates de reconnaissance spécifique à chaque caractère Amazighe.
Nous avons testé notre application sur une base de caractères Amazighes imprimés que nous avons crée. Cette approche a connu un taux de reconnaissance de 93,48% sur 660 caractères Amazighes segmentés. Les erreurs proviennent de la forme de certains caractères dont le squelette comporte des segments non orthogonaux et non reconnus par le codage de Freeman.
La limite de cette approche est qu'elle n'est pas applicable pour les caractères non segmentés. Pour remédier à ces limites, nous avons proposé, dans un deuxieme temps, une nouvelle approche qui tient compte de tous les caractères Amazighes (Es Saady et al., 2011). En effet, nous avons développé un système automatique de reconnaissance de l'écriture Amazighe à base de linge centrale de l'écriture. L'architecture générale du système proposé se présentera dans la figure 3 ci-dessous.

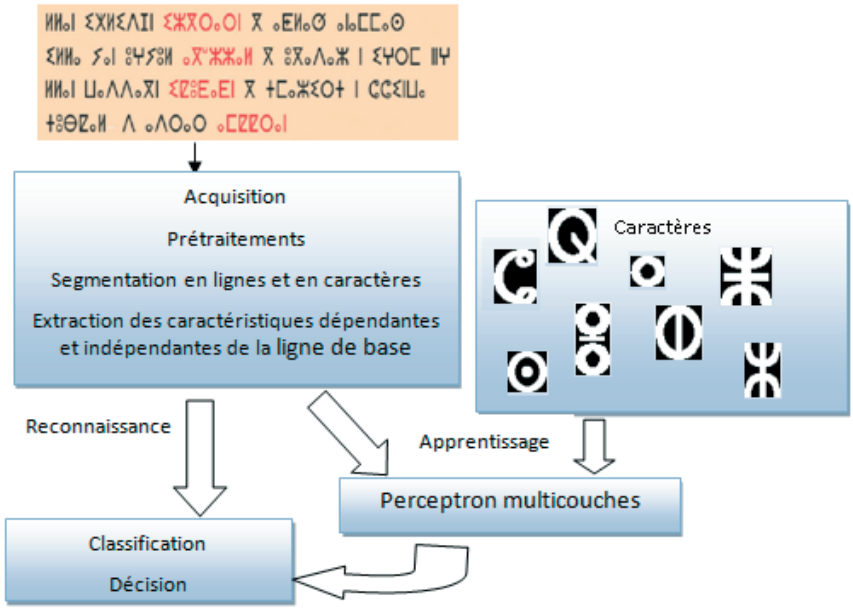


Figure 3: L'architecture du système de reconnaissance proposé dans (Es Saady et al., 2011).

Après des prétraitements sur l'image, le texte est segmenté en lignes et puis en caractères en utilisant les techniques d'analyse d'histogramme de projections

horizontales et verticales. Les positions des lignes de base du caractère (une ligne centrale, ligne supérieure et inférieure de l'écriture) sont utilisées pour obtenir un ensemble de caractéristiques indépendantes et dépendantes à ces lignes. Ces caractéristiques sont liées aux densités de pixels et sont extraites sur les images binaires des caractères en se basant sur l'utilisation de la technique des fenêtres glissantes. Ces primitives alimenteront un réseau de neurones multicouches dans les phases d'apprentissage et de reconnaissance. Le système a montré de bonnes performances sur une base de 19437 paternes Amazighes imprimés élaborée par Ait Ouguengay (Ait Ouguengay 2006). Les résultats trouvés montrent une amélioration significative du taux de reconnaissance lorsqu'on intègre les caractéristiques dépendantes de la ligne de base. Parmi les travaux futurs de ce travail, nous allons ajouter d'autres caractéristiques qui améliorent les résultats pour certains caractères dont le taux de reconnaissance est faible par apport aux restes. En plus, nous allons appliquer notre approche sur la base de données manuscrite développée localement.

Principales bases de données existantes

Dans le domaine de la reconnaissance automatique de l'écriture Amazighe, les bases de données d'images annotées sont inexistantes. Ce domaine a ainsi péché de l'absence d'une base de données de référence, qui permette des comparaisons objectives entre les différents systèmes. Tous les travaux publiés dans ce domaine, cités précédemment, ont été expérimentés sur des bases de données locales, qui contiennent un nombre restreint de l'alphabet Amazighe.
Parmi ces bases de caractères, on cite en premier lieu, celle élaborée par Ait Ouguengay (Ait Ouguengay 2006), qui contient à peut prés vingt mille caractères imprimés et elle a été utilisée pour tester les approches citées dans (Ait Ouguengay 2006) et (Es Saady 2011).
En second lieu, nous avons crée une base de caractères Amazighes manuscrits qui sera rendu public prochainement.
Dans les deux sous sections suivantes, nous présentons ces deux bases de caractères.

Base de caractères imprimés

La base des patterns de la graphie Amazighe élaborée par Ait Ouguengay (Ait Ouguengay 2006), est une base des patterns de différentes fontes Amazighes et de tailles variées. Elle contient au total 12 polices de caractères et les tailles du 10 points au 28 points pour chaque modèle. Les patterns sont fournis sous forme d'images bitonales de tailles variables. La taille maximale est de 102×129 pixels, tandis que la taille minimale est de 19×2 pixels. Une telle disparité s'explique par le fait que le caractère 'a' (ya) est un petit cercle, et est donc beaucoup plus petit

que les autres caractères. Outre le cas particulier du caractère ‘ya’, la base est constituée des patterns de différentes fontes Amazighes et de tailles variées, qui ne sont pas normalisées.
La manière dont sont stockées les images des patterns, dans cette base, ne permet pas la possibilité de renormaliser leur taille en une taille moyenne fixe. En effet, Ceci peut être gênant en particulier à cause de la ressemblance des caractères ’a’ (ya) et ’r’ (yar), qui ne se différencient que par la taille: le caractère ‘ya’ est un petit cercle, tandis que le caractère ‘yar’ est un grand cercle. Dans certains cas, on aura une confusion réelle entre des images de ces deux classes. Ce problème aura une influence sur les résultats des tests.

Base de caractères manuscrits

Nous avons réalisé une base de caractères Amazighes manuscrits recensés aux prés de plus de 50 scripteurs de différents âges, différents sexes, de différentes fonctions, de différents niveau d’etude, de différents moments et sur différents supports. Chaque scripteur nous a donné 13 × 33 caractères. Les échantillons ont été collectés en demandant aux participants d'écrire sur un formulaire 13 exemples pour chaque caractère Amazighe. Nous avons recueilli 350 documents, la figure 4 montre un exemple de formulaire utilisé pour la collecte des données. Après la numérisation de ces documents, nous avons développé un système automatique qui traite le document et le segmente en caractères isolés. Au total, nous avons obtenu 650 modèles pour chaque caractère. Par conséquent, la base contient 21450 caractères (650 × 33). Certains échantillons des caractères Amazighes sont donnés dans le tableau 1.

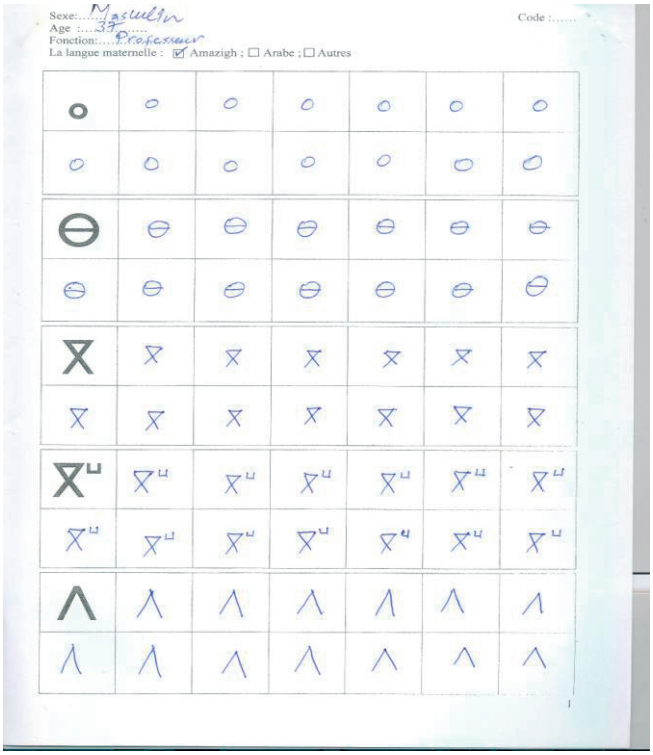


Figure 4: Exemple du formulaire utilisé pour la collecte des données.

Conclusion et Perspectives

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

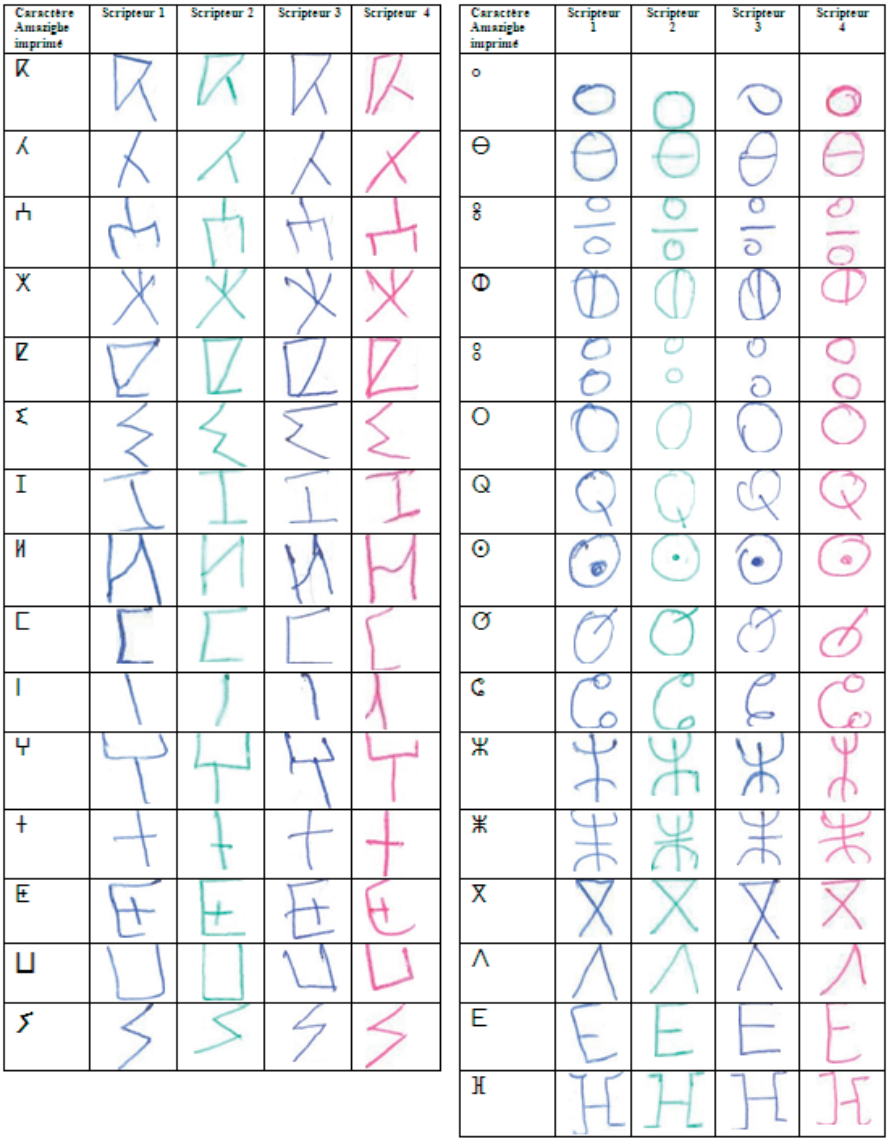


Tableau 1 : Certains échantillons des caractères Amazighes manuscrits.

Références

Ait Ouguengay Y., Taalabi M. (2008), Elaboration d'un réseau de neurones artificiel pour la reconnaissance optique de la graphie amazighe, Phase d'apprentissage, Actes de *SITA'08*, INPT, Maroc.

Amrouch M., Es Saady Y., Rachidi A., Elyassa M., Mammass D. (2009), Printed Amazigh Character Recognition by a Hybrid Approach Based on Hidden Markov Models and the Hough Transform, International Conference on Multimedia Computing and Systems, Actes de *ICMCS'09*, Ouarzazate, Maroc.

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

Bouikhalene, M.Fakir B., Safi S., El Kessab B. (2009), Reconnaissance des Caractères Tifinaghe Par L'utilisation Des Réseaux de Neurones Multicouches, Actes de *SITACAM'09*, Agadir, Maroc.

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

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

El Yachi R. and Fakir M. (2009), Recognition of Tifinaghe Characters using Neural Network, International Conference on Multimedia Computing and Systems, Actes de *ICMCS'09*, Ouarzazate, Maroc.

El Yachi R., Moro K., Fakir M., Bouikhalene B. (2010), On the Recognition of Tifinaghe Scripts, *Journal of Theoretical and Applied Information Technology*, Vol.20, No.2, pp.61-66.

Es Saady Y., Rachidi A., Elyassa M., Mammass D. (2008), Une méthode syntaxique pour la reconnaissance de caractères Amazighes imprimés, Actes de *CARI'08*- Maroc.

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

Es Saady Y., Rachidi A., El Yassa M., Mammass D. (2011), Reconnaissance Automatique de l'Ecriture Amazighe à base de Ligne Centrale de l'Écriture, *4ème Atelier international sur l'amazighe et les TIC*, 24-25 février 2011, IRCAM, Rabat.

Oulamara A., Duvernoy J. (1988). An application of the Hough transform to automatic recognition of Berber characters. *Signal Processing*, vol.14, no.1, pp.79-90.

Rachidi A., Mammass D. (2007), Vers un système de traduction automatique en ligne des documents Amazighes fondé sur les graphes UNL, Revue internationale sur les technologies de l'information, *revue E-TI*, ISSN 1114-8802, Vol.4.

Zhang T.Y. and Suen C.Y. (1984), A fast parallel algorithm for thinning digital patterns, *Communications of the ACM*, 27(3), pp.236–240.

Application de la géométrie riemannienne à la reconnaissance des caractères Tifinagh

O.BENCHAREF(1), M.FAKIR(1), N.IDRISSI(1)
B.BOUIKHALEN(2), B. MINAOUI(3),
(1) Equipe de Traitement de l'Information et Télécommunication
Département d'Informatique, Faculté des Sciences et Techniques,
Université Sultan Moulay Slimane, Béni Mellal - Maroc
bencharef98@gmail.com, fakfad@yahoo.fr
(2) Equipe de Traitement de l'Information et Télécommunication
Faculté Poly disciplinaire, Université Sultan Moulay Slimane
bbouikhalene@yahoo.fr
(3) Equipe de Traitement de l'Information et Télécommunication
Département de physique, Faculté des Sciences et Techniques,
Université Sultan Moulay Slimane, Beni Mellal – Maroc
Bra_minaoui@yahoo.fr

Résumé – Abstract

Dans ce travail, nous utilisons des descripteurs de formes appelés descripteurs métrique, basés sur le calcul de la métrique riemannienne. Ces descripteurs ont montré une fiabilité considérable vis-à-vis, le changement d'échelle, l'existence de bruit et les déformations géométriques. Pour la classification on utilise une méthode basée sur les séparateurs à vaste marge et une autre sur les réseaux de neurones. Nous illustrons cette approche sur la reconnaissance des caractères Tifinagh manuscrits et imprimés.

In this paper, we present shape descriptors that we call metric descriptors, based on the computation of the Riemannian metric. They give considerable reliability to the change of scale, the existence of noise and to geometric distortions. An approach based on the Support vector machine and another one on neural networks are used for the learning and recognition step. We illustrate the proposed approach to the Tifinagh manuscripts and printed character recognition.

Keywords – Mots Clés

Reconnaissance des caractères manuscrits, Caractères Tifinagh, métrique Riemannienne, SVM, Réseaux de neurones, descripteurs métrique.
Recognition of handwritten characters, characters Tifinagh, Riemannian metric, SVM, Neural network, Metric descriptors.

Introduction

Récemment, la vision par ordinateur est devenue l'un des domaines les plus attrayants de la recherche. La reconnaissance de formes représente l'un des principaux piliers de cette science. Dans le schéma classique de reconnaissance de forme, on peut lister deux grandes étapes : l'extraction des descripteurs et la classification :

L'extraction des descripteurs est une forme particulière de réduction de la taille, qui consiste à simplifier le montant des ressources nécessaires pour décrire un large ensemble de données avec précision. Par ailleurs différentes techniques ont été utilisées et pour plus de détails on peut consulter (F. ALT, 1962), (H. Drilten, 1977) et (Chee-way, 2003).
Dans ce papier nous proposons une nouvelle approche basée sur le calcul de la métrique Riemannienne (E. Kalsen, 2004), (X. Gu, 2004). Les variétés munies d'une structure riemannienne, dite aussi variétés riemanniennes, sont des outils très puissants de la géométrie différentielle. Elles apparaissent dans de nombreux contextes pour compenser l'insuffisance de la géométrie euclidienne. On les utilise à titre d'exemple pour le calcul de la longueur d'un chemin s'inscrivant sur une sphère, Estimation du volume, La génération de maillage adaptatif et la représentation d'objets 3D. Poursuivant les investigations sur ces outils nous avons essayé de les adapter à la reconnaissance des caractères manuscrits Tifinagh.
Pour tester notre approche, nous avons voté pour les deux classificateurs connu par leur robustes, les Réseaux de neurones et Les séparateurs a vaste marge (SVM).
Ce papier est organisé comme suit: La deuxième section présente un aperçu sur les caractères Tifinagh. Dans la troisième partie nous donnons quelques notions de base de la géométrie riemannienne et nous décrivons la méthode proposée. La section quatre présentes la classification. Et enfin la cinquième section qui est consacré aux résultats expérimentaux.

Les caractères Tifinagh et la Base de données test

Historiquement les caractères Tifinagh sont connus chez les foqha (théologiens) marocains sous nom de ((khath RRaml) caractères des sables, ça veut dire les

écritures des caravaniers du désert, ils sont utilisés pour échanger les messages entre les nomades, qui dessinent les signes sur les routes sahariennes, et ensuite ils deviennent caractères quasi-magiques; a cause de l'importance de la communication et les itinéraires des voyages et ensuite partie de l'histoire. Ces caractères sont conservés chez les sahariens et sont aujourd'hui l'écriture ancestrale des Touarègues, les chercheurs ont découvert les textes en tifinagh avec différentes formes ; formes des êtres humains; des formes géométriques; et d'autres symboles de la sacralité .avec des ressemblances avec les caractères d'autres peuples du mondes comme des phéniciens, les russes, les hiéroglyphes et araméen.
Certains connaisseurs des symboles tifinagh explique le terme disant qu'il se compose en Tamazight de deux mots (TIFI : découverte et nagh: de soi) ce mot composé veut dire: découverte de nous même.
Actuellement dans le monde entier différentes équipe de recherches sont penchées sur le traitement automatique de la culture amazighe en l'occurrence l'équipe TIT de la FST de Béni Mellal dans le domaine de la reconnaissance des caractères et documents tifinagh.
L'institut royal de la culture amazigh ICRAM propose une standardisation des caractères tifinagh composé de 33 caractères (Figure 1)

Figure 1: Les caractères Tifinagh.

Extraction des descripteurs Métriques

Notion de base de la géométrie riemannienne

La notion de variété reflète l'idée que l'espace peut être courbé et avoir une topologie compliquée, mais qu'il peut être assimilé localement à R^n. l'assimilation ne signifie pas que la métrique est la même, mais que les notions de base de l'analyse comme ensembles ouverts, fonctions et coordonnées sont les mêmes. La

variété complète est alors reconstituée en raccordant sans discontinuité toutes ces régions locales à savoir la sphère le tore et le plan (Figure 2). (E.KALSEN & al(2004))&(J.KERL (2008))

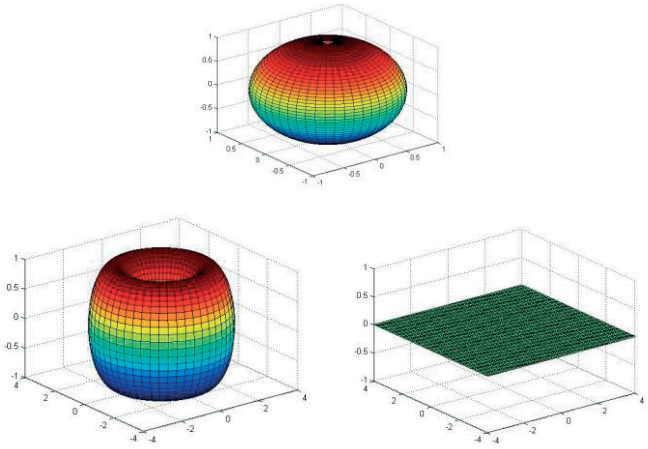


Figure 2 : Exemple de variété géométrique (sphère, tore, plan).

Pour plus de détails on présente quelques définitions nécessaires en géométrie riemannienne :

Définition 1 : On appelle espace métrique euclidien un espace vectoriel muni d'un produit
scalaire $<,>_M$ défini par un tenseur de métrique : c'est-à-dire une matrice symétrique définie
positive M. On le note (R^n,M). Le produit scalaire de deux vecteurs u et v est donné Par $<u, v>_M = t_u M v$. La distance entre deux points a et b est notée par $d_M(a, b)$.

Définition 2 : On appelle variété riemannienne ou espace métrique riemannien, toute variété
Continue Ω incluse dans R^n munie d'une métrique M On la note (x, M(x)) avec x un élément de R^n. La restriction de la métrique M(.) en un point x de la variété définit un produit scalaire sur l'espace tangent $T_x M$. Muni de cette structure, l'espace tangent a une structure d'espace métrique euclidien. Un exemple de variété riemannienne est donné dans la figure 3 :

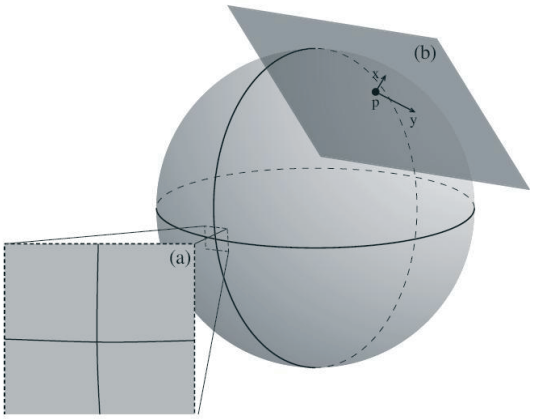


Figure 3: La sphère (variété riemannienne).

En chaque point, il existe un plan tangent qui dépend des plans (a) et (b). Dans un plan (a), les propriétés métriques de la sphère sont approchées par les propriétés métriques du plan euclidien.

Méthode proposée

Un système de reconnaissance est composé de trois phases, le prétraitement, l'extraction de descripteurs et la classification :

1. Le prétraitement **:** la métrique riemannienne se calcule à partir des dérivées partielles en chaque point. Pour cela simuler une image contenant des lettres manuscrites (souvent représentées en noir et blanc) par une variété et calculant sa métrique ne donne pas les résultats souhaitées. (La dérivée partielle en chaque point est nulle et par conséquent la métrique est nulle en ce point). Pour remédier à ce problème nous avons enrichi le contenu de l'image en la superposant séparément avec deux matrices Maillées (figure 4). Le maillage est une technique souvent utilisée dans le traitement des objets 3D soit pour améliorer l'affichage ou pour relier les squelettes d'un objet (figure 2). Pour l'implémentation de cette méthode on utilise le programme ci-dessous

Programme N°1 (Matlab)

```
% nous allons créer deux matrices x et y vide qui recevrons d'une manière uniforme et croissante les valeurs entre 0 et 127

Fonction [x, y, z, ds, dt] = mail()
smin = 127;
smax =  0;
```

1	2	3	4	5	6	7
1	2	3	4	5	6	7
1	2	3	4	5	6	7
1	2	3	4	5	6	7
1	2	3	4	5	6	7
1	2	3	4	5	6	7
1	2	3	4	5	6	7

Figure 4 : Exemple de l'effet de la fonction Meshgrid (0 :1 :7): Matrice (s)

```
ns   = 64;
ds   = (smax-smin)/ns;

% t d'une manière ascendante.
% (Du haut vers le bas)
tmin = 127.0;
tmax =  0.0;
nt   = 64;
dt   = (tmax-tmin)/nt;
% pour le maillage on peut utiliser la fonction mishgrid() de Matlab
[s,t] = meshgrid(smin:ds:smax, tmin:dt:tmax);
% lecture d'image contenant le caractère
% on divise les valeurs de ses pixels par 2
i=double (imread('z8.bmp'))/2;
%% on applique le changement du paramètre pour obtenir les cordonnées 3D
x  = s+i;
y  = t+i;
z  = s-s;
surf(x,y,z);  % projection des résultats.
```

La figure 5 présente un exemple des résultats projetés du premier algorithme

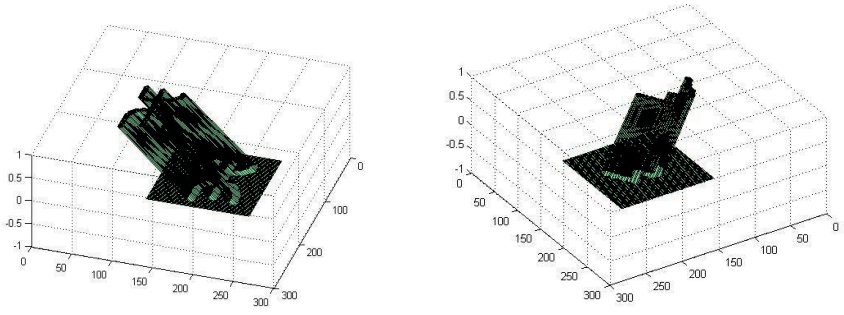


Figure 5 : Exemple de la projection des résultats obtenus par le programme 1, de deux caractères Tifinagh.

- Calcul de la Métrique

Par la suite chaque caractère est considéré comme une surface particulière pour lequel on calcule la métrique en chaque point.

La métrique est représentée par un tenseur (une matrice défini positive) souvent appelé G
est donnée par :

$$Gij = \begin{pmatrix} g11 & g12 \\ g21 & g22 \end{pmatrix} \quad (1)$$

Pour les vecteurs tangents t et s les coefficients métriques seront calculés comme suite :

$$g11=<DDs , DDs> ; g12=<DDs , DDt> ;$$
$$g21=<DDt , DDs> ; g22=<DDt , DDt> ; \quad (2)$$

Avec : DDs = $\left[\frac{\partial x}{\partial s}, \frac{\partial y}{\partial s}, \frac{\partial z}{\partial s}\right]$ et DDt = $\left[\frac{\partial x}{\partial t}, \frac{\partial y}{\partial t}, \frac{\partial z}{\partial t}\right]$ (3)

Calcule des descripteurs : les étapes à suivre sont comme suit :

- Le premier descripteur métrique est la somme de la déterminante de G en chaque point de la surface (chaque pixel de l'image) :

$$D1= \sum_{i=0,n} \sum_{j=0,m} G_{ij} \quad (4)$$

Avec n et m : le nombre de lignes et colonnes,

CGj la somme de la déterminante de G en chaque pixel j^{éme} colonne :

$$CGj= \sum_{i=0,n} G_{ij}$$

Et LGi la somme de la déterminante de G en chaque pixel i^{éme} ligne :

$$LGi= \sum_{j=0,m} G_{ij}$$

- On note successivement le deuxième et troisième descripteur métrique la somme des CGj allons de 0 à n/2 et de n/2+1 à n :

$$D2= \sum_{j=0,\frac{n}{2}} CGj \quad (5)$$

$$D3= \sum_{j=\frac{n}{2}+1,n} CGj \qquad (6)$$

- Le quatrième et cinquième descripteur métrique sont respectivement la somme des LGi allons de 0 à m/2 et de m/2+1 à n

$$D4= \sum_{j=0,\frac{n}{2}} CGj \qquad (7)$$

$$D5= \sum_{j=\frac{n}{2}+1,n} CGj \qquad (8)$$

La Figure 6 illustre des résultats obtenus pour les Cinque descripteurs décrit dans cet article :

	D1	D2	D3	D4	D5
	7.1017	4.5116	2.5901	3.5621	3.5396
	3.9435	2.5255	1.4181	2.4242	1.5193
	3.3036	2.0339	1.2697	1.8063	1.4974
	4.6660	2.7093	1.9566	3.3298	1.3362
	5.1201	3.4737	1.6464	2.4777	2.6423

Figure 6 : les résultats obtenus pour quelques caractères Tifinagh

On remarque que le premier descripteur permet de distinguer les caractères qui ont une morphologie distincte, tandis que le reste des descripteurs, après une normalisation par rapport à leur somme ils ont permis :

- de distinguer entre les caractères géométriquement proches (voir figure 7) exemple des caractères, Yarr et Yass.
- Et ils ont montré une résistance remarquable au changement d'échelle (voir figure 8)

	D1	D2	D3	D4	D5
	0.333	0.1538	0.1796	0.2070	0.1264
	0.333	0.1937	0.1396	0.1879	0.1455

Figure 7 : Les descripteurs métriques pour les deux caractères Yass et Yarr.

	D2	D3	D4	D5
	0.1443	0.1890	0.2250	0.1083
	0.1356	0.1978	0.2062	0.1271
	0.1402	0.1931	0.2175	0.1158

Figure 8 : les descripteurs métriques du caractère Yass calculé pour différents tailles.

Classification

Pour tester notre approche nous avons choisi deux classificateurs connu pour leur robustes et leur caractéristiques unique : Les Réseaux de neurones et Les SVM (les séparateurs a vaste marge)

Les Réseaux de Neurones

Dans notre approche, nous avons utilisé un réseau de neurones multicouches (deux couches) à apprentissage supervisé, entraîné par la rétro propagation du gradient, cela consiste à déterminer l'erreur commise par chaque neurone puis à modifier la valeur des poids pour minimiser cette erreur.

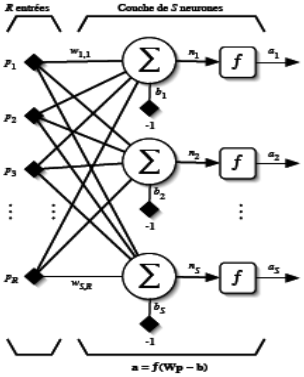


Figure 9 : Architecture du reseau de neurone (f : la fonction d'activation, Wi,j : les poids entre les neurones).

Les séparateurs à vaste marge SVM

Les SVM sont une généralisation des classifieurs linéaires. Ils reposent sur deux idées clés : la notion de marge maximale et la notion de fonction noyau. Les descripteurs métriques calculés précédemment ainsi que la classe de chaque caractère forment un ensemble d'apprentissage *{(xi, yi)} i=1..M*, où chaque xi $\in R^d$ et yi $\in$ {1..N} dans le cas où nous cherchons à reconnaître N classes différentes. Notre objectif est de construire une fonction f(x) qui estime les dépendances entre les exemples xi et les classes yi et qui minimise le risque d'erreur de classification pour un point x donné n'appartenant pas à la base d'apprentissage.

Résultats expérimentaux

Nous avons testé notre approche de reconnaissance des caractères Tifinagh sur deux bases de données la première de (Y. Ouguengay, M. Taalabi 2009), elle est composée de 2175 caractères Tifinagh imprimés avec différents tailles et styles (Figure 10). La deuxième est une base de données locale composée par 330 caractères manuscrits (Figures 6&7)

Figure 10 : Exemple d'un caractère Tifinagh de la base de 'Y.Ouguengay' pour différents styles et tailles.

Chaque caractère va être identifié en utilisant ses descripteurs métriques et la reconnaissance par SVM et Réseaux de neurones. Nous avons testé notre approche pour différentes nombres de caractères de la base d'apprentissage. La figure 11 représente les taux de reconnaissances des objets de la base de données.
On peut noter que malgré la taille de la base d'apprentissage qui n'a pas dépassé 15 échantillons par caractères .Les descripteurs proposés ont donné de très bons résultats pour les nombres de caractères inférieurs à 20. (Pour les deux classificateurs).et de bonne résultats pour la totalité des caractères Tifinagh (pour les SVM).
Les résultats son obtenu pour des images centrées.

Nombre de caractères imprimés a identifié	SVM	Réseaux de neurones
10	99%	99%
20	88%	67%
2	83%	--
33	78%	--

Figure 11 : Evaluation de taux de reconnaissance en fonction du nombre de caractères imprimé a identifié.

Nombre de caractères imprimés a identifié	SVM	Réseaux de neurones
10	97%	98%
20	84%	67%
25	81%	--
33	71%	--

Figure 12 : Evaluation de taux de reconnaissance en fonction du nombre de caractères manuscrit.

Altérations / %	Cm	Lu	Bg	Dg
Taux de reconnaissances avec les SVM	97%	93%	95%	92%

Figure 12 : Taux de reconnaissance (%) sur des images avec différentes altérations

Avec : Cm : caractère manuscrit, Lu : changement de luminance, Bg : bruit Gaussien, Dg : déformation géométrique.

Remarque : les erreurs remarqué parvient des styles d'écritures testé qui sont des fois assez embrouillés (Figure 10) et aux ressemblances géométriques provenant de l'écriture manuscrit dans l'exemple du Figure 13.

Figure 13 : Exemple ressemblances géométriques provenant de l'écriture manuelle de (yak) et (yarr).

Conclusion et perspectives

Dans ce travail nous avons utilisé la géométrie riemannienne comme nouvelle approche d'extraction de descripteurs, les réseaux de neurones et les SVM pour faire la classification

La robustesse de notre système de reconnaissance est illustrée sur une base de données Tifinagh complétées par des images présentant différentes altérations, comme : la déformation géométrique des caractères, la variation de luminance et la présence de bruit blanc gaussien de variance (%10).

- Le taux de reconnaissance pour une base d'apprentissage constituée de 15 échantillons pour 10 caractères est 98%.
- La robustesse obtenue est excellente vis à vis de la présence de bruit et bonne dans le cas de déformation géométrique et variation de luminance.

Les résultats obtenus peuvent être améliorés, en agissant sur plusieurs paramètres :

- l'usage en parallèle de d'autres descripteurs de forme

- l'intégration des caractéristiques discriminantes des différents caractères.
- l'augmentation de la base d'apprentissage.

Références

F.L. Alt.(1962), Digital pattern recognition by moments, *J. ACM*, pp. 240–258.

S.A. Dudani,(1977), Aircraft identification by moment invariants, *IEEE Trans. Comput*.,pp 39–45.

Chee-Way Chonga, P. Raveendranb and R. Mukundan,(2003),Translation invariants of Zernike moments", *Pattern Recognition* , pp 1765– 1773.

X.Gu ,(2004) Genus Zero Surfaceconformal apping,*IEEE TANSACTIONS ON MEDICAL IMAGING*,VOL.23 NO.8.

E.KALSEN & al (2004) Analysis of planar shapes using geodesic paths on shape spaces, *IEEE TRANSACTIOS ON PATTERN ANALYSIS AND MACHINEINTELIGENCE* , vol .26, No 3 .

Y.A.Ouguengay, M.Taalabi,(2009) "Elaboration d'un réseau de neurones artificiels pour la reconnaissance optique de la graphie amazighe : Phase d'apprentissage", *5e Conférence internationale sur les "Systèmes Intelligents : Théories et Applications", Paris* : Europia, cop. (impr. au Maroc), ISBN 978-2-909285-55-3 .

Oren Boiman, Eli Shechtman and Michal Irani(2008), In Defense of Nearest-Neighbor Based Image Classification, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

A. Bosch, A. Zisserman, and X. Munoz.(2007), Image classification using random forests and ferns.*In ICCV*.

SIMARD P., STEINKRAUS D., PLATT J. C.(2005), Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis, *ICDAR*, pp. 958-962

C.MAAOU ,(2005)'' Reconnaissance et détection robuste d'objets couleur '' *20th colloque GRETSI*, pp :727-730 .

E.KALSEN & al(2004), ''Analaysis of planar shapse using geodesic paths on shape spaces'',*IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MATCHING INTELIGENCE* , VOL 26, NO 3.

J.KERL (2008), Numerical differential geometry in Matlab, *Graduate student Colloquium,* university of Arizona.

يعتبر اعتماد التكنولوجيات الحديثة في تعليم اللغة الأمازيغية مسألة بديهية لضمان حضورها الكامل والقوي في العالم الرقمي. ومن المعلوم أن الأمازيغية تعد من اللغات القليلة الموارد، لذا فقد انطلقت أعمال بحث علمية ولغوية في هذا الاتجاه لإغناء الأمازيغية بالموارد اللغوية. وفي هذا الإطار فإن تطوير تطبيقات قادرة على معالجة المعطيات اللغوية الطبيعية بطريقة آلية، أصبحت من الحاجيات الضرورية للنهوض بالثقافة الأمازيغية وتطويرها.

إن الأعمال المنشورة في هذا الكتاب، والتي قدمت في إطار الدورة الثانية من الندوة SITACAM'11، تتعرض لمواضيع رئيسية مثل : تقنيات الترجمة الآلية والنقل الحرفي بين الأمازيغية والعربية واللغات اللاتينية، مقاربات التعرف الضوئي على حروف تيفيناغ، التحليل المعجمي والنحوي للغة الأمازيغية.

L'application des technologies de l'information et de communication (TIC) à l'apprentissage de la langue Amazighe est une condition sine qua non pour une intégration pleine et entière dans le monde informatisé. L'amazighe fait partie des langues peu dotées; par conséquent, des recherches scientifiques et linguistiques sont lancées dans ce sens pour améliorer la situation actuelle. La conception et la réalisation d'applications capables de traiter automatiquement des données linguistiques, exprimées dans la langue naturelle amazighe deviennent de plus en plus des besoins nécessaires pour le développement de la culture Amazighe.

Les travaux retenus dans cette 2ème édition du SITACAM et publiés dans cet ouvrage portent sur des thèmes majeurs, dont notamment : les techniques de traduction et la translittération automatique entre l'amazighe, le latin et l'arabe, les approches et systèmes de reconnaissance des caractères amazighes, la linguistique et l'analyse lexicale et syntaxique de la langue amazighe.

21 DH